THE INTERNET ENCYCLOPEDIA

Volume 2

Hossein Bidgoli

Editor-in-Chief

THE INTERNET ENCYCLOPEDIA

Volume 2
G–O

Hossein Bidgoli
Editor-in-Chief
California State University
Bakersfield, California

WILEY
John Wiley & Sons, Inc.

This book is printed on acid-free paper. ♾

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. The publisher is not engaged in rendering professional services, and you should consult a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books. For more information about Wiley products, visit our web site at www.Wiley.com.

Library of Congress Cataloging-in-Publication Data:

The Internet encyclopedia / edited by Hossein Bidgoli.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-22202-X (CLOTH VOL 1 : alk. paper) – ISBN 0-471-22204-6
(CLOTH VOL 2 : alk. paper) – ISBN 0-471-22203-8 (CLOTH VOL 3 : alk.
paper) – ISBN 0-471-22201-1 (CLOTH SET : alk. paper)
1. Internet–Encyclopedias. I. Bidgoli, Hossein.
TK5105.875.I57I5466 2003
004.67′8′03–dc21
2002155552

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To so many fine memories of my brother, Mohsen, for his
uncompromising belief in the power of education.

About the Editor-in-Chief

Hossein Bidgoli, Ph.D., is Professor of Management Information Systems at California State University. Dr. Bidgoli helped set up the first PC lab in the United States. He is the author of 43 textbooks, 27 manuals, and over four dozen technical articles and papers on various aspects of computer applications, e-commerce, and information systems, which have been published and presented throughout the world. Dr. Bidgoli also serves as the editor-in-chief of *Encyclopedia of Information Systems*.

Dr. Bidgoli was selected as the California State University, Bakersfield's 2001–2002 Professor of the Year.

Editorial Board

Contents

Volume 1

Volume 2

Volume 3

Chapter List by Subject Area

Applications

Design, Implementation, and Management

Electronic Commerce

Foundation

Infrastructure

Legal, Social, Organizational, International, and Taxation Issues

Marketing and Advertising on the Web

Security Issues and Measures

Supply Chain Management

Web Design and Programming

Wireless Internet and E-commerce

Contributors

Tarek Abdelzaher
University of Virginia
Web Quality of Service

Charles Abzug
James Madison University
Linux Operating System

Patricia Adams
Education Resources
Strategic Alliances

Carol A. Akerelrea
Colorado State University
Usability Testing: An Evaluation Process for Internet Communications

Gary C. Anders
Arizona State University West
Online Auctions

Amy W. Apon
University of Arkansas
Public Networks

Pierre A. Balthazard
Arizona State University West
Groupware

Ashok Deo Bardhan
University of California, Berkeley
Real Estate

Joey Bargsten
University of Oregon
Multimedia

Hossein Bidgoli
California State University, Bakersfield
Computer Literacy
Internet Literacy

Gerald Bluhm
Tyco Fire & Security
Patent Law

Robert J. Boncella
Washburn University
Secure Sockets Layer (SSL)

J. Efrim Boritz
University of Waterloo, Canada
XBRL (Extensible Business Reporting Language): Business Reporting with XML

Sviatoslav Braynov
State University of New York at Buffalo
Data Mining in E-commerce
Personalization and Customization Technologies

Randy M. Brooks
Millikin University
Online Publishing

Colleen Brown
Purdue University
History of the Internet

Tara Brown-L'Bahy
Harvard University
Distance Learning (Virtual Learning)

Linda S. Bruenjes
Lasell College
Internet2

Gerard J. Burke
University of Florida
Supply Chain Management

L. Jean Camp
Harvard University
Peer-to-Peer Systems

Charles J. Campbell
The University of Memphis
Java

Janice E. Carrillo
University of Florida
Inventory Management

Michael A. Carrillo
Oracle Corporation
Inventory Management

Lillian N. Cassel
Villanova University
Wireless Application Protocol (WAP)

J. Cecil
New Mexico State University
Virtual Enterprises

Haluk Cetin
Murray State University
Geographic Information Systems (GIS) and the Internet

Henry Chan
The Hong Kong Polytechnic University, China
Consumer-Oriented Electronic Commerce

C. Janie Chang
San José State University
Public Accounting Firms

Camille Chin
West Virginia University
Cybercrime and Cyberfraud

T. Matthew Ciolek
The Australian National University, Australia
Online Religion

Timothy W. Cole
University of Illinois at Urbana-Champaign
Visual Basic Scripting Edition (VBScript)

Fred Condo
California State University, Chico
Cascading Style Sheets (CSS)

David E. Cook
University of Derby, United Kingdom
Standards and Protocols in Data Communications

Marco Cremonini
Università di Milano, Italy
Disaster Recovery Planning

Mary J. Cronin
Boston College
Mobile Commerce
Jaime J. Dávila
Hampshire College
Digital Divide
Chris Dede
Harvard University
Distance Learning (Virtual Learning)
Victoria S. Dennis
Minnesota State Bar Association
Law Firms
Lynn A. DeNoia
Rensselaer Polytechnic Institute
Wide Area and Metropolitan Area Networks
Nikhilesh Dholakia
University of Rhode Island
Gender and Internet Usage
Global Diffusion of the Internet
Ruby Roy Dholakia
University of Rhode Island
Gender and Internet Usage
Global Diffusion of the Internet
Vesna Dolnicar
University of Ljubljana, Slovenia
Benchmarking Internet
Rich Dorfman
WebFeats! and Waukesha County Technical College
Extensible Markup Language (XML)
Magda El Zarki
University of California—Irvine
Wireless Internet
Larry P. English
Information Impact International, Inc.
Information Quality in Internet and E-business Environments
Roman Erenshteyn
Goldey-Beacom College
ActiveX
Ray Everett-Church
ePrivacy Group, Inc.
Privacy Law
Trademark Law
Patrick J. Fahy
Athabasca University, Canada
Web-Based Training
Gerald R. Ferrera
Bentley College
Copyright Law
Daniel R. Fesenmaier
University of Illinois at Urbana–Champaign
Travel and Tourism
C. Patrick Fleenor
Seattle University
Feasibility of Global E-business Projects
Marcia H. Flicker
Fordham University
Securities Trading on the Internet
Immanuel Freedman
Dr. Immanuel Freedman, Inc.
Video Compression
Borko Furht
Florida Atlantic University
Interactive Multimedia on the Web
Jayne Gackenbach
Athabasca University, Canada
Health Issues
Alan Gaitenby
University of Massachusetts, Amherst
Online Dispute Resolution
Bruce Garrison
University of Miami
Online News Services (Online Journalism)
G. David Garson
North Carolina State University
E-government
Roger Gate
IBM United Kingdom Ltd., United Kingdom
Electronic Funds Transfer
Mario Giannini
Code Fighter, Inc., and Columbia University
C/C++
Julia Alpert Gladstone
Bryant College
International Cyberlaw
Mary C. Gilly
University of California, Irvine
Consumer Behavior
Robert H. Goffman
Concordia University
Electronic Procurement
James E. Goldman
Purdue University
Firewalls
Sven Graupner
Hewlett-Packard Laboratories
Web Services
Robert H. Greenfield
Computer Consulting
Circuit, Message, and Packet Switching
Ulrike Gretzel
University of Illinois at Urbana–Champaign
Travel and Tourism
Paul Gronke
Reed College
Politics
Jim Grubbs
University of Illinois at Springfield
E-mail and Instant Messaging
Mohsen Guizani
Western Michigan University
Wireless Communications Applications
Jon Gunderson
University of Illinois at Urbana–Champaign
Universally Accessible Web Resources: Designing for People with Disabilities
Babita Gupta
California State University, Monterey Bay
Global Issues
Louisa Ha
Bowling Green State University
Webcasting

Kirk Hallahan
Colorado State University
Online Public Relations

Diane M. Hamilton
Rowan University
Business-to-Consumer (B2C) Internet Business Models

Robert W. Heath Jr.
The University of Texas at Austin
Digital Communication

Geert Heijenk
University of Twente, The Netherlands
Wireless Internet

Jesse M. Heines
University of Massachusetts Lowell
Extensible Stylesheet Language (XSL)

Rodney J. Heisterberg
Notre Dame de Namur University and
Rod Heisterberg Associates
Collaborative Commerce (C-Commerce)

Steven J. Henry
Wolf, Greenfield & Sacks, P.C.
Open Source Development and Licensing

Julie Hersberger
University of North Carolina at Greensboro
Internet Censorship

Kenneth Einar Himma
University of Washington
Legal, Social, and Ethical Issues

Matthias Holweg
Massachusetts Institute of Technology
Managing the Flow of Materials Across the Supply Chain

Russ Housley
Vigil Security, LLC
Public Key Infrastructure (PKI)

Yeong-Hyeon Hwang
University of Illinois at Urbana–Champaign
Travel and Tourism

Robert E. Irie
SPAWAR Systems Center San Diego
Web Site Design

Linda C. Isenhour
University of Central Florida
Human Resources Management

Hans-Arno Jacobsen
University of Toronto, Canada
Application Service Providers (ASPs)

Charles W. Jaeger
Southerrn Oregon University
Cyberterrorism

Dwight Jaffee
University of California, Berkeley
Real Estate

Sushil Jajodia
George Mason University
Intrusion Detection Techniques

Mark Jeffery
Northwestern University
Return on Investment Analysis for E-business Projects

Andrew Johnson
University of Illinois at Chicago
Virtual Reality on the Internet: Collaborative Virtual Reality

Ari Juels
RSA Laboratories
Encryption

Bhushan Kapoor
California State University, Fullerton
ActiveX Data Objects (ADO)

Joseph M. Kayany
Western Michigan University
Internet Etiquette (Netiquette)

Doug Kaye
RDS Strategies LLC
Web Hosting

Chuck Kelley
Excellence In Data, Inc.
Data Warehousing and Data Marts

Diane Ketelhut
Harvard University
Distance Learning (Virtual Learning)

Chang-Su Kim
Seoul National University, Korea
Data Compression

Wooyoung Kim
University of Illinois at Urbana-Champaign
Web Services

Jerry Kindall
Epok Inc.
Digital Identity

Brad Kleindl
Missouri Southern State University–Joplin
Value Chain Analysis

Graham Knight
University College London, United Kingdom
Internet Architecture

Craig D. Knuckles
Lake Forest College
DHTML (Dynamic HyperText Markup Language)

Jim Krause
Indiana University
Enhanced TV

Peter Kroon
Agere Systems
Speech and Audio Compression

Gary J. Krug
Eastern Washington University
Convergence of Data, Sound, and Video

Nir Kshetri
University of North Carolina
Gender and Internet Usage
Global Diffusion of the Internet

C.-C. Jay Kuo
University of Southern California
Data Compression

Stan Kurkovsky
Columbus State University
Common Gateway Interface (CGI) Scripts

Pamela M. H. Kwok
Hong Kong Polytechnic University, China
Wireless Marketing

Jennifer Lagier
Hartnell College
File Types

Thomas D. Lairson
Rollins College
Supply Chain Management and the Internet
Gary LaPoint
Syracuse University
International Supply Chain Management
Haniph A. Latchman
University of Florida
Managing a Network Environment
John LeBaron
University of Massachusetts Lowell
Internet2
Kenneth S. Lee
University of Pennsylvania
Wireless Internet
Jason Leigh
University of Illinois at Chicago
Virtual Reality on the Internet: Collaborative Virtual Reality
Margarita Maria Lenk
Colorado State University
Guidelines for a Comprehensive Security System
Nanette S. Levinson
American University
Developing Nations
Edwin E. Lewis Jr.
Johns Hopkins University
E-business ROI Simulations
David J. Loundy
DePaul University
Online Stalking
Robert H. Lowson
University of East Anglia, United Kingdom
E-systems for the Support of Manufacturing Operations
Supply Networks: Developing and Maintaining Relationships and Strategies
David Lukoff
Saybrook Graduate School and Research Center
Health Issues
Kuber Maharjan
Purdue University
Downloading from the Internet
Julie R. Mariga
Purdue University
Mobile Devices and Protocols
Mobile Operating Systems and Applications
Oge Marques
Florida Atlantic University
Interactive Multimedia on the Web
Prabhaker Mateti
Wright State University
TCP/IP Suite
Bruce R. Maxim
University of Michigan–Dearborn
Game Design: Games for the World Wide Web
Blayne E. Mayfield
Oklahoma State University
Visual C++ (Microsoft)
Cavan McCarthy
Louisiana State University
Digital Libraries
Patrick McDaniel
AT&T Labs
Authentication
David E. McDysan
WorldCom
Virtual Private Networks: Internet Protocol (IP) Based
Daniel J. McFarland
Rowan University
Client/Server Computing
Matthew K. McGowan
Bradley University
Electronic Data Interchange (EDI)
Nenad Medvidovic
University of Southern California
JavaBeans and Software Architecture
Nikunj R. Mehta
University of Southern California
JavaBeans and Software Architecture
John A. Mendonca
Purdue University
Organizational Impact
Weiyi Meng
State University of New York at Binghamton
Web Search Technology
Mark S. Merkow
E-commerce Guide
Secure Electronic Transactions (SET)
Mark Michael
King's College
HTML/XHTML (HyperText Markup Language/ Extensible HyperText Markup Language)
Physical Security
Brent A. Miller
IBM Corporation
Bluetooth^{TM}—A Wireless Personal Area Network
Robert K. Moniot
Fordham University
Software Piracy
Joseph Morabito
Stevens Institute of Technology
Online Analytical Processing (OLAP)
Roy Morris
Capitol College
Voice over Internet Protocol (IP)
Alec Nacamuli
IBM United Kingdom Ltd., United Kingdom
Electronic Funds Transfer
Annette Nellen
San José State University
Public Accounting Firms
Taxation Issues
Dale Nesbary
Oakland University
Nonprofit Organizations
Dat-Dao Nguyen
California State University, Northridge
Business-to-Business (B2B) Internet Business Models
Peng Ning
North Carolina State University
Intrusion Detection Techniques

Mark E. Nissen
Naval Postgraduate School
Intelligent Agents
Won Gyun No
University of Waterloo, Canada
XBRL (Extensible Business Reporting Language): Business Reporting with XML
Eric H. Nyberg
Carnegie Mellon University
Prototyping
Jeff Offutt
George Mason University
Software Design and Implementation in the Web Environment
Donal O'Mahony
University of Dublin, Ireland
Electronic Payment
Robert Oshana
Southern Methodist University
Capacity Planning for Web Services
Dennis O. Owen
Purdue University
Visual Basic
Raymond R. Panko
University of Hawaii at Manoa
Computer Security Incident Response Teams (CSIRTs)
Digital Signatures and Electronic Signatures
Internet Security Standards
Anand Paul
University of Florida
Inventory Management
Thomas L. Pigg
Jackson State Community College
Conducted Communications Media
Paul S. Piper
Western Washington University
Research on the Internet
Benjamin R. Pobanz
Purdue University
Mobile Devices and Protocols
Richard E. Potter
University of Illinois at Chicago
Groupware
Dennis M. Powers
Southern Oregon University
Cyberlaw: The Major Areas, Development, and Provisions
Paul R. Prabhaker
Illinois Institute of Technology
E-marketplaces
Etienne E. Pracht
University of South Florida
Health Insurance and Managed Care
Frederick Pratter
Eastern Oregon University
JavaServer Pages (JSP)
Robert W. Proctor
Purdue University
Human Factors and Ergonomics
Jian Qin
Syracuse University
Web Content Management
Zinovy Radovilsky
California State University, Hayward
Enterprise Resource Planning (ERP)
Jeremy Rasmussen
Sypris Electronics, LLC
Passwords
Peter Raven
Seattle University
Feasibility of Global E-business Projects
Amy W. Ray
Bentley College
Business Plans for E-commerce Projects
Julian J. Ray
Western New England College
Business-to-Business (B2B) Electronic Commerce
Pratap Reddy
Raritan Valley Community College
Internet Navigation (Basics, Services, and Portals)
Drummond Reed
OneName Corporation
Digital Identity
Vladimir V. Riabov
Rivier College
Storage Area Networks (SANs)
Nick Rich
Cardiff Business School, United Kingdom
Managing the Flow of Materials Across the Supply Chain
Malu Roldan
San Jose State University
Marketing Plans for an E-commerce Project
Constantine Roussos
Lynchburg College
JavaScript
Akhil Sahai
Hewlett-Packard Laboratories
Web Services
Eduardo Salas
University of Central Florida
Human Resources Management
Atul A. Salvekar
Intel Corp.
Digital Communication
Pierangela Samarati
Università di Milano, Italy
Disaster Recovery Planning
J. Christopher Sandvig
Western Washington University
Active Server Pages
Robert J. Schalkoff
Clemson University
Rule-Based and Expert Systems
Shannon Schelin
North Carolina State University
E-government
William T. Schiano
Bentley College
Intranets
Roy C. Schmidt
Bradley University
Risk Management in Internet-Based Software Projects

E. Eugene Schultz
University of California–Berkley Lab
Denial of Service Attacks
Windows 2000 Security

Steven D. Schwaitzberg
Tufts-New England Medical Center
Medical Care Delivery

Kathy Schwalbe
Augsburg College
Project Management Techniques

Mark Shacklette
The University of Chicago
Unix Operating System

P. M. Shankar
Drexel University
Propagation Characteristics of Wireless Channels

John Sherry
Purdue University
History of the Internet

Carolyn J. Siccama
University of Massachusetts Lowell
Internet2

Judith C. Simon
The University of Memphis
Java
Law Enforcement
Law Firms

Robert Simon
George Mason University
Middleware

Nirvikar Singh
University of California, Santa Cruz
Digital Economy

Clara L. Sitter
University of Denver
Library Management

Robert Slade
Consultant
Computer Viruses and Worms

Erick D. Slazinski
Purdue University
Structured Query Language (SQL)

Mark Smith
Purdue University
Supply Chain Management Technologies

Lee Sproull
New York University
Online Communities

Charles Steinfield
Michigan State University
Click-and-Brick Electronic Commerce
Electronic Commerce and Electronic Business

Edward A. Stohr
Stevens Institute of Technology
Online Analytical Processing (OLAP)

Dianna L. Stone
University of Central Florida
Human Resources Management

David Stotts
University of North Carolina at Chapel Hill
Perl

Judy Strauss
University of Nevada, Reno
Marketing Communication Strategies

Wayne C. Summers
Columbus State University
Local Area Networks

Jamie S. Switzer
Colorado State University
Virtual Teams

Dale R. Thompson
University of Arkansas
Public Networks

John S. Thompson
University of Colorado at Boulder
Integrated Services Digital Network (ISDN): Narrowband and Broadband Services and Applications

Stephen W. Thorpe
Neumann College
Extranets

Ronald R. Tidd
Central Washington University
Knowledge Management

Herbert Tuttle
The University of Kansas
Video Streaming

Okechukwu C. Ugweje
The University of Akron
Radio Frequency and Wireless Communications

Asoo J. Vakharia
University of Florida
Supply Chain Management

Robert Vaughn
University of Memphis
Law Enforcement

Vasja Vehovar
University of Ljubljana, Slovenia
Benchmarking Internet

Kim-Phuong L. Vu
Purdue University
Human Factors and Ergonomics

Jordan Walters
BCN Associates, Inc.
Managing a Network Environment

Siaw-Peng Wan
Elmhurst College
Online Banking and Beyond: Internet-Related Offerings from U.S. Banks

Youcheng Wang
University of Illinois at Urbana–Champaign
Travel and Tourism

James. L. Wayman
San Jose State University
Biometric Authentication

Scott Webster
Syracuse University
International Supply Chain Management

Jianbin Wei
Wayne State University
Load Balancing on the Internet

Ralph D. Westfall
California State Polytechnic University, Pomona
Telecommuting and Telework

Pamela Whitehouse
Harvard University
Distance Learning (Virtual Learning)

Dave Whitmore
Champlain College
Multiplexing

Russell S. Winer
New York University
Customer Relationship Management on the Web

Raymond Wisman
Indiana University Southeast
Web Search Fundamentals

Paul L. Witt
University of Texas at Arlington
Internet Relay Chat (IRC)

Mary Finley Wolfinbarger
California State University, Long Beach
Consumer Behavior

Peter R. Wurman
North Carolina State University
Online Auction Site Management

Cheng-Zhong Xu
Wayne State University
Load Balancing on the Internet

Qiang Yang
Hong Kong University of Science and Technology, China
Machine Learning and Data Mining on the Web

A. Neil Yerkey
University at Buffalo
Databases on the Web

Clement Yu
University of Illinois at Chicago
Web Search Technology

Daniel Dajun Zeng
University of Arizona
Intelligent Agents

Yan-Qing Zhang
Georgia State University
Fuzzy Logic

Xiaobo Zhou
University of Colorado at Colorado Springs
Load Balancing on the Internet

Donald E. Zimmerman
Colorado State University
Usability Testing: An Evaluation Process for Internet Communications

Preface

The Internet Encyclopedia is the first comprehensive examination of the core topics in the Internet field. *The Internet Encyclopedia*, a three-volume reference work with 205 chapters and more than 2,600 pages, provides comprehensive coverage of the Internet as a business tool, IT platform, and communications and commerce medium. The audience includes the libraries of two-year and four-year colleges and universities with MIS, IT, IS, data processing, computer science, and business departments; public and private libraries; and corporate libraries throughout the world. It is the only comprehensive source for reference material for educators and practitioners in the Internet field.

Education, libraries, health, medical, biotechnology, military, law enforcement, accounting, law, justice, manufacturing, financial services, insurance, communications, transportation, aerospace, energy, and utilities are among the fields and industries expected to become increasingly dependent upon the Internet and Web technologies. Companies in these areas are actively researching the many issues surrounding the design, utilization, and implementation of these technologies.

This definitive three-volume encyclopedia offers coverage of both established and cutting-edge theories and developments of the Internet as a technical tool and business/communications medium. The encyclopedia contains chapters from global experts in academia and industry. It offers the following unique features:

1) Each chapter follows a format which includes title and author, chapter outline, introduction, body, conclusion, glossary, cross references, and references. This unique format enables the readers to pick and choose among various sections of a chapter. It also creates consistency throughout the entire series.
2) The encyclopedia has been written by more than 240 experts and reviewed by more than 840 academics and practitioners chosen from around the world. This diverse collection of expertise has created the most definitive coverage of established and cutting edge theories and applications in this fast-growing field.
3) Each chapter has been rigorously peer reviewed. This review process assures the accuracy and completeness of each topic.
4) Each chapter provides extensive online and offline references for additional readings. This will enable readers to further enrich their understanding of a given topic.
5) More than 1,000 illustrations and tables throughout the series highlight complex topics and assist further understanding.
6) Each chapter provides extensive cross references. This helps the readers identify other chapters within the encyclopedia related to a particular topic, which provides a one-stop knowledge base for a given topic.
7) More than 2,500 glossary items define new terms and buzzwords throughout the series, which assists readers in understanding concepts and applications.
8) The encyclopedia includes a complete table of contents and index sections for easy access to various parts of the series.
9) The series emphasizes both technical and managerial issues. This approach provides researchers, educators, students, and practitioners with a balanced understanding of the topics and the necessary background to deal with problems related to Internet-based systems design, implementation, utilization, and management.
10) The series has been designed based on the current core course materials in several leading universities around the world and current practices in leading computer- and Internet-related corporations. This format should appeal to a diverse group of educators, practitioners, and researchers in the Internet field.

We chose to concentrate on fields and supporting technologies that have widespread applications in the academic and business worlds. To develop this encyclopedia, we carefully reviewed current academic research in the Internet field at leading universities and research institutions around the world. Management information systems, decision support systems (DSS), supply chain management, electronic commence, network design and management, and computer information systems (CIS) curricula recommended by the Association of Information Technology Professionals (AITP) and the Association for Computing Management (ACM) were carefully investigated. We also researched the current practices in the Internet field used by leading IT corporations. Our work enabled us to define the boundaries and contents of this project.

TOPIC CATEGORIES

Based on our research we identified 11 major topic areas for the encyclopedia:

- Foundation;
- Infrastructure;
- Legal, social, organizational, international, and taxation issues;
- Security issues and measures;
- Web design and programming;
- Design, implementation, and management;
- Electronic commerce;
- Marketing and advertising on the Web;

- Supply chain management;
- Wireless Internet and e-commerce; and
- Applications.

Although these 11 categories of topics are interrelated, each addresses one major dimension of the Internet-related fields. The chapters in each category are also interrelated and complementary, enabling readers to compare, contrast, and draw conclusions that might not otherwise be possible.

Although the entries have been arranged alphabetically, the light they shed knows no bounds. The encyclopedia provides unmatched coverage of fundamental topics and issues for successful design, implementation, and utilization of Internet-based systems. Its chapters can serve as material for a wide spectrum of courses, such as the following:

- Web technology fundamentals;
- E-commerce;
- Security issues and measures for computers, networks, and online transactions;
- Legal, social, organizational, and taxation issues raised by the Internet and Web technology;
- Wireless Internet and e-commerce;
- Supply chain management;
- Web design and programming;
- Marketing and advertising on the Web; and
- The Internet and electronic commerce applications.

Successful design, implementation, and utilization of Internet-based systems require a thorough knowledge of several technologies, theories, and supporting disciplines. Internet and Web technologies researchers and practitioners have had to consult many resources to find answers. Some of these sources concentrate on technologies and infrastructures, some on social and legal issues, and some on applications of Internet-based systems. This encyclopedia provides all of this relevant information in a comprehensive three-volume set with a lively format.

Each volume incorporates core Internet topics, practical applications, and coverage of the emerging issues in the Internet and Web technologies field. Written by scholars and practitioners from around the world, the chapters fall into the 11 major subject areas mentioned previously.

Foundation

Chapters in this group examine a broad range of topics. Theories and concepts that have a direct or indirect effect on the understanding, role, and the impact of the Internet in public and private organizations are presented. They also highlight some of the current issues in the Internet field. These articles explore historical issues and basic concepts as well as economic and value chain concepts. They address fundamentals of Web-based systems as well as Web search issues and technologies. As a group they provide a solid foundation for the study of the Internet and Web-based systems.

Infrastructure

Chapters in this group explore the hardware, software, operating systems, standards, protocols, network systems, and technologies used for design and implementation of the Internet and Web-based systems. Thorough discussions of TCP/IP, compression technologies, and various types of networks systems including LANs, MANS, and WANs are presented.

Legal, Social, Organizational, International, and Taxation Issues

These chapters look at important issues (positive and negative) in the Internet field. The coverage includes copyright, patent and trademark laws, privacy and ethical issues, and various types of cyberthreats from hackers and computer criminals. They also investigate international and taxation issues, organizational issues, and social issues of the Internet and Web-based systems.

Security Issues and Measures

Chapters in this group provide a comprehensive discussion of security issues, threats, and measures for computers, network systems, and online transactions. These chapters collectively identify major vulnerabilities and then provide suggestions and solutions that could significantly enhance the security of computer networks and online transactions.

Web Design and Programming

The chapters in this group review major programming languages, concepts, and techniques used for designing programs, Web sites, and virtual storefronts in the e-commerce environment. They also discuss tools and techniques for Web content management.

Design, Implementation, and Management

The chapters in this group address a host of issues, concepts, theories and techniques that are used for design, implementation, and management of the Internet and Web-based systems. These chapters address conceptual issues, fundamentals, and cost benefits and returns on investment for Internet and e-business projects. They also present project management and control tools and techniques for the management of Internet and Web-based systems.

Electronic Commerce

These chapters present a thorough discussion of electronic commerce fundamentals, taxonomies, and applications. They also discuss supporting technologies and applications of e-commerce inclining intranets, extranets, online auctions, and Web services. These chapters clearly demonstrate the successful applications of the Internet and Web technologies in private and public sectors.

Marketing and Advertising on the Web

The chapters in this group explore concepts, theories, and technologies used for effective marketing and advertising on the Web. These chapters examine both qualitative and quantitative techniques. They also investigate the emerging technologies for mass personalization and customization in the Web environment.

Supply Chain Management

The chapters in this group discuss the fundamentals concepts and theories of value chain and supply chain management. The chapters examine the major role that the Internet and Web technologies play in an efficient and effective supply chain management program.

Wireless Internet and E-commerce

These chapters look at the fundamental concepts and technologies of wireless networks and wireless computing as they relate to the Internet and e-commerce operations. They also discuss mobile commerce and wireless marketing as two of the growing fields within the e-commerce environment.

Applications

The Internet and Web-based systems are everywhere. In most cases they have improved the efficiency and effectiveness of managers and decision makers. Chapters in this group highlight applications of the Internet in several fields, such as accounting, manufacturing, education, and human resources management, and their unique applications in a broad section of the service industries including law, law enforcement, medical delivery, health insurance and managed care, library management, nonprofit organizations, banking, online communities, dispute resolution, news services, public relations, publishing, religion, politics, and real estate. Although these disciplines are different in scope, they all utilize the Internet to improve productivity and in many cases to increase customer service in a dynamic business environment.

Specialists have written the collection for experienced and not-so-experienced readers. It is to these contributors that I am especially grateful. This remarkable collection of scholars and practitioners has distilled their knowledge into a fascinating and enlightening one-stop knowledge base in Internet-based systems that "talk" to readers. This has been a massive effort but one of the most rewarding experiences I have ever undertaken. So many people have played a role that it is difficult to know where to begin.

I should like to thank the members of the editorial board for participating in the project and for their expert advice on the selection of topics, recommendations for authors, and review of the materials. Many thanks to the more than 840 reviewers who devoted their times by proving advice to me and the authors on improving the coverage, accuracy, and comprehensiveness of these materials.

I thank my senior editor at John Wiley & Sons, Matthew Holt, who initiated the idea of the encyclopedia back in spring of 2001. Through a dozen drafts and many reviews, the project got off the ground and then was managed flawlessly by Matthew and his professional team. Matthew and his team made many recommendations for keeping the project focused and maintaining its lively coverage. Tamara Hummel, our superb editorial coordinator, exchanged several hundred e-mail messages with me and many of our authors to keep the project on schedule. I am grateful to all her support. When it came to the production phase, the superb Wiley production team took over. Particularly I want to thank Deborah DeBlasi, our senior production editor at John Wiley & Sons, and Nancy J. Hulan, our project manager at TechBooks. I am grateful to all their hard work.

Last, but not least, I want to thank my wonderful wife Nooshin and my two lovely children Mohsen and Morvareed for being so patient during this venture. They provided a pleasant environment that expedited the completion of this project. Nooshin was also a great help in designing and maintaining the author and reviewer databases. Her efforts are greatly appreciated. Also, my two sisters Azam and Akram provided moral support throughout my life. To this family, any expression of thanks is insufficient.

Hossein Bidgoli
California State University, Bakersfield

Guide to the Internet Encyclopedia

The Internet Encyclopedia is a comprehensive summary of the relatively new and very important field of the Internet. This reference work consists of three separate volumes and 205 chapters on various aspects of this field. Each chapter in the encyclopedia provides a comprehensive overview of the selected topic intended to inform a board spectrum of readers ranging from computer professionals and academicians to students to the general business community.

In order that you, the reader, will derive the greatest possible benefit from *The Internet Encyclopedia,* we have provided this Guide. It explains how the information within the encyclopedia can be located.

ORGANIZATION

The Internet Encyclopedia is organized to provide maximum ease of use for its readers. All of the chapters are arranged in alphabetical sequence by title. Chapters titles that begin with the letters A to F are in Volume 1, chapter titles from G to O are in Volume 2, and chapter titles from P to Z are in Volume 3. So that they can be easily located, chapter titles generally begin with the key word or phrase indicating the topic, with any descriptive terms following. For example, "Virtual Reality on the Internet: Collaborative Virtual Reality" is the chapter title rather than "Collaborative Virtual Reality."

Table of Contents

A complete table of contents for the entire encyclopedia appears in the front of each volume. This list of titles represents topics that have been carefully selected by the editor-in-chief, Dr. Hossein Bidgoli, and his colleagues on the Editorial Board.

Following this list of chapters by title is a second complete list, in which the chapters are grouped according to subject area. The encyclopedia provides coverage of 11 specific subject areas, such as E-commerce and Supply Chain Management. Please see the Preface for a more detailed description of these subject areas.

Index

The Subject Index is located at the end of Volume 3. This index is the most convenient way to locate a desired topic within the encyclopedia. The subjects in the index are listed alphabetically and indicate the volume and page number where information on this topic can be found.

Chapters

Each chapter in *The Internet Encyclopedia* begins on a new page, so that the reader may quickly locate it. The author's name and affiliation are displayed at the beginning of the article.

All chapters in the encyclopedia are organized according to a standard format, as follows:

- Title and author,
- Outline,
- Introduction,
- Body,
- Conclusion,
- Glossary,
- Cross References, and
- References.

Outline

Each chapter begins with an outline indicating the content to come. This outline provides a brief overview of the chapter so that the reader can get a sense of the information contained there without having to leaf through the pages. It also serves to highlight important subtopics that will be discussed within the chapter. For example, the chapter "Computer Literacy" includes sections entitled Defining a Computer, Categories of Computers According to Their Power, and Classes of Data Processing Systems. The outline is intended as an overview and thus lists only the major headings of the chapter. In addition, lower-level headings will be found within the chapter.

Introduction

The text of each chapter begins with an introductory section that defines the topic under discussion and summarizes the content. By reading this section the readers get a general idea about the content of a specific chapter.

Body

The body of each chapter discusses the items that were listed in the outline section.

Conclusion

The conclusion section provides a summary of the materials discussed in each chapter. This section imparts to the readers the most important issues and concepts discussed within each chapter.

Glossary

The glossary contains terms that are important to an understanding of the chapter and that may be unfamiliar to the reader. Each term is defined in the context of the particular chapter in which it is used. Thus the same term may be defined in two or more chapters with the detail of the definition varying slightly from one to another. The encyclopedia includes approximately 2,500 glossary terms.

For example, the article "Computer Literacy" includes the following glossary entries:

Computer A machine that accepts data as input, processes the data without human interference using a set of stored instructions, and outputs information. Instructions are step-by-step directions given to a computer for performing specific tasks.

Computer generations Different classes of computer technology identified by a distinct architecture and technology; the first generation was vacuum tubes, the second transistors, the third integrated circuits, the fourth very-large-scale integration, and the fifth gallium arsenide and parallel processing.

Cross References

All the chapters in the encyclopedia have cross references to other chapters. These appear at the end of the chapter, following the text and preceding the references. The cross references indicate related chapters which can be consulted for further information on the same topic. The encyclopedia contains more than 2,000 cross references in all. For example, the chapter "Java" has the following cross references:

JavaBeans and Software Architecture; Software Design and Implementation in the Web Environment.

References

The reference section appears as the last element in a chapter. It lists recent secondary sources to aid the reader in locating more detailed or technical information. Review articles and research papers that are important to an understanding of the topic are also listed. The references in this encyclopedia are for the benefit of the reader, to provide direction for further research on the given topic. Thus they typically consist of one to two dozen entries. They are not intended to represent a complete listing of all materials consulted by the author in preparing the chapter. In addition, some chapters contain a Further Reading section, which includes additional sources readers may wish to consult.

G

Game Design: Games for the World Wide Web

Bruce R. Maxim, *University of Michigan–Dearborn*

INTRODUCTION

Programming computers to play games attracted the interest of such computer scientists as Babbage, Turing, and Shannon long before the first personal computers (PCs) came into existence. When personal computers and dedicated game consoles became widely available, in the 1970s, large numbers of computer game titles appeared in the marketplace almost immediately. Some people have argued that the computational expectations of modern computer game players pushed PC hardware manufacturers to increase the multimedia processing power of the home computer. The advent of the Internet and its rapid spread to homes throughout the world similarly has attracted the attention of game designers seeking to create multiplayer games among widely dispersed users. Creating games intended to run on the World Wide Web requires a good understanding of the principles used to design games for stand-alone computers, as well as knowledge about network architecture and client/server computing. This chapter discusses some of the design principles and challenges posed by developing games for the World Wide Web.

ECONOMIC IMPACT OF COMPUTER GAME INDUSTRY

The computer game industry has had gross sales in excess of 7 billion dollars since 1998 (Sawyer, Dunne, & Berg, 1998). Some industry analysts have predicted that sometime during the next ten years computer game play will be the most popular means of family entertainment. Indeed, by some estimates, computer game sales may have exceeded motion picture revenues for the year 2001. A recent survey showed that the average age of computer game buyers is about 31 years old (WINNY, 2000). Computer game development is serious business. Major game development projects often have production budgets exceeding 2 million dollars for a single title and sequels abound.

The Internet has the potential to allow game developers to increase their share of the entertainment market as the number of households having access to the World Wide Web continues to grow. Users of the World Wide Web can download commercial games to play on their personal computers, participate in multiplayer games online, and find support communities for their favorite games. There are several online game services that derive their sole income from providing access to Internet games for a fee (Morrison, 1996). Many of the console game manufacturers have promised some type of Internet connectivity in their new product lines. A number of successful educational computer games are being ported to the Internet.

A recent survey of information systems managers found that 90% of them have access to games at work and that 58% play games at work several times a week (WINNY, 2000). Easy access to the Internet has encouraged software piracy by some people. In addition to lost worker productivity, chief information officers are finding that their corporate servers are being used by some employees as repositories for the exchange of pirated software, which exposes the company to prosecution for distributing unlicensed copies of software. Many people feel that downloading software from the Internet is a major source of computer viruses in the work place. Computer viruses have a reputation for destroying information stored on computers, and a significant amount of time is often required to remove them and have their damage repaired.

There is some evidence that computer game playing can be psychologically addictive, encourages violent behavior among young children, and can cause repetitive stress injuries. There is also some evidence that playing computer games can increase hand–eye coordination, raise a person's self-esteem, and help children learn to deal with the complexities found in real-life situations (WINNY, 2000).

ELEMENTS OF COMPUTER GAMES

Not every game lends itself easily to online play. Multiplayer games in which players take turns moving and single-player games in which players complete against a computer opponent possessing artificial intelligence seem to be the easiest to adapt to Internet play. One reason for choosing to play a game on the Internet is to experience the interaction with game players around the world. The key to developing a good game is to give players interesting challenges to overcome. The solutions to these challenges should be intuitive to the players and should gradually increase in difficulty. This allows a game to be learned as the player's skill increases. A game that produces random results for a given input may not be perceived as an entertaining game.

Genre

The term genre, as used in this chapter, describes a game type. There are many types of computer games. Use of the term genre is a bit controversial among game designers. Some game designers do not believe it is possible to classify every game as belonging to a single genre. Games from the same genre tend to share many attributes (role of the player, use of animation, user interface style, etc.). Although there is no single agreed-upon list of genre names, a list proposed by LaMothe (1999) appears in Table 1.

Story Line

Most games use some sort of story line to enhance the game. The story line may provide background to give players contexts for their actions (e.g., shooters) or it may actually serve as a script (e.g., role-playing games or interactive fiction) to govern the interactive flow of the game (Sawyer et al., 1998). Linear story lines have only one possible path to a successful outcome. Nonlinear story lines allow players to choose different routes to follow during the course of game play and may even have different endings, based on the decisions made by each player. It is usually good to make the player feel that his or her actions during the game affect the game's (and the story's) ultimate outcome.

Puzzles

In stand-alone games, people are more likely to evaluate a game according to the quality of its story line and graphics rather than the cleverness of its puzzles. However, ingenious puzzles are the heart of interactive fiction or role-playing games. Good puzzles can enhance the story line for any game and can assist with character development. Good puzzles can make a good game great, by giving players small challenges to keep them interested in the game while working toward game completion. Bad puzzles can interfere with the player's progress toward goals and can prevent players from immersing themselves in the story.

Several types of puzzles can be found in computer games. Puzzles may require players to push buttons to copy sequences or discover patterns. Players may be required to start or stop sequences of actions within a specified time period. A game may require users to construct a puzzle solution by combining clues or following a trail of evidence like a detective. Sometimes puzzles may take the form of word searches, magic squares, or riddles. These types of puzzles challenge players, and their performance is likely to improve with practice. This keeps players coming back to play again. Puzzles that rely solely on brute force solutions or excessive trial and error tend to discourage players from returning to play again. Puzzles with random results that depend solely on luck and little skill can also make players lose interest in a game.

User Interface

Players expect games to have graphical user interfaces, even if the game does not involve direct manipulation of game objects (Brown, 2000). Most games found on the Internet are designed to run either from inside a Web browser window or from a dedicated game server. Games designed to run inside a browser window allow the game designer to make use of a user platform that is relatively machine independent with regard to displaying game graphics and animation. Games designed to run from a dedicated server often require a platform-dependent client program to be running on each player's machine. The client program would typically handle user input and display-of-game information sent by the server.

Games that involve graphical simulations of real-world actions need to have believable behavior, but designers need to remember that they do not need to build a perfect physical model to create fun games. Most physical actions can be approximated assuming simple Newtonian physics and rigid body elastic collisions. In many games, designers may be able to simplify the detection of collisions between screen objects by focusing on an object's center of mass and surrounding it with a rectangular bounding box. In a two-dimensional (2D) graphical system, object collisions are indicated any time the bounding box surrounding one object enters the area surrounded by the bounding box on another object. This allows the game program to speed up the graphical display of the effects of object interactions. Using discrete-simulation (constant step) techniques can

Table 1 Selected Game Genres

GENRE	EXAMPLE	CHARACTERISTICS
Interactive Fiction	Tomb Raider	May be either text based or involve 3D graphics; games are scripted to show that the solution of each puzzle leads players down predetermined paths until the game ends or the user quits
Fighting Games	WWF Warzone	Usually involves one or two players locked in mortal combat; requires fast computer response time and usually contains realistic graphics to depict violent actions
Puzzle and Card Games	Tetris	Classic turn-based board games, card games, or puzzles played by one or more players using a computer keyboard or mouse for input and 2D graphics displays
Role-Playing Games	Ever Quest	One or more players assume the roles of characters who are allowed to interact with each other and with objects in the environment; artificial intelligence techniques are employed to create computer-controlled nonplayer characters who interact with the human-controlled characters; combat between players and solution of puzzles is often part of the game; often players are required use their own judgment to equip themselves prior to beginning a grand quest
God Games	Roller Coaster Tycoon	Often implemented as turn based, graphical simulations in which the player has a chance to set the parameters for simulation and where computer animation is used to display the impact of these decisions over a compressed time period
Simulations	Microsoft Flight Simulator	Driving- or flight-simulator-type games; often makes use of 3D graphics and requires moderately fast real-time display of the response to player actions
Arcade Games and Shooters	Pac Man	Computer versions of classic arcade games; usually involve 2D graphics and have one or two players shoot at objects or each other with laser-type weapons
Sports Games	Madden NFL	Graphical simulations of individual or team-based sporting events; computer uses artificial intelligence to control most players in team sports; make heavy use of multimedia in user interface designs
Military/Strategy Games	Final Fantasy	Similar to god games or interactive fiction with 2D graphical output; may involve more than one player

also help with synchronization of simultaneous events. In discrete simulation, the effects of simultaneous events in a given time unit are modeled offline and displayed for the user as if they occurred at the same time (LaMothe, 1998).

User interaction devices for Internet games tend to be limited to the mouse and the keyboard, because these represent the base-level devices in the Unix world, where a majority of the World Wide Web servers live. For multiplayer games, text and voice communication between players seems to be an important social aspect of games (Munkki, 1997). This communication may take the form of requests for technical support or it may simply be "trash talking." Although text communication only requires the use of the keyboard, voice communication requires the presence of a microphone and speakers on each player's machine.

Besides graphics and computer animation, multimedia elements that have become part of computer games include sound, music, and video. Sound and music can enhance the story line of a game in the same way they do movies. Typically, making use of most multimedia formats in a browser window requires the user to download and install a plug-in on his or her local machine for each media type used. One of the challenging aspects of working with both video and sound in computer games is synchronizing video stream with audio stream so that they seem to be taking place at the same time. The network bandwidth required for streaming either video or audio is substantial for local area networks. The latency or lag time associated with the Internet further complicates the task of media synchronization (Ng, 1997a,b).

Artificial Intelligence

Developing computer opponents for turn-based games was one of the earliest applications of artificial intelligence (AI) by computer scientists. Many of these early computer opponents make use of a static evaluation function that takes a proposed game state as input and returns a numeric value indicating the value of the position of the state. By applying the static evaluation function to each new game state that is reachable from the current game

Table 2 State Transition Table for Robot Guard

CURRENT GAME STATE	EVENT TRIGGER	RESULTING GAME STATE
Move East	Alarm Silent	Move East
Move East	Path Blocked	Move West
Move East	Alarm Activated	Attack Intruder
Attack Intruder	Alarm Activated	Attack Intruder
Attack Intruder	Alarm Silent	Move East
Move West	Alarm Silent	Move West
Move West	Path Blocked	Move East
Move West	Alarm Activated	Attack Intruder

state, the computer opponent can select the alternative with the best numeric value. The performance of computer opponents can be improved by looking ahead two or more moves before applying the static evaluation function. An algorithm like minimax is used to combine the results of the "look ahead" moves (recognizing the fact that each player will be looking for the best move possible).

Another task confronting opponents in combat or sports games is path finding through the game objects displayed on the screen. Nothing exposes weaknesses faster than when a computer opponent allows itself to take the long way around an obstacle when an obviously shorter path is available. Heuristic search techniques like A* (pronounced A star) can be used to avoid time-consuming trial and error searching for paths between game positions (LaMothe, 1999). Sometimes the positions of objects for a particular game state are known in advance and all routes can be precomputed and stored in the game database. This is not an AI technique, but it definitely is one way to speed up game play.

Pursuit of opponents and running away from opponents can often be handled by heuristic techniques based on the distances between player pieces on the game map. If the distance is getting bigger, a pursuit heuristic might be to alter direction to make the distance smaller. If the distance is getting smaller, an evasion heuristic might call for a computer-controlled object to do whatever is needed to increase the distance from the pursuing piece. This type of heuristic is fairly effective for simple objects such as rocks or missiles. For slightly more intelligent objects (such as birds or spaceships), some random behavior might be added every few moves to make the computer-controlled pieces seem more realistic. Movement patterns might also be used to govern the behavior of patrolling guards or prowling animals (LaMothe, 1999).

For important game elements, such as computer-controlled characters in role-playing games or interactive fiction, more advanced AI techniques might be used. Finite-state machines or production systems are good for modeling event-driven computing systems. Finite-state machines only require knowledge of the current game state and the most recent user input or game event to determine the next state of computation. This requires careful analysis of the system states and their relationships, but it can be a very effective way to make a computer-controlled character appear to react intelligently to each game situation. Finite-state machines can be represented graphically or by using state transition tables. Table 2 contains a state transition table that might be used for a robot guard in a combat game. One way to avoid repetitive behavior is to give a computer-controlled character some means of remembering previous actions and allowing them to vary them from time to time (LaMothe, 1999).

Many people would insist that an intelligent computer system should be capable of learning from its mistakes. Neural networks and genetic algorithms used in computer games allow computer opponents to improve their performance based on past game experiences. Even early checkers-playing programs made use of numerical learning techniques similar to these approaches to improve game play.

ROLE OF THE WORLD WIDE WEB IN COMPUTER GAME PRODUCTION

Game Archives

Using any Internet search engine and typing in the search string "computer games," one is likely to find several million hits. Some sites may simply allow visitors to download games to be installed on their personal computers. Some of these games may be freeware, given away by developers, or they may be shareware, in which developers trust users to send voluntary donations to offset development costs. These games may be free demonstration versions of games that can purchased either online or in local retail stores. Some of these sites also provide support resources for both game players and game developers. Some of these sites may provide visitors with opportunities to play games online without needing to install them on their own personal computers.

User Support Communities

Many companies make Web-based user support resources available to registered owners of their products. Game companies are no exception. Online registration of products is an option commonly included during the installation of many computer games. Game owners are often put on electronic company mailing lists and are promised notification regarding software bug fixes (patches) and new product releases. Software companies may also provide access to technical support to assist users with game installation problems or help them download software patches and missing or updated device drivers. Some games may use the Internet to automatically download and install new versions of game components on the user's machine. Many game companies may also provide online bulletin boards to allow users to share game-playing tips with other players or contact technical support representatives for instant responses to questions. Some multiplayer games make use of the Internet to provide communication between players during actual game play. Game manufacturers are finding that adding online communications components to their games is becoming important for success in the market place (Bernier, 2000). Many games are so popular that the users themselves set up Web sites to share reviews of new games and provide game cheats that show scripted solutions for beating complicated game puzzles. Simply typing the name of a popular game (e.g., Tomb Raider) into a search engine is likely to yield thousands of hits.

Developer Resources

Most game developers insist that knowing their potential users is the key to the success of any game. Communication between geographically widely distributed developer and game users can be easily supported by the World Wide Web. Online entertainment requires interactive participation by its users. Games must evolve, and developers need to filter user feedback into the design of their improved game. The process is very much like user-centered software design. The production value realized by developers of online games and Internet development processes involves game design quality improvement, evolution of multiplayer game technologies, creation of communications mechanisms, and increasing media density or fidelity in the game (Palumbo, 1998). Some companies underestimated expected customer support demands during the launch of a game and found that games sales exceeded the established customer support system. Companies can reduce or eliminate many support problems by having the foresight to create a Web site dedicated to late-breaking game news and user communication (Firor, 2002).

Beta testing is a type of testing where the preliminary release of a software product is made available to a subset of its intended users, who then make use of the product in their own personal computing environments. The beta testers ideally report problems and make suggestions to the software developers prior to product launch. This early user feedback has proved to be an effective means of reducing the number of bugs found after product launch. The use of the Internet can facilitate the distribution of the beta versions of the software, as well as the collection of user reactions to it.

There are several Web sites dedicated to communication between game developers themselves (e.g., Gamasutra). These Web sites allow experts to share technical tips among themselves. Many times, developers will share with the Web audience the lessons learned while developing successful (or unsuccessful) products. The commercial game development community is widely scattered and very volatile. Many companies make use of contract employees on a project-by-project basis. Some Web sites allow visitors to post employment ads and resumes to enable potential employers and employees to find one another as new game development projects are undertaken.

The World Wide Web can also provide sources for software development tools and training for developers. The training may take the form of short, free tutorials or full, semester-long, online courses. These courses may be offered by universities or by commercial companies (e.g., http://www.gameinstitute.com). Java is a popular development language for networked games and is distributed free of charge over the Internet. There are companies (e.g., Mpath) that provide dedicated game service over the Internet (Morrison, 1996). This service is usually provided on a pay-to-play basis to support multiplayer game play. Some companies may license the use of their online game engines or servers to other developers to improve their revenue stream.

Online Games

Online games can be delivered via the Internet or through dedicated game servers on wide area networks with locally installed client software on a player's personal computer. Single-player games are often accessed directly from a Web site and a player communicates with a game from inside his or her Web browser. Game play for single-user games using a Web browser interface is much like game play for single-user games designed for stand-alone PCs, because the game logic may in fact be running on the local PC and not on the server. To display multimedia content, a user is usually required to have the appropriate media plug-ins installed on his or her personal computer. Standard Web technologies (HTML, Javascript, common gateway interface scripts, or Java) might be used to deliver the game output to the Web browser and to process the player's input. To communicate with a dedicated game server, the client program may be implemented using any programming language capable of supporting socket programming. Socket programming is a technique for allowing the exchange of information between two programs running on the same computer network. Browser interfaces to online games have the advantage of allowing for multiplatform game implementation with very little additional effort on the part of the developers.

When people think of online games, they often envision multiplayer games, with several players distributed at distant locations on a local area network or globally connected by the World Wide Web. There are two big problems confronting designers of multiplayer games. The first is difficulty in synchronizing game-state information when multiple instances of the game client software are running simultaneously. The second difficulty is related to the first in that the latency (or response time lags) that occurs as the user inputs and the resulting game consequences need to be shared with players using client software running on independent machines connected by a relatively slow network (the Internet). The types of multiplayer games that are implemented most easily on the Internet are turn-based games or event-based games, because in these, the lags in response times inherent in Internet applications are most easily hidden (Morrison, 1996).

Interactivity and social diversity are two of the most powerful elements of the Internet (Palumbo, 1998). The Internet gives designers a chance to build community followings for their games. This was not done easily before the advent of online games. For some games (e.g., Quake or Diablo), players created their own communities (Min, 2000). Some games (e.g., Fireteam) are based on players recruiting teams to play against similarly recruited teams across the Internet. Many teams stay together for long periods of time and enjoy their inclusion in a cohesive group. Teams may create their own Web sites to coordinate their efforts.

The World Wide Web provides a number of resources and opportunities to both game developers and game players during all phases of the software development cycle. Software can be delivered directly to game players. Games can be played over the Internet. Developers can communicate with each other and their potential user groups during the game development process. Users can make use of the Internet to form their own support groups.

DESIGNING COMPUTER GAMES

Brainstorming

Sometimes the toughest part of building a game is coming up with an idea. Game ideas may come from popular books, movies, sport contests, or board games. Designers often choose one of three possible starting points: genre, technology, or story line (Rouse, 2001). Choosing a genre, many times, determines the type of technology that will be used in the game. The technology used to implement a game may also determine the type of genre that can be used most effectively for a particular game. Often designers choose game technologies because they already exist or because the particular technologies are what the target audience for the game expects.

As with all software engineering endeavors, in game design it is very important to understand the desires of the prospective users. Failing to appreciate the cultural biases of potential players may result in designing a game that sells well in one country and poorly in another. Poor-quality artwork and mediocre animation sequences can cause a game to be rejected by potential purchasers. Poor game play, often caused by the use of inconsistent and unintuitive user interfaces, can also cause a game to fail in the market place. In commercial game sales, the first game of its kind to reach the marketplace and deliver all its promises is usually the big winner financially. Many failed games have been based on obsolete technology or were delivered too late to the marketplace to beat the competition.

Documents

Modern game designers are finding that it is important to treat computer game design in the same way other software engineers approach projects involving large numbers of people and significant investments of time. Game developers are likely to use evolutionary software process models (e.g., rapid prototyping) to mange their development risks and reduce their project completion times. This means that the game may be built incrementally, starting with a small version of the game and adding features to each increment. Each increment will have defined requirements, design goals, and project and testing goals. This requires that formal project documents be developed early in the project and revised often.

The first project document developed for a game is sometimes called a game treatment. The purpose of the game treatment is to sell the game idea to potential investors or to a development team considering which project to undertake next (Lewindowski, 1999). The most important point to convey is what it is like to actually play the game. The player's role and allowed actions in the game should be described, as well as the behavior of computer-controlled characters and objects. The intended audience for the game also needs to be identified. The game story line should be summarized, along with the major and minor goals to be achieved by players. The proposed game-delivery platforms (hardware and software) should be described. The technology (production tools) required to implement the game should be described as well. The introduction of multimedia elements, such as 3D graphics or CD-quality music, increases the complexity of the development task, as do plans to make use of client/server technology to allow multiplayer, networked game play. It is very important to ensure that the proposed game-playing environment be readily accessible to the intended game audience.

The game treatment must be written in sufficient detail to allow potential investors or collaborators to decide whether or not to move forward with this project. The game treatment plays a role similar to a feasibility document in a software engineering project. Many games are developed using a rapid prototyping process. This means that the game treatment will be used as the initial-requirements definition for the game. Prior to accepting the game as defined in the game treatment, the developers and investors discuss the risks involved. Risks are threats to successful completion of the software development project. Several types of risks need to be considered. Business risks are concerned with the marketing and profitability of a product. Technical risks are concerned with the implementation of the software product. Project risks are concerned with threats to the project budget and development schedule. The game treatment may be modified or rejected based on consideration of each of these risks.

All software projects need to have written design documents to work from before any programming work begins, and games are no exception. The design document created for evolutionary process models used for software projects such as computer game programming or Web site design will most likely to be a dynamic living document that changes with the project as implementation proceeds. So it is not necessary to specify every detail of every game algorithm or data structure before programming activities begin. However, to avoid unnecessary rework of software products, a clear understanding of the game's overall direction must be held by all development team members as they begin working on the project.

For games that have more than a single action screen, a storyboard is often developed. Animators and movie writers use storyboards to provide a visual layout of the key scenes that will be shown to the audience. Game designers make use of storyboards to show a sequence of 6–12 key scenes from the game, along with notes describing the multimedia elements and game actions that will accompany each scene.

A flowchart may be used to show how the linear sequence of screens displayed in the storyboard may be interrupted by user actions or computer-generated game events during the course of game play. Users find games to be more interesting if they are able to vary the plot sequence of the game each time they play. The flowchart defines those places in the game that plot may be varied without destroying the story line completely.

If games have complex characters, a character bible is created, containing every observable fact and describing each character the player will interact with during the game. Sometimes bibliographies are written for each character so that as a game unfolds, a game designer will not make mistakes when displaying these characters and their actions to game players. This helps the user become more immersed in the game as fantasy. Memorable characters can lead to game sequels or can become marketing icons in their own right.

A game script, which is not likely as books or movies to be linear in nature, would be difficult to describe without the tools mentioned previously. Two important elements of the game script include level designs and puzzle details. Game levels are added to give players a sense of goals completion on their way to finishing the game. Level design has become an important specialization in game development. From an economic standpoint, adding levels to an existing game is far less costly than designing a new game.

User interface design takes on a high level of importance in game development projects. One representation of the user interface design consists of the screen displays seen by the user at any given point of the game. A second dimension of the user interface design is creating a flow diagram containing enough detail so that all user actions that trigger game events and their consequences are included. Software engineers call this type of diagram a state transition diagram. Often the user interface is developed as a working prototype for the game. Web page development tools can be particularly helpful in this regard. It is often easy to link Web pages containing the multimedia elements experienced by the user in the nonlinear order they would experience during the course of the game. Animation sequences can be added as they are created. If the game is to be implemented as an Internet game, the amount of throwaway code is minimized.

Game development projects often involve many developers independently producing code simultaneously for different parts of the game system. It is essential to develop a project management plan early in the design cycle to avoid production delays and unnecessary reworking. The project plan needs to clearly identify the people required to complete the game project and the skills that should be possessed by each team member. The hardware, software tools, and consumable materials required for development need to be identified so they can be utilized in a timely manner during game implementation. Once the people and resources required to complete the project are identified, it should be possible to begin to build a budget and a schedule for the project. Software engineers have techniques at their disposal that allow them to accurately predict the cost and duration of software development projects. It is important to utilize these techniques when building a project plan for a game, because games are costly to develop and missing delivery deadlines can be detrimental to a game's success. For example, failing to complete a game tied to a popular movie in time for it to be sold during the Christmas shopping season can make the game an economic failure. Wise game developers will also take the time to analyze their development risks and identify ways of preventing these risks from becoming threats to the completion of the game project. Some of these risks include marketing concerns, technology availability, programming skills, funding, and resource issues.

CREATING GAME CONTENT

Platforms

Online computer games (either single or multiplayer) by definition rely on network communication in one form or another. There are two common models for network communication, peer to peer and client/server. Hybrid models combining the benefits of both also exist (Ng, 1997b). With the client/server communication model, the game logic and database operations tend to be executed on the server machine. The player's machine often only needs to be running a client program that is capable of displaying the current game state based on data sent by the server and of collecting player input to send to the server as needed. Modem-to-modem communication might be thought of as a special case of peer-to-peer networking. In peer-to-peer networking, each player's machine needs to be running its own copy of a game program that is capable of handling both the game logic and the input/output operations required to play the game. Some hybrid models make use of peer-to-peer networking communication, but one of the machines is given the additional task of routing messages among all the game players.

The target delivery platform has great impact on the type of game content that can reasonably be included in a particular game. A game designed for a single player using the client/server communications model may be limited only by the speed of the network connection to the game server. Consequently, for a single user online, a game designer may need only to decide which operating systems and which processor types he or she is interested in supporting. The Windows- and Unix-type operating systems are the most popular target platforms in terms of numbers of users in 2002. Windows operating systems are designed to run on machines that make use of Intel-type processors. Unix operating systems are available for almost every personal computer processor and have nearly identical resources in each implementation. Each operating system tends to support graphics programming in different ways, because the graphic libraries are based on different standards. Each program will need to be recompiled to run on each type of processor, because each processor requires programs to be written using its own unique machine language. Similarly, each operating system supports different types of multimedia. So even if a designer is relying on browser technology that tends to be machine independent, he or she still needs to be aware of which media types will be supported by an operating system and which will not.

The choice of server technology can also be important. Dedicated game servers may give game designers the most control over the delivery of the game content.

They may also make it easier to control access to a game and charge users for its use. If a dedicated game server is not accessible from the Internet, it is harder for users to access without incurring long-distance phone charges. Once game players are connected via the Internet to multiple game servers, the problems of communication latency will begin to appear. This can make it tough to synchronize the display of action sequences that appear to all users at the same time, or even to process user input in a timely manner.

Compilers and Game Engines

Most multimedia games are not developed from scratch. Developers usually rely on a development library that contains tools needed to create graphical user interfaces and manipulate multimedia game objects. Most games written for the PCs are written using Microsoft's Visual C++ programming language and make use of Microsoft's DirectX game-development library. One of the components of DirectX (called DirectPlay) supports peer-to-peer network communication over the Internet. DirectX is not a complete game-development library. To make game programming more efficient, many developers create game engines that contain complete subsystems (like game AI or sprite animation and collision detection or level design templates) out of development library components provided in such products as DirectX. The Microsoft .NET software development tools may make it easier to unify network and desktop computing resources for users of Microsoft products.

DirectX is not a platform or machine-independent software product. Because Unix is such a popular Web-development platform, there is a great deal of interest in using open-systems software products (such as Linux, OpenGL, and gcc) for game development. Indeed, many console games are written using these technologies, so there is a great deal of interest in adding networking to existing console games.

The use of software libraries for game physics and 3D models is a new area of interest for game developers. The use of libraries has not been popular among game developers in the past, because it was felt that their use would reduce the entertainment value of games by making them appear to be shallow copies of one another. Three-dimensional-game architectures, DVD playback systems, and sound systems are changing so rapidly these days that developers cannot afford to develop game software completely from scratch within the desired eight-month development cycle. Consequently, game developers are beginning to view the use of libraries more favorably.

Technologies used for Web-based computing have also been used to implement online games. Web-based computing depends on each computer having a fairly standard Web browser as its primary user interface. This tends to have the advantage of allowing games developed using one operating system or hardware platform to be used on others. Java has been used to implement both single and multiplayer multimedia action games. Javascript and Flash can be used to implement single-player games that have modest graphics and user input requirements. C++ and Perl are popular languages for implementing CGI programming. CGI programming has the advantage of allowing programs to run on the server machine and requiring the client machine to have only a Web browser running on it. This, of course, makes it possible for players to access game programs from several different types of machines.

Multimedia Tools

Computer games typically include three types of multimedia content: visual (pictures, animation, and video), aural (recorded or generated sounds and music), and textual. Multimedia game content is usually created using specialized tools (scanners, digitizers, synthesizers, etc.) as part of the development process. Artwork is often created using computer-based drawing tools or photographs to allow more realistic images to be included in the game display in a shorter time period. Real-time animation sequences may take too long for the computer to generate during game play, and it may be better to display them as prerecorded videos. Designers often need to choose between the level of player control and the amount of time it takes to create the animation dynamically and display it during the game. With sound the opposite may be true; it is sometimes faster to synthesize speech or music rather than record it (even though the fidelity may not be as good).

When a game is to be played over the Internet, designers may find themselves with fewer options for art, video, and sounds. Web browsers may only support the display of .gif or .jpg graphics files and only .au files for sounds. Requiring players to have multimedia plug-ins (e.g., Quicktime video or Microsoft's MediaPlayer) installed on their local machines can mitigate some of these media format limitations.

In many multimedia-computing applications, elements such as sound and music are digitized and stored in separate files. It can be challenging to combine and synchronize the two media streams so that the viewer sees a coherent presentation. Coordinating real-time multimedia streams is not easy and can take a lot of bandwidth. Some of the challenges are related to communications and processing delays that become even more obvious when the Internet is involved. Some techniques for synchronization involve interleaving the data streams, and some may involve embedding time codes to be used by the media player to combine the streams at the appropriate time. The use of technologies such as Quicktime and Realmedia can make the developer's task much simpler.

MULTIPLAYER GAMES AND NETWORKING

There are several attractions to using the Internet to play games. The Internet can make games available to large numbers of users on a round-the-clock basis. The Internet can provide users with the opportunity to build global support communities for particular games and allow direct marketing of game-related items to consumers. (Ng, 1997a). It is harder to add Internet capability to a game after it is written than during the design phase

(Lincroft, 1990). There are several networking issues that must be taken into a count when a multiplayer game is being designed.

Session-Oriented Multiplayer Games

Not every multiplayer game is amenable to delivery on the Internet. In many ways, turn-based games are easier to deploy than fighting games because of inherent latency problems with Internet communications. Even for turn-based games, designers need to decide whether new players can drop in at any time or whether they can join only at fixed points in the game. In session-oriented multiplayer games, new players can join the game only when a new round or a new level of play begins. Sometimes this decision is related to whether the online game is designed to be short in duration or whether it will be a persistent game, which never ends and in which players drop in and out. Designers need to consider what will happen to the lost entities created as players or servers drop out of games. Should an AI take over for a lost entity? Can a backup server replace a lost network connection? In games that last longer than 3 hr, it is important for users to be able to save incomplete games or characters they developed (Ng, 1997a). Whether a game should be saved on the server or on a client machine may depend on the size of the game-state information structure.

Synchronization Problems

With networked game communications, the key problem is synchronizing the sharing of game-state information among several users in a timely manner. One solution is to use state synchronization, which requires each player's game instance to send a copy of its current state information to each of the other game instances. This requires a lot of bandwidth and takes a lot of time in the peer-to-peer model. A second solution is to use input synchronization. This requires each game instance to communicate its player's input to each of the other game instances. This reduces the amount of bandwidth required. If a game includes random events that do not depend on user inputs, the input synchronization system can fall apart. When random events are allowed in game play, a combination of these two synchronization techniques may need to be used. This means that a game instance may need to share information on its user's input as well as on changes in its game-state information with each other game instance following a random event (Morrison, 1996).

Information-Sharing Concerns

The problem of sharing information in a timely manner among widely scattered Internet users has two other dimensions: bandwidth and latency. Bandwidth issues can be dealt with after a game is written by using such techniques as data compression or by limiting the number of game players. Latency must be addressed during the design. If peer-to-peer networking is used for a multiplayer game, it is sometimes wise to make one of the peer machines take on the added responsibility of relaying messages to all game players, much like a server would do in the client/server model. Intelligent routing schemes might also need to be used. It is doesn't matter whether each player in a multiplayer game has exactly the same view of the game state or not. The only thing that matters is that players do not miss seeing important events (e.g., when one player shoots at another). In modern operating systems, it is possible to separate game communication from graphics displays (renderings) by using separate threads for communication and rendering (Ng, 1997b).

Dead reckoning is used to render game objects based on their last known information (e.g., heading and velocity). This reduces the amount of information that must be shared among players. This also helps deal with unreliable network connections. If a connection is lost, dead reckoning can be used to predict the game-state changes until the connection is restored. This assumes that no major changes take place during the time the connection is missing. Sometimes it is wise to build in delays to the time required to process each user input, so that when delays occur, the users do not notice them as much (Munkki, 1997). In the real world, humans cannot see objects beyond the horizon or in the next room. It is not necessary to render these objects until the player could reasonably expect to see them. Sometimes the client needs to keep two copies of the game state. One copy contains changes based on the actual actions taken by all players, and one copy contains changes made by dead reckoning. When information is received indicating that the copies are no longer acceptably in synch, the state discrepancies need to be fixed by the client when it is convenient (Lincroft, 1999). Sometimes these discrepancies can be hidden from the player by scheduling game-state changes two or three virtual time periods in the game's future story line and waiting to render them on the screen until their correctness is verified.

Network Protocol Trade-offs

Sometimes designers need to consider carefully which networking protocols provide the best means of communicating among game clients. Much of the time, Internet communication is done using the TCP protocol, which is reliable, stream oriented, connection persistent, and slow. Part of the reason it is slow is that although data packets are always delivered, the packets must be assembled before the message can be delivered. The UDP/IP protocol is unreliable and has small packets and very low overhead. The UDP/IP protocol can be much faster than TCP, but the programmer has the responsibility of checking information to see that it is received in the proper order (Bernier, 2000). Simply using faster network connections (e.g., Internet2) does not eliminate the need to consider these trade-offs between TCP and UDP/IP.

CONCLUSION

Two computing applications that involve extremely large numbers of users are computer games and Web-based applications. It is only natural to want to consider using the Internet to support the development and delivery of computer games. As high-speed Internet connections (e.g., DSL, cable modems, Internet2) become more widely available and affordable to home users, interest in developing online games will continue to increase. It is important to remember that the computer game industry, which is only 30 years old, has already grown to a multibillion-dollar-a-year industry. Its growth was accelerated by the advent of the affordable home computer. The Internet is fewer than 10 years old and has grown at an even faster rate. As game developers have better and faster network technologies (e.g., Internet2) to exploit, the quality of online multiplayer games is likely become closer to that of today's console games. It is important to note that new techniques for event synchronization will also be required, because part of the problem is displaying games simultaneously to several users (not just bandwidth and latency issues).

GLOSSARY

Bandwidth The potential amount of information able to flow simultaneously through a computer network.

CGI Common gateway interface; a protocol for allowing remote users to run computer programs on a machine housing a Web server.

Client/Server Computing A type of networked computing in which one computer (the client) requests data and/or services from another computer (the server).

Demos An executable subset of a larger program, intended to give users an idea of what it is like to use the full product.

DirectX Microsoft game-development-tool library.

First-Person Game A game that displays the screen view as seen through the eyes of the game's main character.

Flash A popular Internet multimedia animation language distributed by Macromedia; requires a browser plug-in to be installed on a client machine.

Freeware Software distributed free of charge by its author.

Game Engine The core code that handles the basic functionality of a game and that may be customized to deliver several variations of a particular game genre.

gcc A free C/C++ complier distributed by the Open Systems Software Group.

Javascript An Internet scripting language supported by most popular Web browsers.

Latency The delay that occurs between the time a network signal is sent and the time the signal is received and acted on.

Linux Freeware version of the Unix operating system; versions are available for most PC hardware platforms.

NPC A nonplayer character controlled by a computer program, usually in a role-playing game.

OpenGL Standard graphics library developed by Silicon Graphic.

Rendering The task of drawing an on screen graphical representation of a data object.

RPG Role-playing game; players in this type of game assume the roles of characters in a story and are allowed to interact with other players, nonplayer characters, and game objects.

Shareware Software distributed by its author on a trial basis with the expectation that users will send payment if they find it worthwhile.

Sprite An animated graphical game object typically capable of independent motion on the game display screen.

Third-Person Game A game that displays the main character interacting with the game environment as if the game player were watching a movie from the perspective of an onlooker.

Thread A separately executing process in a multiprogramming execution environment.

Unix A popular computer operating system often used by Web programmers and Internet server administrators.

CROSS REFERENCES

See *C/C++; Client/Server Computing; Common Gateway Interface (CGI) Scripts; Interactive Multimedia on the Web; Linux Operating System; Multimedia; Unix Operating System; Virtual Reality on the Internet: Collaborative Virtual Reality.*

REFERENCES

Aronson, J. (2000). *Using groupings for networked gaming*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/20000621/aronson_01.htm

Bernier, Y. (2000). *Half-Life and Team-Fortress networking: Closing the loop on scalable network gaming backend services*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/20000511/bernier_pfv.htm

Bettner, P., & Terrano, M. (2001). *GCD 2001: 1500 archers on a 28.8: Network programming in Age of Empires and beyond*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/200110322/terrano_01.htm

Brown, D. (2000). *From Underground to World Renowned: Following up on the First Annual Independent Games Festival finalists*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/20001023/brown_pvf.htm

Firor, M. (2002). *Postmortem: Mythic's Dark Age of Camelot*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/20010213/firor_pfv.htm

Holder, W., and Bell, D. (1998). *Java game programming for dummies*. New York: Hungry Minds Inc.

LaMothe, A. (1999). *Tricks of the Windows game programming gurus*. Indianapolis, IN: Sams.

LaMothe, A. (1998). *Windows game programming for dummies*. New York: IDG Books.

Lewindowski, J. (1999). *Developer's guide to computer game design*. Plano, TX: Wordware Publishing Inc.

Lincroft, P. (1999). *Designing fast-action games for the Internet*. Retrieved April 26, 2002, from http://www.gamasutra. com/features/19990903/lincroft_01.htm

Min, A. (2000). *Postmortem: Multitude's Fireteam*. Retrieved April 26, 2002, from http://www. gamasutra.com/ features/20000105/fireteam_01.htm

Morrison, M. (1996). *Teach yourself Internet game programming with Java in 21 days*. Retrieved April 26, 2002, from http://www.njnet.edu.cn/info/ebook/java/javagame

Munkki, J. (1997). *Design and implementation of networked games*. Retrieved April 26, 2002, from http://www. hut.fi/~jmunkki/netgeames/

Ng, Y. (1997a). *Designing fast-action games for the Internet*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/19970905/ng_01.htm

Ng, Y. (1997b). *Internet game design*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/19970801/ng.htm

Palumbo, P. (1998). *People vs. pictures: Why online games take the focus off of production values*. Retrieved April 26, 2002, from http://www.gamasutra.com/features/19980130/process.htm

Rouse, R. (2001). *Game design: Theory and practice*. Plano, TX: Wordware Publishing Inc.

Sawyer, B., Dunne, A., & Berg, T. (1998). *Game developers marketplace*. Scottsdale, AZ: Coriolis Group Books.

WINNY (2000). *Game—The psychological and physical impact*. Retrieved April 26, 2002, from http://multimedia.design.curtin.edu/cache/g/0007/

Gender and Internet Usage

Ruby Roy Dholakia, *University of Rhode Island*
Nikhilesh Dholakia, *University of Rhode Island*
Nir Kshetri, *University of North Carolina*

INTRODUCTION

Modern information and communications technologies (ICTs) such as the Internet arguably have the potential to offer greater benefits to women than men (Carter & Grieco, 2000; United Nations Conference on Trade and Development [UNCTAD], 2002). For instance, new ICT applications such as changing a health service appointment electronically can "return" the time lost by single mothers under conventional arrangements that require travel to a clinic for such tasks (Carter & Grieco, 1998). In reality, however, a significant gender bias toward men exists in the adoption of modern ICTs, including the Internet. For instance, the user survey conducted by Graphic, Visualization, & Usability Center (GVU) in 1998 found that 66.4% of the Internet users in the world were men and 33.6% were women (GVU Center, 1998). In 2000, the bias still persisted in most parts of the world, with the exception of the United States and Canada. Understanding of the Internet's evolution over time and space, therefore, is incomplete unless we understand the role and influence of gender in the processes of Internet adoption and usage.

Most available statistics on Internet usage tend to use the terms *gender* and *sex* interchangeably, even though sex is a biological characteristic and gender is the "social constructed roles ascribed to males and females. These roles, which are learned, change over time and vary widely within and between cultures" (United Nations definition, World Bank, 2002). Although the current reporting of data on the Internet and gender does not adequately capture the cultural and changing differences in gender roles, the data do provide a view of some of the differences in ways men and women use the Internet. They underscore the importance of "gender and Internet usage" as a topic and the need for an explanation of phenomena such as the following (based on a variety of sources, including Brisco, 2002; Hafkin, 2001; UNCTAD, 2002):

- In the United States and Canada, proportions of Internet users seem to parallel gender proportions in the population, but in other advanced economies, such as Sweden and Japan, men still account for a larger portion of the Internet-using population.
- In some countries, women Internet users have increased dramatically. For example, between 1999 and 2000, the proportion of women Net users jumped from 33% to 42% in Mexico and from 25% to 43% in Brazil.
- While women represent nearly 50% of the labor force in Asia and own more than one third of small and medium-sized businesses in the region, in 2000 they accounted for only 22% of Internet users on average.
- In Africa, women's participation in Internet usage continues to be low, ranging from 12% in Senegal to 38% in Zambia.

In this chapter, we address the following questions:

- What is the nature of gender bias in Internet adoption and usage?
- Is the bias changing over time and space?
- How can the bias be profiled and expressed effectively?
- What factors explain the variation of the bias within and across countries?
- Why do men and women tend to use the Internet for different purposes?

We use country-level Internet usage data obtained from various sources to address these questions. These multiple sources have used a variety of methods to collect, analyze, and represent data and such heterogeneity in data imposes several limitations on their use and interpretation. Lacking a single, comprehensive study based on a consistent approach to the collection of data globally, we have relied on available heterogeneous data to examine the gender relationships. The remainder of the chapter (a) examines the nature of gender bias in the adoption of the Internet, (b) discusses various possible sources of the bias, and (c) provides some conclusions. In this chapter, we use the terms gender and sex interchangeably, to reflect the availability of data.

Table 1 Gender Distribution of Internet Users Worldwide

REGION/COUNTRY	INTERNET USERS M:F (RATIO 2000)	GDP PER CAPITAL ($, 2001)	SOURCE
Western Europe	58:42	22,311	Jupiter MMXI (2002)
Denmark	64:36	31,090	Pastore (2001)
Germany	60:40	23,700	Stewart (2002)**
Ireland	60:40	23,060	Jupiter MMXI (2002)
Italy	70:30	24,340	Jupiter MMXI (2002)
Spain	56:44	20,150	Pastore (2001)
Sweden	61:39	25,400	Jupiter MMXI (2002)
United Kingdom	57:43	24,460	Jupiter MMXI (2002)
USA	50:50	34,870	Nua Internet Surveys (2000)
Latin America	60:40	4,896	E-Marketer (2001b)
Argentina	57:43	6,960	E-Marketer (2001b)
Brazil	57:43	3,060	E-Marketer (2001b)
Mexico	58:42	5,540	E-Marketer (2001b)
Asia	78:22	1,925	Hafkins &Tagger (2001)
China	70:30	890	China Internet Network Information Center (2001)
India	77:23	460	E-Marketer (2001a)
Japan	67:33	35,990	Hafkins & Tagger (2001)
Korea	53:47	9,400	Deok-hyun (2001)
Singapore	57:43	24,740	Pastore (2001)
Oceania	54:46	19,080	Pastore (2001)
Australia	55:45	25,780	Pastore (2001)
New Zealand	52:48	12,380	Pastore (2001)
Middle East	94:6	4,089	Hafkins & Tagger (2001)

Note. GDP = gross domestic product. GDP data are from United Nations Development Program (2000), World Bank (2003), and authors' calculation.
**Data for January 2002.

NATURE OF GENDER BIAS IN INTERNET ADOPTION

Global Variation

Precise data on Internet usage by gender are extremely difficult to obtain, especially from developing countries, and the available data lack reliability and comparability (UNCTAD, 2002). Nonetheless, statistics on Internet access and use across countries (Table 1) reveal gender as one of the most important factors influencing Internet adoption and usage. The degree of gender bias in the adoption of the Internet varies widely across the world. As Table 1 indicates, among Internet users, the male–female ratio ranges from 94:6 in the Middle East to 78:22 in Asia, 75:25 in Western Europe, 62:38 in Latin America, and finally 50:50 in United States. There is great variability even within a region. In Western Europe, for instance, the male–female ratio in Internet use varies from 70:30 in Italy to 55:45 in Germany.

In general, the proportion of female Internet users tends to be higher in countries with higher per capita incomes. For instance, countries in Western Europe and Oceania had higher proportions of female Internet users than lower income countries in Asia and Latin America. In the developing countries, women usually account for much smaller proportion of the total Internet population (Table 1). Adoption of any technology, including the Internet, requires investments in fixed capital as well as recurring variable costs. A lower income level implies that users have to invest a higher proportion of their income in acquiring and using a technology. (In developing countries, Internet availability in public venues such as Internet cafes obviates the need for fixed investments by individuals. There are, however, serious gender barriers that prevent women from going to such public locations and using the Internet. Therefore, for women, lack of funds to buy a computer continues to be a barrier.) The total perceived sacrifice in adopting the technology thus is higher for individuals with lower income levels. Women's per capita GDP (gross domestic product) worldwide is lower than that of men (see Table 4, later in this chapter) and women's per capita GDP as a proportion of men's per capita GDP varies widely across the world. The income disparity across sexes thus explains some of the gender disparities and global variations in Internet adoption and usage.

The proportion of female users is also higher among countries with longer history of Internet usage. For instance, in the early 2002, female Internet users were 46% in Sweden, 42% in Britain, and 39% in Germany and France. This proportion drops to 31% in Italy and 29% in Spain ("Men Still Dominate," 2002a). Similarly, Brazil and Mexico—early-adopting Latin American nations—

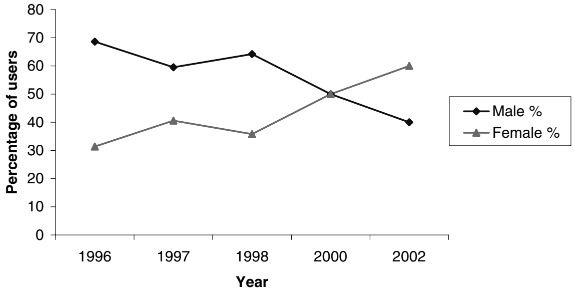


Figure 1: Gender distribution of Internet users in USA.

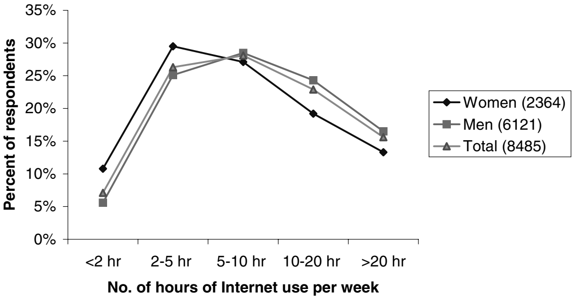


Figure 2: Gender and Internet usage in Quebec, Canada (1997).

had higher proportion of female Internet users than other Latin American nations.

The gender gap in Internet adoption has been diminishing over time. Some statistics suggest that the gender gap may even be reversed in the United States and Canada but not in other countries. As presented in Figure 1, the gender bias in Internet adoption in the United States disappeared in 2000 and also started showing a reverse trend after that—with more women than men online.

Width and Depth of Internet Adoption

Despite some evidence of the narrowing and even reversing gender bias (Figure 1; Rainie, 2002), further analysis indicates that the bias still persists in terms of the width and depth of adoption—even in countries where the gender gap has seemingly disappeared. The concepts of width and the depth of technology adoption help extend the analysis of gender bias in Internet adoption. Gatignon and Robertson (1991) defined the width of adoption as the "number of people within the adoption unit who use the product, or the number of different uses of the product" and the depth as "the amount of usage or the purchase of related products" (p. 468). Thus, higher width of Internet usage is associated with greater number of individuals within a household using the Internet, as well as greater number of different uses of Internet by a specific member of that household. For instance, a possible measure of the width of Internet adoption may be the number of activities or applications (e.g., education, communication, information search, entertainment, etc.) for which Internet is used.

For a multifunctional technology such as the Internet, "depth" can have at least two measures: one related to the usage of the technology for performing a particular function (functional depth) and the other related to the total usage of the technology (overall depth). In case of Internet, for instance, a possible measure of functional depth of "Internet adoption for shopping" could be the number of times per month an individual uses the Internet for shopping or the total time per month spent shopping online. Total or overall depth of adoption of the Internet, on the other hand, could be measured by the total time spent using the Internet per month.

It is not surprising to find that men still outpace women on the Internet in terms of various usage measures ("Demographics: European Women Surf," 2002), and this bias is evident even in countries where the overall ratio of Internet users has become gender-neutral. In 1997, a study conducted in Quebec, Canada, indicated that men were heavier users of the Internet than women (Figure 2). The study indicated that men spent 16% more time online than women, viewed more pages, and went online more often. As a result, Internet adoption tends to be much deeper for men than for women. This was also evident in a December 2000 U.S. study indicating that men went online 20 times, spent 10 hours and 24 minutes online, and viewed 760 pages per month. The corresponding figures for women were 18 sessions, 8 hours and 56 minutes, and 580 pages (Nua Internet Surveys, 2001). Similarly, in Asia male Internet users in 2001 spent 14.5 hours per month online compared with 12 hours spent by women (Nua Internet Surveys, 2002b; Silicon.com, 2001).

The functional depth of Internet adoption depends on the type of usage. For instance, compared with Spanish men, Spanish women tend to favor instant messaging sites and file sharing. In Asia, the functional depth of Internet adoption for visiting sports sites is higher for men than women. In Hong Kong, for instance, 10% more men visit soccer-related sites than women (Nua Internet Surveys, 2002a).

Visits to various Web sites also show significant gender differences in the United States and Canada, as well as rest of the world (Table 2). Only in the case of Yahoo! did more U.S. and Canadian women than men visit a site; the differences are more pronounced in the rest of the world ("Demographics: European Women Surf," 2002b).

A survey conducted in 2001 indicated that although 100% of the female respondents using a computer in Rhode Island used e-mail applications, there were variations in usage locations (Dholakia, Dholakia, Mundorf, & Xiao, 2002). Men, more than women, used e-mail both at home and work; women tended to e-mail mostly at work. For online shopping, the difference was even more pronounced, with more men reporting being Internet shoppers and conducting shopping transactions from home as well as work (Table 3). When we examined shopping sites such as eBay and Amazon.com, men used these sites more than women in all parts of the world, including the United States and Canada. The only exception, Amazon.com, was used by more women than men, but only in the United States and Canada (see Table 2).

Other studies report similar gender bias within the United States. Unilever (2001) suggested that online shoppers are more likely to be men than women; men dominate

Table 2 Gender Difference in Usage of Leading Web Sites (%)

	UNITED STATES AND CANADA		EUROPE		ASIA PACIFIC		EMERGING MARKETS	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
Yahoo	49	51	64	36	57	43	64	36
Microsoft	53	47	69	31	61	39	68	32
Google	52	48	67	33	62	38	70	30
eBay	53	47	70	30	63	37	NA	NA
Amazon	44	56	63	37	61	39	62	38

Source: Men Still Dominate Worldwide Internet Use, 2002.

all shopping categories except health and apparel (Ebates.com, 2000). Men not only report more incidences of Internet shopping, they also tend to be "heavy buyers," spending over $500 more online than women (48%) (Bhatnagar, Misra, & Roy, 2000). Men also exhibit considerable variation in Internet shopping behaviors, whereas women tend to stick to "click-and-mortar" behavior patterns—"shopping" online but buying offline ("What Kind of Dot-Shopper Are You?," 2000).

The functional depth of Internet adoption varies across age groups within as well as across countries. Among Internet users in the age group 18 to 24 years in the United States, women prefer to visit news and entertainment sites, whereas men prefer search engines and sports sites (Nua Internet Surveys, 2002b). Similarly, a survey conducted in the United Kingdom indicated that about 80% of men aged 55 and older used the Internet for searching information or for pursuing their hobbies, whereas 86% of women in the same age range preferred to use the Internet to communicate with friends and family (Nua Internet Surveys, 2002c).

Similarly, gender composition of the online registrants at a local U.S. newspaper indicated that, for all age groups over 25, more men than women tend to register online (Figure 3).

SOURCES OF GENDER BIAS IN INTERNET ADOPTION

Because the adoption of the Internet is a relatively new phenomenon, there are limited data on changes over time, particularly for countries outside the United States. Based on available data, it is likely that the gender bias will largely disappear over time. For instance, data from online registrants at a newspaper site in the United States in 2002 (see Figure 3), clearly show the narrowing gender gap at younger ages until the pattern reverses itself for the 13–24 age group. For this egalitarian pattern to emerge for all age and income groups and across all countries, several forces must operate in a convergent manner. Figure 4 provides an overall framework that offers a perspective on various factors that contribute to the gender bias in the adoption of modern ICTs such as the Internet.

There are several variables at the macro level that influence the overall environment in which decisions are made regarding technology—decisions regarding who designs the technology, what features are included in the design—that influence the usefulness and adoption patterns of technology. There also are several factors at the micro level that influence evaluation and adoption of specific technologies. A complex interplay between these forces shape the gender–technology interactions and lead to gender symmetry or asymmetry in the overall adoption of a specific technology like the Internet as well as the width and depth of adoption.

Sociocultural Factors

Culture can be viewed as one of the components of external variables in the technology acceptance model (TAM) proposed by Davis (1989). Although men and women everywhere differ behaviorally in some ways, they do not always differ in the same ways or to the same degree across countries. Cultural factors explain a significant

Table 3 Location (Home or Workplace) of Internet Use in Rhode Island, USA

	MALE ($n = 80$) (%)			FEMALE ($n = 56$) (%)		
Application Used	None	At least in one locations	Both locations	None	At least in one locations	Both locations
E-mail	2.5	17.7	79.7	0	26.8	73.2
Online information services	3.8	22.8	73.4	3.6	23.2	73.2
Shopping	12.8	48.7	38.5	20.4	59.3	18.5

Source: Dholakia, Dholakia, Mundorf, and Xiao (2002).

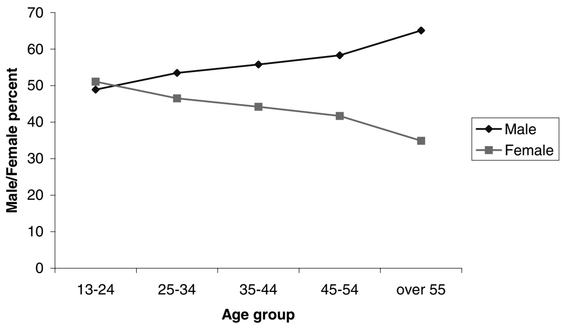


Figure 3: Gender distribution of online registration at a U.S. newspaper site.

proportion of such variations (Segall, Dalon, Berry, & Poortinga, 1990). Such differences are likely to result in gender bias in Internet adoption in several ways. First, cultural factors determine the level of involvement of men and women in technology-related decisions. Second, culture is one of the important factors in determining the likelihood of learning the skills required for Internet use. Third, culture influences the adoption of a technology by making the alternative technologies more or less attractive in performing a function. We describe each of these factors in the following paragraphs.

Women's Involvement in Decision Making

Cultural factors influence women's involvement in decision making at various levels: household, organization, and national levels. Rahman and Kumar (1998) found limited participation of women in technology adoption decisions in Bangladesh. Truman and Baroudi's (1994) analysis indicated that women receive lower salaries than men even after controlling for job level, age, education, and work experience. These same factors constitute some of the causes of women's underrepresentation in the information-systems-related occupations in the United States (in 1991, 33.7% of system analysts and 34 % of U.S. programmers were women; Truman & Baroudi, 1994). The data from United Kingdom show similar relationships. In 1991, British women accounted for only 10% of the professional membership of the British Computer Society, 3% of the data processing managers, 20% of senior system analysts, and 25% of programmers (Beech, 1991). Women are also underrepresented in national-level decision making. As Table 4 indicates, women's proportion of seats in parliaments varies from 3.5% in Arab states to 21.2% in East Asia. The differences in women's involvement in decision making may be both a cause and an effect of differences in terms of job preferences, educational specializations and Internet usage.

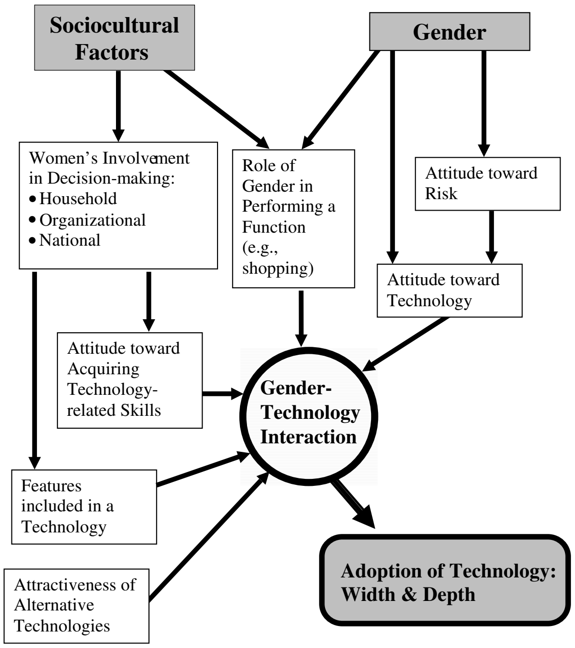

Sociocultural Factors
Gender
Women's Involvement in Decision-making:
• Household
• Organizational
• National
Role of Gender in Performing a Function (e.g., shopping)
Attitude toward Risk
Attitude toward Technology
Attitude toward Acquiring Technology-related Skills
Gender-Technology Interaction
Features included in a Technology
Attractiveness of Alternative Technologies
Adoption of Technology: Width & Depth

Figure 4: Gender-culture-technology interaction.

Influence of Culture on Attitudes Toward Acquiring Technology-Related Skills

Sociocultural factors influence the perceived benefits of acquiring technology-related skills. In general, women tend to face higher barriers than men in accessing education and training required to use ICTs (UNCTAD, 2002). Moreover, such barriers vary widely across cultures. Consider, for instance, two countries at about the same level of economic development, India and Egypt. In India, it is easier for parents to find better husbands for their daughters if the daughters are educated, hence they emphasize girls' education (Kumar, 1991). Egyptians, on the other hand, do not give importance to women's education because Egyptian men seek less educated wives (Mianai, 1981).

Availability of skills required to use a technology influences perceived ease of use and perceived usefulness of the technology and hence attitude toward its adoption. In the case of the Internet, for instance, a potential adopter must have a computer and English language skills. In some cases, the required skills may be learned only for the purpose of using the technology, whereas in other cases they are learned for some other (more general) purposes. For instance, it is unlikely that a potential adopter would learn English and acquire computer skills solely for the purpose of using the Internet. Thus, additional perceived benefits of learning the skills influence the attitude toward learning them (Fishbein, 1967) and hence adopting the technology. Because of job-related reasons, men in many countries may learn some English, and this skill also incidentally helps in using the Internet.

Features Included in the Internet

It is not surprising that because of these sociocultural influences, the features included in the design of a technology tend to favor men over women. Features unappealing to women users are likely to exist for several reasons. First, women are highly underrepresented in administration, managerial, and technical jobs in most of

Table 4 Gender-Related Indicators Influencing Internet Use

REGION	GENDER-RELATED DEVELOPMENT INDEX	SEATS IN PARLIAMENT HELD BY WOMEN (% OF TOTAL)	LITERACY RATE (%)		GDP PER CAPITA ($ 1998)	
			F	M	F	M
Arab states	0.612	3.5	47.3	71.5	1837	6341
East Asia	0.710	21.2	75.5	91.1	2788	4297
Latin America and Caribbean	0.748	12.9	86.7	88.7	364.	9428
South Asia	0.542	8.8	42.3	62.7	1147	3021
South East Asia and Pacific	0.688	12.7	85	92.4	2316	4154
Afr	0.459	11	51.6	68	1142	2079
Eastern Europe and Commonwealth of Independent States (CIS)	0.774	8.4	98.2	99.1	4807	7726
Organization for Economic Cooperation and Development (OECD)	0.889	15.1	96.7	98.2	14165	26743

Source: United Nations Development Program (2000).
Note. F = female; GDP = gross domestic product; M = male.

the countries. For instance, whereas the proportion of women administrators and managers in the United States is 44.4%, women are virtually nonexistent in such jobs in many of the countries (United Nations Development Program, 2000). Such lack of top-level representation is compounded by manufacturers' unawareness of women's needs. Even when market research data are available, designers tend to ignore such evidence of actual needs and rely instead on speculative ideal types or stereotypes of women users (Cockburn & Dilic, 1994). Moyal (1992) argued as follows:

> It is the male engineer, technician, manager, strategic planner and policy-maker who have devised and installed the telegraph, the telephone, radio and television broadcasting stations, the communications satellite, and the advanced digital and mobile communication infrastructures and networks that link and continue to upgrade our spiraling telecommunication world.

In the case of the Internet, several technology-related factors tend to favor male users, including the following (Abernathy, 1999; Biersdorfer, 2001; Herring, 1999; Marketing to Women, 2001):

- Stereotyped views of female users;
- Male-oriented aggressive formats of computer games;
- Largely male-oriented online discussion groups, lacking elements of civility and online etiquette that women desire; and
- Prevalence of Internet pornography.

Features designed into a technology influence perceived usefulness and perceived ease of use (Davis, 1989), the major predictors of the likelihood of the adoption of a technology. Lack of features that appeal to women means lower perceived usefulness and perceived ease of use for female users, leading to lower width and lower overall depth of Internet adoption among women. If more features that appeal to women are included, then women's likelihood of adoption as well as the width and the depth of adoption are likely to increase.

Attractiveness of Alternative Technologies

Cultural factors could influence the propensity to adopt a given technology in performing a certain function by making the "alternative technology" less or more attractive. For example, consider two relatively affluent Middle Eastern countries, Saudi Arabia and Kuwait. Whereas the proportion of female Internet users in the Arab states is estimated to be between 4% (Internet.com, 2001) to 6% (Hafkin & Tagger, 2001), the proportion is more than 66% in Saudi Arabia ("How women beat the rules," 1999; "International Online," 2000) and more than 50% in Kuwait (Wheeler, 1998). The rapid increase in the number of female Internet users in Saudi Arabia is driven mainly by women's use of the Internet for business and personal reasons. In Muslim countries such as Saudi Arabia, women are not allowed to drive cars. Also, open and public interactions between men and women are highly constrained outside of marriage and family in Muslim societies. The Internet helps overcome such barriers. Wheeler (1998), for example, reported that several people in Kuwait had fallen in love and subsequently married through use of Internet relay chat technology. The Internet proved to be much better than the existing alternatives for women in the Middle Eastern cultural settings, given the social norms against women driving cars and the restricted public intermingling of the two sexes.

Gender-Related Factors

As social products, any technology including the Internet is not culturally neutral or value-free. The degree of compatibility between the values and norms of the technology and those of a potential adopter largely determine the adoption patterns of the technology (Rogers, 1983). The

Table 5 Summary of the Nature of Gender Bias in Terms of Several Indicators

INDICATOR	RESEARCH FINDINGS	REMARKS
Proportion of female Internet users	Lower than the proportion of male Internet users in most markets Tends to increase with the maturity of the Internet market.	Has reversed in the United States and Canada; female users outnumber male users in these countries
Width of adoption	Tends to be higher for men.	Men tend to use Internet for more applications than women.
Functional depth of adoption	Depends upon the type of Internet application	For instance, in the United States, Internet adoption for visiting news and entertainment sites is deeper for women and that for visiting search engines and sports sites is deeper for men
Overall depth of adoption	Tends to be higher for men	Men tend to spend more time online than women and to visit more pages

Sources: Nua Internet Surveys (2000); Graphic, Visualization, & Usability Center Center (1998); O'Leary (2000).

"values" of most of the ICT products and services, including the Internet, tend to be more masculine than feminine, which partly explains the existing gender-related digital divide (Herring, 2000).

Men and women are also "specialized" for different tasks. In the United States, for instance, "going shopping" is associated with women whereas men "work" (Firat & Dholakia, 1998). These specializations partly explain why American women still assume major responsibility for shopping (Dholakia, 1999) and why 70% of women make most of the household and purchase decisions (Hawfield & Lyons, 1998). These gender-related stereotypes persist, even among young adults (Dholakia & Chiang, 2003). Such differences in specialization tend to result in gendered adoption patterns and gendered width and depth of applications of a given technology. As Internet applications have expanded to include roles and responsibilities that have traditionally fallen on women—including greater range of shopping and family communications—the proportion of women users on the Net has increased.

Attitude Toward Risk and Attitude Toward Technology

Men and women differ significantly in terms of attitude toward risk (Slovic, 1966) and toward technology in general (Brunner & Bennett, 1998). Because so-called innovators of a new technology have more favorable attitudes toward risk (Gatignon & Robertson, 1991), women's risk-averse behavior is likely to result in lower rates of technology adoption. Men and women also differ in terms of their attitudes toward technology. In an experiment with school children, Brunner and Bennett (1998) found that girls were more ambivalent about technologies, were more likely to get bored with a bad technology experience, and were less likely to fix a technology if it breaks, compared with boys in the same age group. Venkatesh and Morri (2000) reported gender differences in the importance assigned to various factors for the adoption of ICT, and Gefen and Straub (1997) reported gender differences in the perception and use of e-mail.

CONCLUSION

In summary, men and women have different "cultures," are "specialized" in different tasks, and have different preferences. Such differences tend to interact with the features found in the Internet and other modern ICTs in ways that intensify their perceived usefulness and the perceived ease of use in favor of men rather than women. Table 5 summarizes the existing gender bias in Internet use in terms of several indicators.

Several limitations characterize the data used for the analysis and interpretation of gender differences in Internet use. First and foremost, our analysis has used a variety of sources, each of which has employed its own methodology of collecting, analyzing, and reporting data. Next, the data use *sex* and *gender* interchangeably such that it is difficult to capture sociocultural differences in the "meanings, roles, values, and status" assigned to the words *male* and *female*. Keeping these limitations in mind, the analysis presented in this chapter seems to support the following conclusions:

- There is a gender gap measured by the proportion of male and female Internet users.
- This gap is decreasing and has even reversed in overall terms in the United States and Canada.
- When the gap is examined in terms of additional constructs (width and depth of adoption), differences in male–female uses of the Internet persist even in countries when the overall gap has reversed. RISQ analysis of Canada (Figure 2) shows time use differences. Tables 2 and 3 show application differences.

Several factors contributing to gender bias in the adoption of the Internet have been discussed in this chapter. Drawing on a general model of gender–technology–

culture interaction, we first reviewed the sociocultural factors that contribute to the Internet being less compatible with the culture and values of women than those of men. Limited involvement of women in technology-related decisions may be both a cause and effect of differences at several levels. With lower preference for disciplines that demand higher intensity of Internet use such as science, engineering, and technology, features included in the Internet and related technologies tend to favor men over women. The two genders also differ in their attitudes toward acquiring required skills, and in valuing alternative technologies. These factors partly explain a lower rate of Internet adoption among women.

Second, we reviewed gender-related factors that influence men's and women's attitudes toward risk and attitudes toward technology in general. Greater risk aversion and a generally more unfavorable attitude toward technology-related problems tend to result in lower adoption rates among women. Because men and women differ in terms of their specializations, the functional depth corresponding to an application (e.g., shopping) tends to be different for the two genders. It is not surprising that men report greater purchases of technical products online and women purchase more apparel. When an application is consistent with gender specialization, such as a woman's greater role in maintaining and enhancing the family's social interaction and communication patterns, then almost universal use by women can be expected, as is becoming evident to some extent with respect to e-mail. Finally, the adoption of the Internet, like all other ICTs, is a function of the income level of the potential adopters. Women's lower per capita income and lower literacy rate partly explain the existing gender bias in terms of the adoption of the Internet.

Managerial and Policy Implications

It is likely that with the passage of time many of these gender differences will diminish. To the extent that sociocultural and gender-related factors act as barriers to the design and use of technology such as Internet, these sources of gender bias can be tackled if appropriate measures are taken at various levels. Because men and women have different "cultures," specializations, and preferences, the prevalent "one-size-fits-both-genders" approach is less likely to work in the design of a technology such as the Internet. Research is needed to understand women's needs in better ways and technology designers should be induced to incorporate features that meet such needs.

At the policy level, greater emphasis is needed on women's education—especially in science, engineering, technology, and administration fields. Women's representation in ICTs is strong in Singapore because of the government's "concerted" state directed ICT training—58% of analyst programmers and 52% of analyst designers in 1987 were women (Webster, 1996). Women educated in such fields are likely to have the skills and propensity to adopt modern ICTs. Women's enrollments in such disciplines would have doubly greater social benefits since women trained in ICTs are likely to be the future designers and incorporate features that are likely to favor women's adoption. The emphasis must be accompanied by new ways of imparting the education, however (Shaffner, 1993, p. 97):

> The whole idea of sitting at a desk in an office and using a computer would have to change. It would have to be more like something that you could do in an environment where you could have children, babies, lovers, and community.

Policies are also needed to impart ICT skills to a broader group of women and enhance existing skills. These include ICT awareness programs, ICT-related training opportunities for women in workforce, and acquisitions of the right and sufficient ICT skills (UNCTAD, 2002). Finally, the extent to which the gender-related digital divide can be bridged depends on women's access to and involvement in technology-related decisions at the household, organizational, national, and international levels.

ACKNOWLEDGMENT

The authors are grateful to multiple reviewers of *The Internet Encyclopedia* for detailed and valuable comments on earlier versions of this chapter. Support for the research from RITIM (Research Institute for Telecommunications and Information Marketing) at The University of Rhode Island is gratefully acknowledged.

GLOSSARY

Adoption of an innovation A micro process that focuses on the stages through which an individual passes when deciding to accept or reject an innovation.

Depth of adoption of a technology The variety and extent of usage of the acquired technology or the purchase of related products.

Functional depth of adoption of a technology The frequency that a multifunctional technology such as the Internet will be used for a particular function (e.g., for shopping).

Gender The "social constructed roles ascribed to males and females. These roles, which are learned, change over time and vary widely within and between cultures" (United Nations definition). Although *sex* and *gender* are often used interchangeably, sex is a biological variable and gender a social construct. The adjectival and noun forms of *male* and *female* are typically employed in a biological sense, whereas the adjectives *masculine* and *feminine* are typically employed in a gendered, social sense.

Gender-related development index (GDI) A composite index measuring average achievement in the three basic dimensions captured in the human development index—a long and healthy life, knowledge, and a decent standard of living—adjusted to account for inequalities between men and women.

Gross domestic product (GDP) The sum of the total value of consumption expenditure, total value of investment expenditure, and government purchases of goods and services.

Information and communications technologies (ICTs) Technologies that facilitate the capturing, processing, storage, and transfer of information.

Internet The "global information system that (i) is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons; (ii) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and (iii) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein" (The Federal Networking Council, 1995).

Total or overall depth of adoption of a technology A measure that can be determined by the total time spent using the technology in a given period of time (e.g., per month).

Width of adoption of a technology The number of people within an adoption unit who use the product, or the number of different uses of the product.

CROSS REFERENCES

See *Digital Divide; Global Diffusion of the Internet; Global Issues; Internet Literacy; Legal, Social and Ethical Issues.*

REFERENCES

Abernathy, D. J. (1999, December). Second and fourth rocks from the sun. *Training and Development,* p. 18.

Beech, C. (1991) Women and WIT. In I. V. Eriksson, B. Kitchenham, & K. G. Tijdens (Eds.), *Women, work and computerization: Understanding and overcoming bias in work and education.* Amsterdam: North Holland.

Bhatnagar, A., Misra, S., & Rao, H. R. (2000). On risk, convenience and Internet shopping behavior. *Communications of the ACM, 43,* 98–105.

Biersdorfer, J. D. (2001, June 7). Among code warriors, women, too, can fight. *New York Times.* Retrieved June 7, 2001, from http://www.nytimes.com/2001/06/07/technology/07WOME.html?

Brisco, R. (2002). Turning analog women into a digital workforce: Plugging women into the new Asia economy. Retrieved January 23, 2003, from http://www.digitaldividenetwork.org/content/stories/index.cfm?key = 135

Brunner, C., & Bennett, D. (1998, February). Technology perceptions by gender. *The Education Digest,* 56–58.

Carter, C., & Grieco, M. (1998). New deals, no wheels: Social exclusion, teleology and electronic ontology. Retrieved March 22, 2000, from http://www.geocities.com/margaret_grieco/working/wheels.html

Carter, C., & Grieco, M. (2000). New deals, no wheels: Social exclusion, tele-options and electronic ontology. *Urban Studies, 37,* 1735–1748.

China Internet Network Information Center (2001, January). *Semiannual survey report on the development of China's Internet.* Retrieved March 22, 2001, from http://www.cnnic.gov.cn/develst/e-cnnic200101.shtml

Cockburn, C., & Dilic, R. F. (1994). *Bringing technology home: Gender and technology in a changing Europe.* Buckingham, U.K.: Open University Press.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly, 13,* 319–339.

Demographics: European women surf to a different drum (2002, March 25). *CyberAtlas.* Retrieved March 22, 2000, from http://cyberatlas.Internet.com/big_picture/demographics/article/0,,5901_997491,00.html

Deok-hyun, K. (2001). ROK shows lowest Internet gender gap in Asia. *The Korea Times.* Retrieved March 22, 2003, from http://www.hk.co.kr/times/200105/t2001050316572540110.htm

Dholakia, R. R. (1999). Going shopping: Key determinants of shopping and motivations. *International Journal of Retail and Distribution Management, 27,* 154–165.

Dholakia, R. R., & Chiang, K. P. (2003). Shoppers in cyberspace: Are they from Venus or Mars and does it matter? *Journal of Consumer Psychology, Special Issue on Consumers in Cyberspace, 13,* 171–176.

Dholakia, N., Dholakia, R. R., Mundorf, N., & Xiao, X. (2002). Interactions of transportation and telecommunications: A survey. Kingston, RI: University of Rhode Island, URI Transportation Center.

Ebates.com (2000). Are you an online window shopper—or an actual buyer? Retrieved March 1, 2001, from http://www.ebates.com/press_release.jsp?press_release = press_releases/press_012.html

E-Marketer (2001a). Profiling Indian Internet users: Part 1. Retrieved March 22, 2001, from the E-Marketer Web site: http://www.emarketer.com/analysis/easia/20010104_india1.html?ref = dn

E-Marketer (2001b). Who's using the Internet in Latin America? Retrieved March 22, 2001, from the E-Marketer Web site: http://www.emarketer.com

Firat, A. F., & Dholakia, N. (1998). The making of the consumer. In *Consuming People* (pp. 13–20). New York: Routledge.

Fishbein, M. (1967). Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement.* New York: Wiley.

Gatignon, H., & Robertson, T. S. (1991). A propositional inventory for new diffusion research. In H. H. Kassarjian & T. S. Robertson (Eds.), *Perspectives in consumer behavior* (4th ed., pp. 461–487). Upper Saddle River, NJ: Prentice Hall.

Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly, 21,* 389–400.

Graphic, Visualization, & Usability Center Center (1998). *GVU's 10th WWW user survey.* Retrieved January 23, 2001, from www.gvu.gatech.edu/user_surveys/survey-1998-10/

Hafkin, N. (2001). Gender, information technology and the digital divide in Africa and other developing areas. Retrieved March 22, 2001, from http://www.worldbank.org/gender/digitaldivide/nhafkin.ppt

Hafkin, N., & Tagger, N. (2001). Gender, information technology, and developing countries: An analytic study. Retrieved March 22, 2001, from the United States Agency for International Development Web site: http://www.usaid.gov/wid/pubs/hafnoph.pdf

Hawfield, K., & Lyons, E. (1998, April). Conventional wisdom about women and Internet use: Refuting traditional perceptions. Prepared for iVillage.com. Retrieved March 12, 2001, from the eLab Web site: http://elab.vanderbilt.edu/research/papers/html/student_projects/women/conventional_wisdom.html

Herring, S. C. (1999). The rhetorical dynamics of gender harassment on-line. *Information Society, 15,* 151–167.

Herring, S. C. (2000, Winter). Gender differences in CMC: Findings and implications. *The CPSR Newsletter, 18.* Retrieved March 12, 2001, from http://www.cpsr.org/publications/newsletters/issues/2000/Winter2000/herring.html

How women beat the rules (1999, October 2). *The Economist,. 48.*

International online (2000, April 29). *The Economist,* pp. 42–44.

Internet.com (2001, January 24). Global digital divide still very much in existence. Retrieved March 12, 2001, from http://cyberatlas.internet.com/big_picture/geographics/article/0,,5911_569351,00.html

Jupiter MMXI (2002). Current Web statistics. Retrieved February 15, 2002, from http://se.jupitermmxi.com/xp/se/home.xml

Kumar, U. (1991). Life stage in the development of Hindu women in India. In L. L. Adler (Ed.), *Women in cross cultural perspective* (pp. 143–148). New York: Praeger.

Marketing to women (2001). Girls know computers but find them boring. *Marketing to Women, 13,* 12.

Men still dominate worldwide Internet use. (2002, January 22). *CyberAtlas.* Retrieved March 12, 2002, from http://cyberatlas.internet.com/big_picture/demographics/article/0,1323,5901_959421,00.html

Mianai, N. (1981). *Women in Islam: Tradition and transition in the Middle East.* London: John Murray.

Moyal, A. (1992). Women calling!: The gendered use of the telephone, *TeleGeography*. Retrieved March 22, 2001, from http://www.telegeography.com/Publications/moyal.html

Nua Internet Surveys (2000, March 30). *In-Stat/MDR*: 60 percent of US households online by 2000. Retrieved February 15, 2003, from http://www.nua.com/surveys/index.cgi?f=VS&art_id=905355687&rel=true

Nua Internet Surveys (2001, January 15). Women outnumber men online in U.S. Retrieved March 12, 2001, from http://www.nua.ie/surveys/?f=VS&art_id=905356838&rel=true

Nua Internet Surveys (2002a, May 27). World Cup fever spreads to the Net. Retrieved July 12, 2002, from http://www.nua.com/surveys/

Nua Internet Surveys (2002b, July 12). Young Americans opt for a variety of sites. Retrieved July 12, 2002, from http://www.nua.com/surveys/

Nua Internet Surveys (2002c, August 21). Gender differences in UK seniors Internet use. Retrieved August 21, 2002, from http://www.nua.com/surveys/

O'Leary, M. (2000, December). Web goes mainstream for everybody. *Online,* 80–82.

Pastore, M. (2001), Internet remains a man's domain, Retrieved February 15, 2003, from http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_809341,00.html

Rahman, S., & Kumar, J. (1998). Technological change and women's participation in crop production in Bangladesh. *Gender, Technology and Development, 2,* 243–267.

Rainie, L. (2002). *Women surpass men as e-shoppers during the holidays.* Retrieved January 1, 2002, from the Pew Internet and American Life Project Web site: http://www.pewinternet.org/reports/toc.asp?Report=54

Réseau D'informations Scientifiques du Québec (1998). The RISQ surveys of Quebec internauts: Survey 4. Retrieved March 22, 2001, from http://www.risq.qc.ca/survey/4/Internet/int_heures.html

Rogers, E. M. (1983). *The diffusion of innovations* (3rd ed.). New York: Free Press.

Segall, M. H., Dalon, P. R., Berry, J. W., & Poortinga, Y. H. (1990). Human behavior in global perspective: An introduction to cross-cultural psychology. Boston: Allyn & Bacon.

Shaffner, J. (1993). Gender and politics in machine: Computer scientists and social change. Unpublished M.Phil. dissertation, University of Edinburgh, Scotland.

Silicon.com (2001, July 4). Women more efficient surfers than men: *Less faffing, more clicking*... Retrieved November 21, 2002, from http://www.silicon.com/news/500019/1/1025498.html

Slovic, P. (1966). Risk-taking in children: Age and sex difference. *Child Development, 37,* 169–176.

Stewart, J. (2002). Information society, the Internet and gender. A summary of pan-European statistical data. Retrieved February 15, 2003, from http://www.rcss.ed.ac.uk/sigis/public/D02/D02_Part2.pdf

The Federal Networking Council (1995). Definition of the Internet. Retrieved November 21, 2000, from http://www.fnc.gov/Internet-res.html.

Truman, G., & Baroudi, J. J. (1994). Gender differences in the information systems managerial ranks: An assessment of potential discriminatory practices. *MIS Quarterly, 18,* 129–141.

Unilever (2001). Women online: Statistics on likes, dislikes. Retrieved March 22, 2001, from http://www.clienthelpdesk.com/statistics_research/women_online.html

United Nations Conference on Trade and Development (2002). *E-commerce and development report 2002.* Retrieved November 21, 2002, from http://r0.unctad.org/ecommerce/docs/edr02_en/ecdr02ch3.pdf

United Nations Development Program (2000). *Human development report.* Geneva, Switzerland.

Venkatesh, V., & Morri, M. G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. *MIS Quarterly, 24,* 115–139.

Webster, J. (1996). *Shaping women's work: Gender employment and information technology.* London: Longman.

What kind of dot-shopper are you? (2000). Retrieved May 22, 2001, from http://www.harrisinteractive.com/news/

Wheeler, D. L. (1998). Global culture and cultural clash: New information technologies in the Islamic world—a view from Kuwait. *Communication Research, 25,* 359–376.

World Bank (2002). Glossary of key terms in social analysis. Retrieved March 22, 2003, from http://lnweb18.worldbank.org/ESSD/essdext.nsf/61DocByUnid/9FDABDCD88D03D6585256BC00065F517?Opendocument

World Bank (2003). *World Development Report 2003.* Retrieved February 15, 2003, from the World Bank Web site http://econ.worldbank.org/wdr/wdr2003/

Geographic Information Systems (GIS) and the Internet

Haluk Cetin, *Murray State University*

INTRODUCTION

A geographic information system (GIS) consists of hardware, software, and users to support the capture, storage, retrieval, update, management, manipulation, analysis, display, and modeling of geospatial data. Before GIS, researchers used films and mylars to manually overlay spatial information. The integration of spatial information with nonspatial attributes (variable) is a critical component of a GIS. Many large complex tasks are accomplished much more quickly by a GIS than by a human working alone. A GIS is a "smart map" system linking databases to digital maps. Several names, such as spatial information system, land information system, natural resource management information system, planning information system, and environmental information system, have been used for GIS, giving a "high-tech" feel to spatial information (Goodchild & Kemp, 1992).

GIS and related technologies are becoming more available to many users in different disciplines through the World Wide Web. GIS and related technologies will help greatly in the management and analysis of worldwide data and allow better understanding of environmental and other processes around the world.

Desktop GIS has been the system of choice, but it has many limitations including high cost, high learning curve, and limited public access of data. On the other hand, an Internet-GIS (i-GIS), which focuses on distributed geographic information services for the decentralization of geographic information management, has many advantages including decentralized data storage, wider user access, and convenience (Plewe, 1997; Tsou & Buttenfield, 2002). The advent of the Internet and computer technology has made it possible to share and analyze data through World Wide Web–based management systems. The Internet and GIS have been used to download, preprocess, review, modify, and analyze up-to-date geographic data since the early 1990s. Although today's Web-based systems offer an excellent opportunity to use GIS beyond the home and office, existing i-GIS map servers do not provide the complex analytical functionality required by many users.

Internet-GIS is becoming one of the most rapidly evolving fields in e-commerce. For organizations and institutions, i-GIS will provide easy access to GIS data with decentralized databases.

Future GISs, including automated processing at offices and homes, will likely be a part of our daily lives. When you walk into your house, the system will know where you are and what kinds of things you might want to do, such as turning on specific appliances or planning a trip using virtual environments.

A BRIEF HISTORY OF GEOGRAPHIC INFORMATION SYSTEMS

GIS technology, which began in the 1960s, has been one of the most rapidly evolving fields during the past two decades. The Canada Geographic Information System (CGIS) and the Urban and Regional Information Systems Association (URISA) were developed in 1963, the latter a nonprofit association using GIS technology in public works and services and local and state planning agencies. The Harvard Laboratory for Computer Graphics established in 1964 pioneered many aspects of GIS. In 1965, Synagraphic Mapping System (SYMAP), an automated computer mapping application system, was developed at the Northwestern Technology Institute and at the Harvard

Laboratory for Computer Graphics. In 1967, the United States Central Intelligence Agency (CIA) developed an Automatic Mapping System (AUTOMAP), a map compilation system at the global level. The establishment of several companies, including the Environmental Systems Research Institute (ESRI) and Intergraph Corporation in the United States and Laser-Scan in the United Kingdom, initiated worldwide commercial applications. In the 1970s, several GIS programs and systems were established and the first Landsat Multispectral Scanner (MSS) satellite (originally known as ERTS-1) was launched. In 1977, the U.S. Geological Survey developed one of the first spatial data formats, the digital line graph (DLG). The establishment of other companies, such as ERDAS, which was founded in 1978, and the first vector based GIS data structure, ODYSSEY GIS, which was developed at the Harvard Lab, provided major advances. In 1981, ESRI introduced one of the earliest commercial GIS software packages, Arc/Info, which has become by far the most widely used GIS software including topological consistency in GIS data sets. During the same year, the global positioning system (GPS) project became operational. In 1982, the development of the Geographic Resources Analysis Support System (GRASS), an open source/free raster–based GIS software, was begun at the U.S. Army Construction Engineering Research Laboratories, a branch of the U.S. Army Corp of Engineers. During the late 1980s, several GIS companies such as MapInfo (1986), SPOT satellite program (1986), Idrisi project (1987); a GIS journal; the International Journal of Geographical Information Systems (1987); a GIS conference; the first GIS/Land Information Systems (LIS) conference (1988); and data formats such as the first public release of the U.S. Bureau of Census TIGER (topologically integrated geographic encoding and referencing) digital data were created. The U.S. Geological Survey defined one of the latest spatial data formats, the Spatial Data Transfer Standard (SDTS), in 1992 (Federal Information Processing Standard 173). During the 1990s, GIS became one of the most rapidly evolving information technology (IT) fields. Today GIS is being used in many disciplines including the hard sciences; the engineering, medical, and social science fields; at every level of government including the local, state, and national levels; and in business and industry.

One of the important issues in GIS is meta-data standards, which describe the origin, content, quality, condition, and other characteristics of geospatial data. The Federal Geographic Data Committee (FGDC) coordinates the development of the National Spatial Data Infrastructure (NSDI). According to the FGDC, "the NSDI encompasses policies, standards, and procedures for organizations to cooperatively produce and share geographic data. The 17 federal agencies that make up the FGDC are developing the NSDI in cooperation with organizations from state, local and tribal governments, the academic community, and the private sector." The FGDC approved the Content Standard for Digital Geospatial Metadata (FGDC-STD-001–1998) in June 1998 (http://www.fgdc.gov/metadata/metadata.html). The FGDC has created a Web-based FGDC Metadata Entry System (http://130.11.52.178/metaover.html) "to stimulate the creation of basic FGDC-compliant meta-data records for the cataloging of spatial data sets." Also, the NSDI/FGDC National Geospatial Data Clearinghouse Information Resource Web site (NSDI/FGDC, 1998) node has been created to provide links to spatial data. The Clearinghouse Activity is a decentralized system of servers located on the Internet which contain field-level descriptions of available geospatial data.

DESKTOP GEOGRAPHIC INFORMATION SYSTEMS

There are several definitions of GIS (Chrisman, 1997). GISs are a form of information system applied to geographic data and consist of several interconnected subsystems including data storage, management, and processing; data analysis and manipulation; and output subsystems used to capture, store, retrieve, update, process, analyze, model, and display spatially referenced (georeferenced) data systematically. A GIS is different from a database management system (DBMS) in terms of capability. DBMSs store and process mostly nonspatial data. On the other hand, a GIS, which may also include relational database management systems (RDBMS), not only processes geospatial data but also link nonspatial (or attribute) information to spatial data. A GIS enables institutions and managers to manipulate geographic data more efficiently and to make informed decisions for planning purposes.

Contributing Technologies and Disciplines

GIS represents a convergence of numerous disciplines and technologies. A GIS has the potential to offer a technology that can be used by diverse disciplines emphasizing spatial data collection, analysis, manipulation and integration, analysis, and modeling. Several disciplines including artificial intelligence (AI), civil engineering, computer science and engineering, earth sciences, electronic engineering, geodesy, geography, mathematics, photogrammetry, remote sensing, and semiotics have contributed to the technology used in today's GIS.

Components of a GIS

A GIS is an integrated system consisting of hardware, software, data, and users (Aronoff, 1995; Bolstad, 2002; Figure 1). GIS hardware includes computers, input devices, data storage, display, and output systems. GIS software is the main part of the system and provides for spatial data management, manipulation, analysis, display, and output. Data input devices include digitizers, scanners (flat-bed or drum), and imaging systems such as digital cameras. Data storage units are tape and disk drives (hard disk and floppy), CD/DVD drives, and other optical and magnetic storage devices. GIS display and output devices are monitors, projectors, printers, and plotters. GIS output can be in hard-copy (paper maps, tables, etc.) or softcopy format (digital data and images).

GIS Data Structures

There are two major data structures used in GISs, vector- and raster-based (Goodchild & Kemp, 1992). Each data

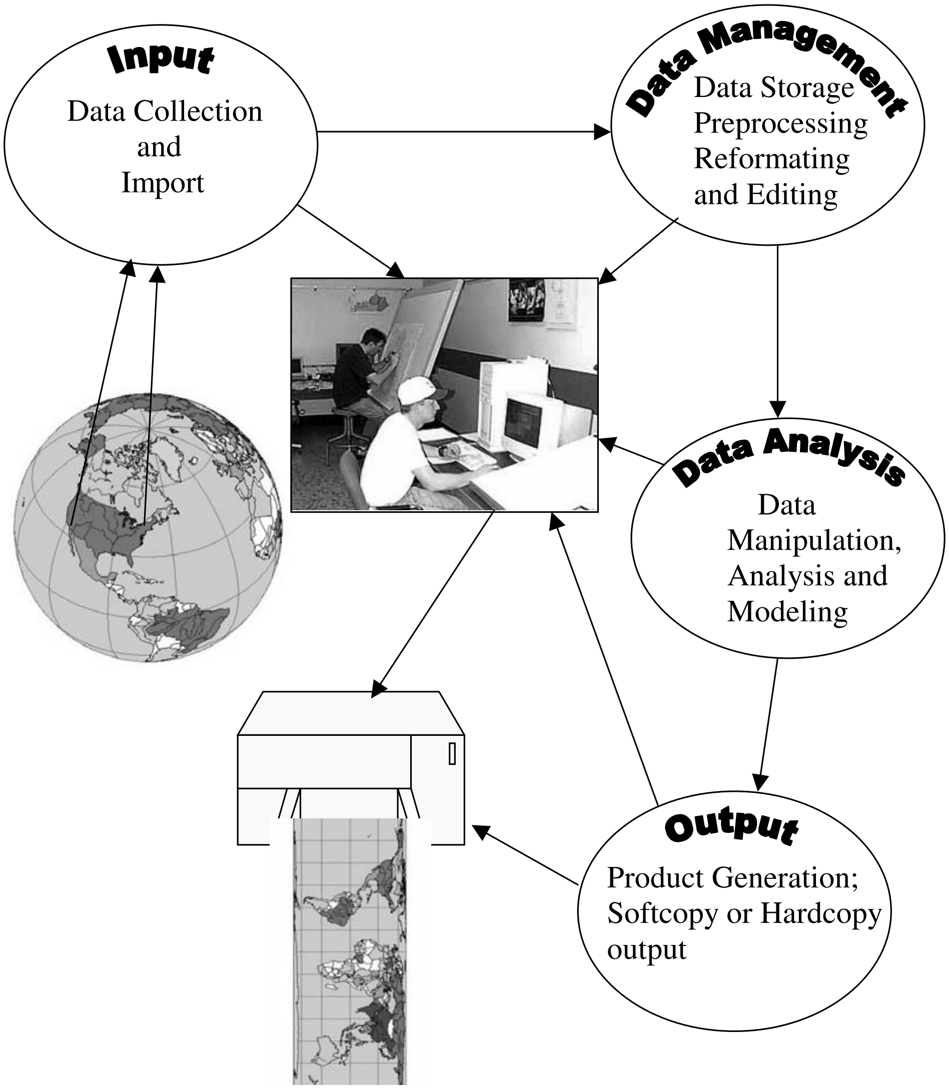

Input
Data Collection and Import
Data Management
Data Storage
Preprocessing
Reformatting and Editing
Data Analysis
Data Manipulation, Analysis and Modeling
Output
Product Generation; Softcopy or Hardcopy output

Figure 1: Components of geographic information systems.

structure has advantages and disadvantages (Figure 2a and 2b). A vector-based GIS is more useful when point (a location), line (roads or rivers), and polygon (fields or parcels) types of data are used (Figure 2a). The raster data model is more suitable for image-based data, such as aerial photos or satellite images (Figure 3). Raster data usually consist of rectangular grid cells (or pixels). Raster data can be manipulated quickly by a computer, but they often take up substantial disk space and are less detailed. Most modern GISs include the two data structures because both are used for most projects (Figure 4).

A GIS incorporates data from different sources and has many functions, including data browsing, retrieving, updating, displaying and querying, mapmaking, address matching and route finding, modeling, and terrain and statistical analysis. A GIS uses map mathematics to deal with many tasks including data conversions, data overlays, and buffers (DeMers, 2000).

GIS has applications in almost every discipline requiring some spatial data handling. Integration of a GIS technology with other systems such as GPS allows locations to be found quickly as in the case of 911 emergency response systems (Xue, Cracknell, & Guo, 2002). These applications are discussed in detail in the next section.

APPLICATIONS OF GIS

The following are some examples of GIS applications in several fields. There are also many GIS applications in the various subcategories, which are not discussed here.

Agriculture Applications

Farmers who have large fields can manage farm data easily with a GIS. They can create base maps that include roads, houses, barns, and farm field boundaries and

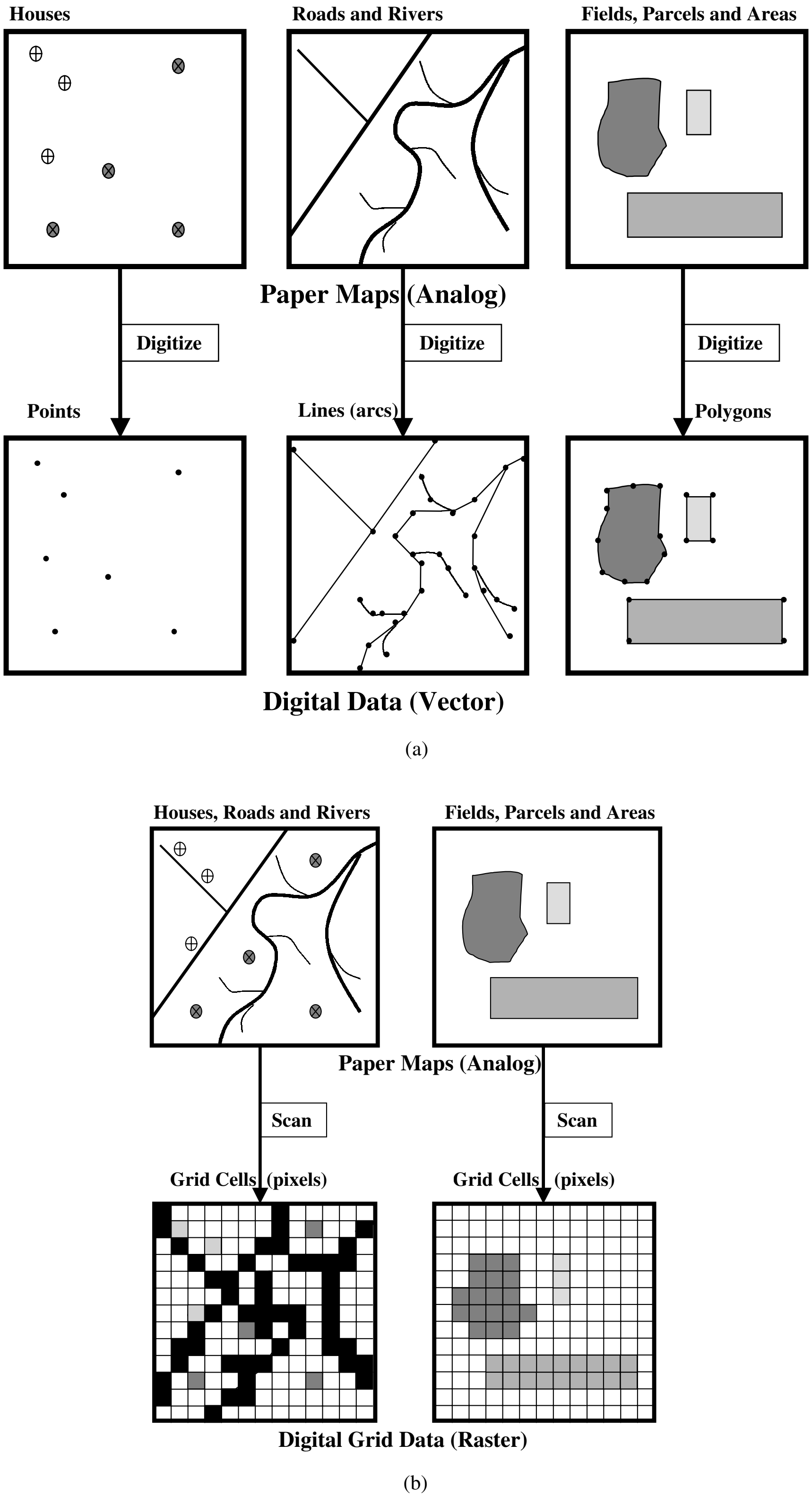


Figure 2: Geographic information system data structures: (a) vector data model and (b) raster data model.

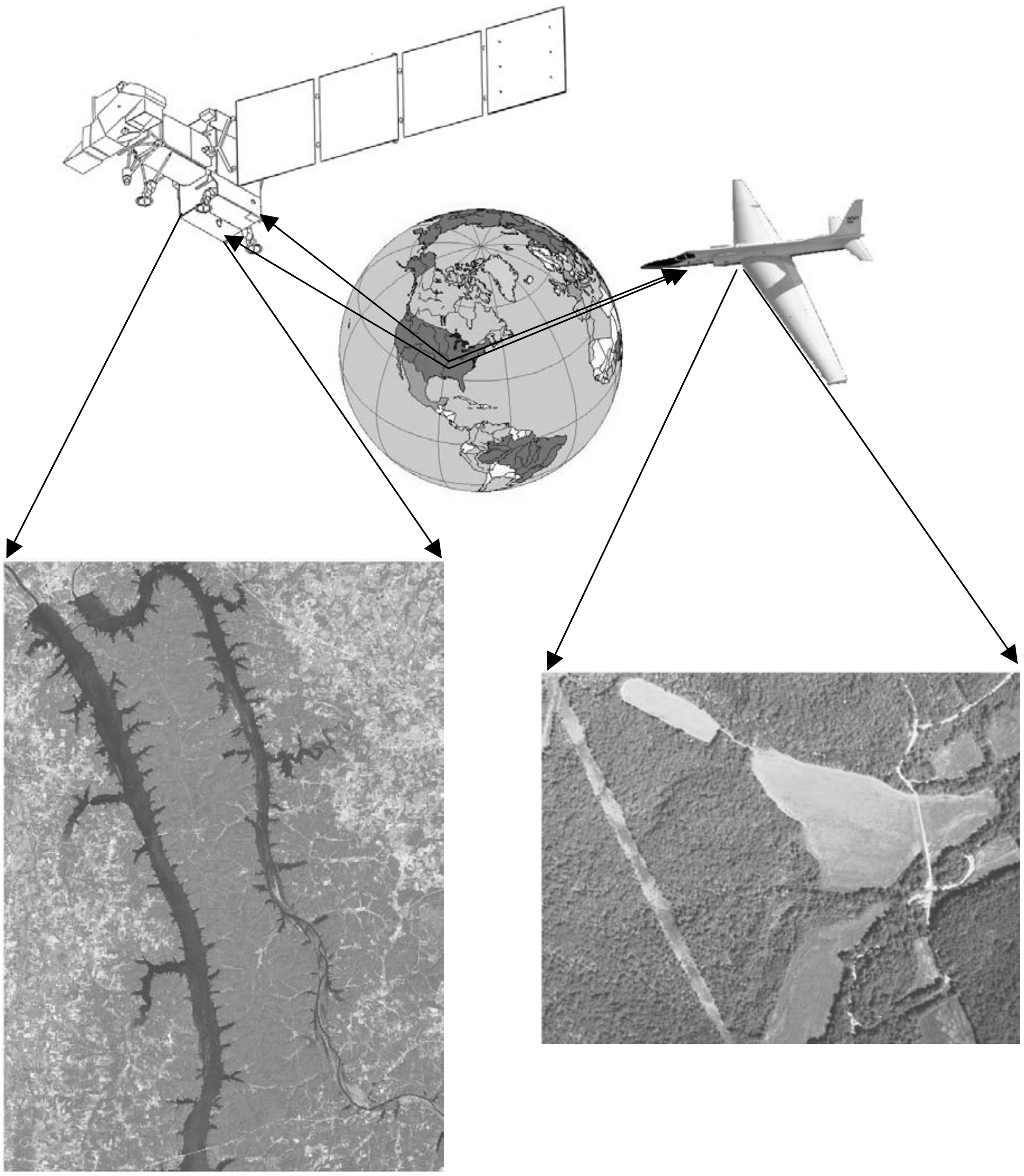


Figure 3: Satellite and airborne remotely sensed data suitable for raster data model (satellite and aircraft images are courtesy of NASA).

overlay them with maps that include drainage systems, soils, land use, pest and disease maps, and topography (elevation, slope, and aspect). The GIS may be used to improve management efficiency and the decision-making process. Many companies and research institutions are developing precision agriculture systems (PAS) using GIS and differential GPS to plan fertilizer, pesticide and herbicide, and irrigation applications at an optimum level to improve crop yield. GIS allows farmers access up-to-date farm data to increase production and reduce costs.

There are several GIS based PASs that allow farmers to use field data such as soil moisture, soil types, and nutrient information to develop prescriptions to apply specific amounts of fertilizer, seed, or other products. For example, seed population maps (seed prescription) can be created using soil moisture capacity of a farm field where different soil moisture conditions may exist. Many studies have shown that variable rate seeding improves yield. Higher rate seeding is applied to areas with higher soil moisture capacity, and lower rate seeding is applied to areas with lower soil moisture capacity. Similarly, variable rate nitrogen for corn or wheat can be applied to areas with differing nutrient deficiencies. A GIS also enables farmers to make quick informed decisions regarding crop

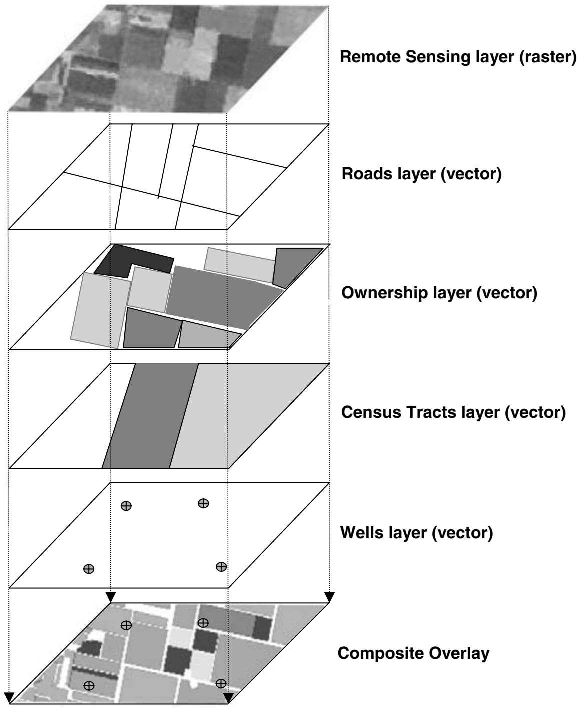


Figure 4: An integrated geographic information systems database.

stress due to disease (fungus, etc.), pest infestation, or water or salinity stress.

Today farmers have greater access to GIS data such as soil type, satellite and aircraft imagery, topography and geomorphology, and moisture condition. Farmers can access the Internet and obtain current satellite imagery, disease and pest infestations, and weather data to improve the management of their farms.

A farmer does not need to own a GIS to map his or her farm. Some companies have developed Internet map servers (IMS) to let farmers manage data. Currently, these IMSs do not provide the full capabilities of a GIS. In the near future, such software will allow the creation of farm maps with a minimum knowledge of GIS and without owning a GIS.

Archaeological and Geological Applications

GIS helps archaeologists examine archaeological sites interactively and model the sites for different scenarios using various human and natural activities that affected the sites. Some sites may include several historical events. GIS is a dynamic system and allows the construction of layers that represent these time periods. These layers can then be studied to reconstruct the history of the site.

GIS has many applications in geology. GIS provides the framework to input, develop, interpret, and analyze complex spatial and tabular datasets used for many mining and petroleum operations. Mineral and petroleum exploration, geological and hydrogeological mapping, karst modeling, earthquake and other natural disaster mapping, and resource assessments are just a few examples of applications in geology. The USGS (http://www.usgs.gov) provides many examples of GIS in geosciences.

A GIS is a powerful tool when integrated with other systems such as GPS technology for geological applications. In the past, geological mapping was labor intensive and time-consuming. The use of GPS with GIS and other technologies such as remote sensing has helped geologists to be more efficient and accurate in mapping as well.

Education and Research Applications

GIS has been one of the fastest growing technologies used in classrooms and research. A GIS is capable of performing analyses of geospatial or attribute data and provides powerful tools for almost any academic discipline. Libraries that provide tools for exploring information and museums that have systems that analyze historical and spatial data use dynamic GIS images. GIS helps students and teachers learn geography-related issues in an interactive way.

The software, hardware, and spatial data required for GIS analyses are readily available to K–12 schools and to universities through the Internet. The traditionally complex character of GIS is no longer a problem because there are several i-GIS application modules available for free or at a low cost, allowing K–12 students to access and process free data over the Internet. These modules minimize the initial learning curve and allow the use of a customized local data set. As a result, students can gain fundamental experience in the analysis of spatial data with the help of remote sensing, GPS, desktop GIS, and Internet-based mapping.

Another application of GIS is in educational administration. Several school districts have developed school attendance decision support systems designed to manage a rapidly growing student population. Students are assigned a specific school based on the location of their home, for example.

Many disciplines and technologies contribute to research in GIS. Software or system designers and engineers, database managers, geographers, cartographers, and scientists create software and algorithms that are used in many GISs. Research universities and private companies have developed GISs and GIS modules that provide new functions for existing information systems. Spatial decision support system (SDSS) models that can be used easily without extensive training have also been developed. There are many differences between GIS and SDSS. GIS sites usually require technical analysts to set up databases and use complex functions, whereas SDSS sites generally require one or two business analysts who want to make spatial decisions with limited GIS capabilities (Daniel, 1992). Some of these models are available on various Web sites.

Several GIS models can be purchased from GIS companies, but other models are available free of charge via the Internet. Some of the models and modules available on the Internet are related to hydrology and soils (e.g., SWAT, 2001) archaeological (e.g., Mn/Model, 1997), and emergency response (E-911). More research is being done to integrate traditional GIS with the Internet to provide data more rapidly. As stated earlier, IMS is one of the current technologies integrating GIS and the Internet.

Emergency 911 Systems and Health Applications

GIS allows 911 operators to access many databases quickly and efficiently, such as the transportation network, housing, the nearest water source, and location information for the nearest hospitals or other medical facilities. When integrated with GPS technology, a GIS allows 911 operators to track emergency vehicle locations and determine the best route. An emergency management GIS system can find the closest response units to the location of an incident and direct them to the nearest emergency medical facility saving critical time by analyzing traffic conditions and other factors.

Several health organizations and federal agencies such as the National Institutes of Health (NIH) and the National Cancer Institute (NCI) provide statistics, epidemiology, and cancer information and data over the Internet. NCI has created a cancer mortality map and graph Web site (http://www3.cancer.gov/atlasplus/), which provides interactive maps, graphs, and tables showing geographic patterns of cancer death rates for more than 40 cancers for the 1950–1994 time period. The cancer data provided at the Web site have been used to analyze spatial relationships between cancer mortality rates and environment using GIS (Ozkirim & Cetin, 2001).

Environmental Management, Forestry, and Conservation Applications

GIS has been adapted and used to assist government agencies, companies, and universities in natural resource management, environmental impact assessment, waste management, site remediation, and forest and fire management. GIS provides natural resource managers powerful analysis and decision-making tools for long-term forecasting. Diverse ecosystems can be analyzed and modeled by combining spatial, spectral, and temporal data through visualization analysis. Foresters, for example, can make long-term forecasts for forest inventory, wildlife habitat management, and fire prevention and management. GIS has been adapted by many coastal and marine researchers to create and manage large coastal ecosystem databases.

The U.S. Environmental Protection Agency (EPA) focuses on assisting businesses and communities with compliance training and guidance and is responsible for enforcing and ensuring compliance with environmental regulations. EPA has created many environmental data sets and uses GIS to regulate and monitor emissions from power plants, factories, and motor vehicles. The EnviroMapper Web site, for example, is one of EPA's Internet sites providing various types of environmental information, including air releases, drinking water, toxic releases, hazardous wastes, water discharge permits, and Superfund sites (http://www.epa.gov/enviro/html/em/). Users can create maps (dynamically) at the national, state, and county levels and link them to environmental text reports.

GIS is also used to monitor environmental problems in watersheds such as chemical or oil spills and wastewater. GIS is a powerful tool for understanding environmental issues and changes.

Remotely sensed data (Figure 3) have become a primary input to GISs used to model the earth and its resources at local, regional, and global scales. Large-scale monitoring and mapping has become possible with the advent of new high-spectral, spatial, and temporal resolution satellite sensors. The integration of remotely sensed imagery with other GIS data allows a synergistic processing of multisource spatial and spectral data for environmental applications.

The United Nations Environment Program has established a GIS-based interactive environmental Web site (http://www.unep.net) providing worldwide environmental data and a forum for technical peer review. The site can be used to create environmental maps with statistical and other up-to-date attribute information for any location or region in the world.

The USGS provides many GIS data including digital raster graphics (topographic maps in digital form), digital orthophotos, DLGs, and Digital Elevation Models that are used in environmental applications. The USGS has created several Web sites to query the data over the Internet (i.e., http://edc.usgs.gov/ and http://edcsns17.cr.usgs.gov/EarthExplorer/).

Federal, State, and Local Government Applications

Many federal agencies use GIS to create national and worldwide databases. The USGS is one of the agencies that pioneered its use. The USGS has developed many GIS formats and data sets that are used by the public, by industry, and by federal, state and local governments. One of the enhanced USGS sites, SDDS uses an IMS to serve the National Elevation Dataset (NED) at http://seamless.usgs.gov/viewer.htm. The U.S. Bureau of Census has created several data formats and data sets such as census data and the TIGER digital data. The Federal Emergency Management Agency (FEMA) has also created many GIS data sets and modules such as flood risk maps, National Hazard Loss Estimation Methodology (HAZUS), and other emergency-related data sets. FEMA has established an Internet-based GIS "Flood Map Store" to allow the public to access flood maps.

State and local governments use GIS for regional and urban planning and development, policy analysis, cultural and resource management, social and demographic studies, transportation, education, environmental health, assessment and waste management, and public utilities. For example, GIS is used to incorporate up-to-date data and sometimes real-time information for public utilities and wastewater and storm-water management.

There are numerous local government applications of GIS including emergency response and management, facilities or properties mapping, site location, building and site permits, land use, land regulation, street planning, election management, and emergency transportation routing. For example, law enforcement and public safety agencies collect large amounts of data including those regarding crimes, citations, arrests, field interviews, chronology, and other law enforcement information. Law enforcement agencies use GIS to create crime location and trend maps. It provides a powerful decision-making tool for police and federal and private investigators by allowing them to analyze crime data graphically in a spatial context. A GIS also allows government lawyers and prosecutors to improve case preparation and presentation and create interactive maps and models for cases including spatial evidence data linked to images, documents, witnesses, and other evidences.

Commercial Applications

Most business data are spatially georeferenced. GIS can be used to determine a business location based on crime data, environmental factors, location of customers and competitors, and transportation infrastructure. GIS can also be used to manage and analyze information about sales, customer (demographic) profiles, markets, distribution networks, inventories, status of processing facilities, service boundaries, and delivery routes. For example, GIS can assist the food processing industry by providing accurate spatial and attribute information about raw materials and inventories and distribution networks.

Companies providing pesticides, fertilizers, seeds, equipment, and services to farmers can benefit from a GIS by analyzing short- or long-term trends and strategies for product distributions. They can use GIS to monitor customer demand based on conditions such as weather to deliver products efficiently.

GIS technology enables pipeline companies and operators to integrate geospatial-based data to monitor and manage pipelines with efficiency and integrity. Applications include demographic analysis, marketing, customer segments, and optimal site locations for pipes and pumping stations. GIS technology enables pipeline companies to integrate geography-based data with other data such as customer, operations, and marketing for planning and problem-solving purposes. Many pipeline companies use GIS to monitor, analyze and evaluate pipeline performances. Pipeline companies can reduce design costs for locating potential pumping stations and pipelines by analyzing spatial data, geology, soils, land ownership, customer, topography, and the transportation network. The companies can also optimize the routing of service vehicles to save time and cost.

Similarly, GIS technology enables electric power companies to monitor, analyze and evaluate power grids. They can use GIS to integrate spatial data with operations, customer demand, and other data to optimize services to customers. When there is a high demand for electricity, companies can analyze and find the least expensive source of electricity from other power companies. Most power companies also use GIS to locate new potential power grids and towers.

Factories and public service companies utilize GIS to monitor raw material input, production, and output. GIS allows companies to analyze production to make long-term predictions. System designers use GIS because its modules or user interfaces can be changed at a low cost compared with factory machinery.

Military and Defense Applications

Defense organizations and military forces use GIS to improve decision support systems. Most military aircraft and ships use GIS and related technologies such as GPS and artificial intelligence systems. At a command control, most decisions are geographical (spatial). A military commander can make up-to-date operational, logistical, tactical, or administrative decisions with a GIS. In the past, most battle decisions were based on paper maps and pushpins. A GIS can provide rapid access to up-to-date spatial data and tools to field commanders or command controls to make quick decisions and can be used to create virtual battlefields. A GIS can also be used to provide logistics to troops or transportation units and handle base management. The use of spatial data standards with a GIS provides interoperability among many branches of military and contractors providing support to the military.

GIS is becoming an integral component in the intelligence field. Up-to-date information, which can be easily provided by a GIS, is important for the field. The CIA was one of the first agencies to use GIS. The National Imagery and Mapping Agency aids the intelligence community by providing many types of geospatial data and maps to the military using GIS.

Real Estate, Financial, Insurance and Business Applications

GIS allows realtors to build spatial databases so that customers can access real estate data quickly. Customers can visit several places in a short period of time using virtual GIS by displaying maps or aerial images of property locations, property photos, and videos. Customers can query data layers including proximity to schools, libraries, fire stations, police stations, highways, dump and toxic waste sites, mines, parks, entertainment facilities, or shopping centers. Title companies, among the first GIS users in the business sector, use GIS to search titles, access parcel ownership data, and make maps in a short period of time. Many title companies, multiple listing services, and the real estate industry have begun to implement Internet-based GIS marketing tools to provide free public access to ownership, locational, and physical data.

Financial institutions use GIS to analyze customers' financial behavior and needs. GIS allows the use of many databases including demographic, real estate, transportation, and other spatial data to establish market share, marketing targets, market penetration, and advertisement strategies. GIS can provide companies cost-effective ways to examine purchasing, travel, and other customer habits. Many investment companies are adopting GIS, which is well suited to analyze competitor information, visualize market situations, and analyze portfolios. GIS is used to find better locations for new bank branches and to analyze how these new branches affect existing branch locations. Many banks and credit card companies use i-GIS or IMS to provide customers with locator tools and interactive maps showing the nearest ATM or branch and the best route to an ATM location.

GIS allows surveyors to improve their tasks and to access many diverse data. Most surveyors are moving from computer aided design–based (CAD) systems to GIS, which provides tools to manage and analyze diverse data and integrate with GPS technology.

Journalists can access GIS and use demographic (census) data, crime locations, crime statistics, and traffic accidents to map many complex scenarios quickly and accurately. GIS provides reporters with more powerful tools, faster data access, and more dynamic mapping capabilities to complete tasks in a short period of time. Reporters can download the data using wireless connections to the Internet at the scene or before they arrive. With the availability of more powerful i-GIS, journalists will be able to analyze, model, and map areas of interest more efficiently in the field.

Several insurance companies use GIS and diverse data such as demographics, proximity to fire stations, and crime and environmental data to analyze property locations and surrounding areas for making accurate predictions and quick decisions.

Telecommunications and Transportation Applications

GIS applications in telecommunications include multiple market and demographic analysis, target marketing, customer segments for communication products and optimal site locations for cable and cellular towers. GIS technology enables telecommunication companies to integrate geospatial data with other data such as customers, operations, and marketing for planning and problem-solving purposes. Many local and long-distance telephone companies use GIS to monitor, evaluate, and analyze network performance and problems. Telecommunications companies can manage and route service vehicles using GIS, optimizing the routing of service vehicles to save time and cost. Call center operators can use GIS to access information about customers and the status of their networks such as equipment or signal quality. GIS provides support tools that can work in a complex environment, which many telecommunications companies require. Many cellular companies use spatial data (land ownership, customer locations, topography, and transportation network) to locate potential antenna sites to reduce design costs.

Airlines and bus companies can map everything from customer data to traffic statistics. Based on weather patterns, an airline company can use GIS to rearrange flights, match customers to particular airplanes, and track the status of airplanes. Bus companies can find the best routes by integrating GIS technology with GPS. By using GPS technology, the companies can track their fleet and analyze customer demand to use their buses more efficiently. Bus drivers can use GPS to find locations. GPS can be useful in rural areas where road signs may not be adequate.

GIS can be used to create worldwide transportation databases to handle daily changes such as monitoring rail systems and road conditions, redirecting ground and air traffic (intelligent transportation system), and maintaining transportation networks. GIS has been successfully used by railroad companies to handle commodity flow analysis, facilities management, emergency response and risk management, capacity planning, and passenger information. Complex road maps can be created using a GIS. Many companies and agencies use GIS for road construction, design, operations, and maintenance.

Integration of GIS technology with GPS allows trucking companies to reach locations quickly. Many companies using the two systems track fleet vehicles and find the best route to deliver goods and services for optimum deployment and cost savings. Moreover, these companies also use GIS for depot and warehouse operations.

Public Participation GIS (PPGIS)

A National Center for Geographic Information and Analysis (NCGIA) workshop was held in Orono, Maine, to explore public participation in GIS (PPGIS) in 1996. The workshop defined the characteristics of an alternative GIS called GIS/2 (http://ncgia.spatial.maine.edu/ppgis/ppgishom.html). In 1998, a specialist meeting called Empowerment, Marginalization and Public Participation GIS (PPGIS) was held in Santa Barbara, California (http://www.ncgia.ucsb.edu/varenius/ppgis/ncgia.html). The PPGIS term was used to include topics related to community interests and GIS technology. This initiative included several topics such as implementations of PPGIS, effects of the use of GIS on local politics, GIS in local surveillance, and the impacts and implications of spatial information on communities. PPGIS sessions were held at the 96th Annual Association of American Geographers (AAG) meeting in Pittsburgh in 2000.

The first "Annual Public Participation GIS Conference" was organized by URISA and held at Rutgers University, New Brunswick, New Jersey, in July 2002 (http://www.urisa.org/ppgis.htm). The conference brought together participants that included citizens and citizens groups, public officials, administrators, technicians, planners, librarians, policy scientists, and researchers. Several issues, including the value of GIS technology in empowering citizen organizations and communities, urban neighborhoods, indigenous peoples, developing nations, environmental organizations, and virtual communities, were discussed at the meeting.

PPGIS is a distributed, Internet-based GIS and has enabled many communities to access and distribute local knowledge among the community. PPGIS is intended for a broad audience of students, academics, planners, policy makers, and GIS practitioners. Several PPGIS projects have been implemented within the context of an academic debate over GIS and society (Weiner, Harris, & Craig, 2001). Users of community GISs do not require any software beyond an ordinary Web-browser, making the involvement of the community all the more possible.

DESKTOP GIS VERSUS INTERNET OR INTRANET GIS

Although desktop GIS has been a choice for many institutions requiring analysis of complex geographic databases, it has many limitations including cost and public access issues. Before the introduction of i-GIS, many companies, educational institutions, and local, state, and federal agencies started to use intranet GIS to overcome some of the limitations of a traditional GIS; however, an intranet-based GIS does not usually allow full public access. On the other hand, i-GIS has many advantages, including decentralized data storage and data maintenance, wider data access, and convenience. Web-based systems offer many opportunities to go beyond the home or office. Existing i-GIS map servers do not facilitate analysis of complex geographic databases, however, because some technological (e.g., software and bandwidth and speed) issues and specific user requirements. Today's Web based IMS systems offer data browsing, retrieving, updating, displaying and querying, mapmaking, address matching, and route finding. Most of them and probably new ones will be able to facilitate analysis of complex geographic databases in the near future when some of the technological limitations that exist today are overcome.

GIS DATA USERS AND THE INTERNET

GIS has been used at local, state, and national levels. The power of a GIS is highly correlated with its accessibility. A desktop GIS is mostly used by experts. Therefore, a desktop system is generally limited to a smaller group of people. A Web-based GIS can be accessed by millions of people using the Internet. Most people do not need to know all the aspects of GIS. Many probably do not realize that they are using an i-GIS (e.g., a map server such as MapQuest [http://www.mapquest.com/] or MapBlast! [http://www.mapblast.com/] providing maps and driving directions) when they request a best route for a destination.

Most GIS data are dynamic, therefore regular updating of data is required in most instances. When a user desires a best route, a map server should not contain outdated data. If a bridge is out, a GIS cannot generate alternate routing unless its database is updated promptly.

During the early days of GIS, data were created and used predominantly by federal agencies. The U.S. Geological Survey (http://www.usgs.gov/), the U.S. Census Bureau (http://www.census.gov/), the Central Intelligence Agency (http://www.cia.gov/), and the National Imagery and Mapping Agency (http://www.nima.mil/) have generated many data formats and databases. During the 1970s and 1980s, GIS data were mostly available on magnetic tapes or floppy disks, which were delivered in days or weeks and were costly. The Internet was used to transfer data to private users, but during the late 1980s and early 1990s the main users of GIS data and the Internet were government agencies and research universities. With the introduction of the Web in the early 1990s, many institutions started to provide data to the public over the Internet. During the past decade, local and state governments, as well as the private sector, helped create many datasets that were available to the public. GIS became a part of our daily life and a multibillion dollar industry in the late 1990s. Most of the GIS data were provided free via the Internet. The USGS stopped serving some of the data that were originally provided from its Web site and transferred some datasets to private companies. These companies provided GIS data free or charged minimal fees for faster service (i.e., faster Internet connections or delivery) because their sites were financially supported by advertisements. When advertising companies started to cut their advertisement spending, GIS data providers started to charge more for their services.

INTERNET VERSUS INTRANET GIS

The term intranet is usually used to describe applications and protocols of the Internet used to share and move information within an organization's boundaries. Intranet systems have been used to serve and share GIS data within institutions because data could not be made public because of security issues and other technical limitations of the Internet. Many large corporations have Intranets for their employees. Intranets may be connected to the outside world via a firewall, which allows a protected gateway between an organization's internal network and the Internet. An Intranet typically has several advantages over the Internet:

- *Secure connections:* Intranet provides private internal networks such as local area network (LAN) or wide area network (WAN), which are protected from outside Internet users by a firewall.
- *Easier control:* Intranet is much easier to handle in terms of protocols and connections.
- *Higher data transfer speed per unit (bandwidth) cost:* Intranet provides higher and broader bandwidth per unit cost than the Internet.

For large companies that have offices in various cities or states, Intranet services may not be sufficient. Several network providers offer virtual private network (VPN) services, which are used on the public or open Internet, to companies that have more than one business location. These companies use point-to-point tunneling protocol (PPTP) to make secure connections between VPN nodes. Because the Internet is an "unsecure" open network, the PPTP is used to transmit data from one VPN node to another securely.

As stated previously, GIS data are dynamic and must be updated regularly. Several county and city agencies started to update and share their data with local and state agencies via the Internet in the mid-to-late 1990s. To use a GIS over the Internet, RDBMS with object-oriented extensions, software systems with cross-platform portability, client-server architecture, and other technologies had to be developed.

With the growth of the Internet, the distribution and viewing of data online are now an integral part of many projects. The integration of GIS with the Internet is an inevitable trend that is becoming a reality. Several companies have developed ActiveX-based distributed applications in addition to the standard IMS-based systems. The ActiveX-based systems are similar to the IMS-based systems but use a Web browser. A user connects to a Web server having "desktop" GIS software using a browser and loads the ActiveX component to run a GIS mapping application software remotely. Using the ActiveX component has many advantages including the use of data on a local computer without using remote disk mount utilities, which can be difficult to maintain. This technique is ideal for local or state governments that may require quick access to data (i.e., ownership or parcel data).

Desktop GISs are not appropriate for modern distributed network environments because of their closed architecture. Distributed GIS services can provide many

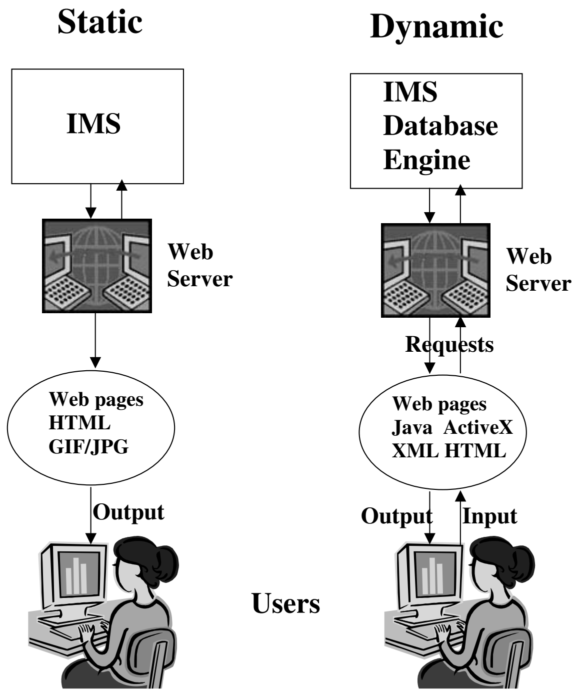


Figure 5: Static and dynamic Internet Map Servers.

capabilities and functions for data storage and management (Brown, 1999; Huang, Jiang, & Li, 2001; Huang & Worboys, 2001; Tsou & Buttenfield, 2002). In addition to the ActiveX-based systems, there are also other technologies, such as Java/VRML (virtual reality modeling language), common gateway interface (CGI), extensible markup language (XML), .NET, and other services used to support the development of i-GIS.

INTERNET MAP SERVERS

It is expensive to maintain and update large GIS databases. One of the goals of IMSs is to access large spatial databases in real time and to provide the data to the public. Most IMSs have been used to browse, display, query, and retrieve spatial and attribute data. Some IMSs can also be used to update GIS data. There are several commercially available IMS software packages (see references for their URL), which offer several GIS functions including routing, regional, and local- or street-level mapping, and simple querying and analysis (Jankowski, Stasik & Jankowska, 2001; Limp, 2001, 2002).

IMSs can be classified into two major groups, static and dynamic (Figure 5). A static IMS provides maps in GIF, JPEG, or other image formats. Map images (usually in the form of a digital raster graphics in TIFF format) are created using desktop GIS software or scanned from existing hard-copy maps in advance and served over the Internet using a Web server and basic Web browser formats (e.g., HTML code). Dynamic systems usually update map images based on users' requests or based on current data to be made available to the users (Figure 6). Only dynamic IMSs provide the functionality of local standalone GIS databases. Some dynamic systems allow users to create their own maps or perform limited geographic analysis. Several local and state governments provide data including parcel tax data, current conditions (roads, crime events, and patterns), environmental conditions such as lake or river levels and fire information, and other maps to the public using IMSs (Table 1).

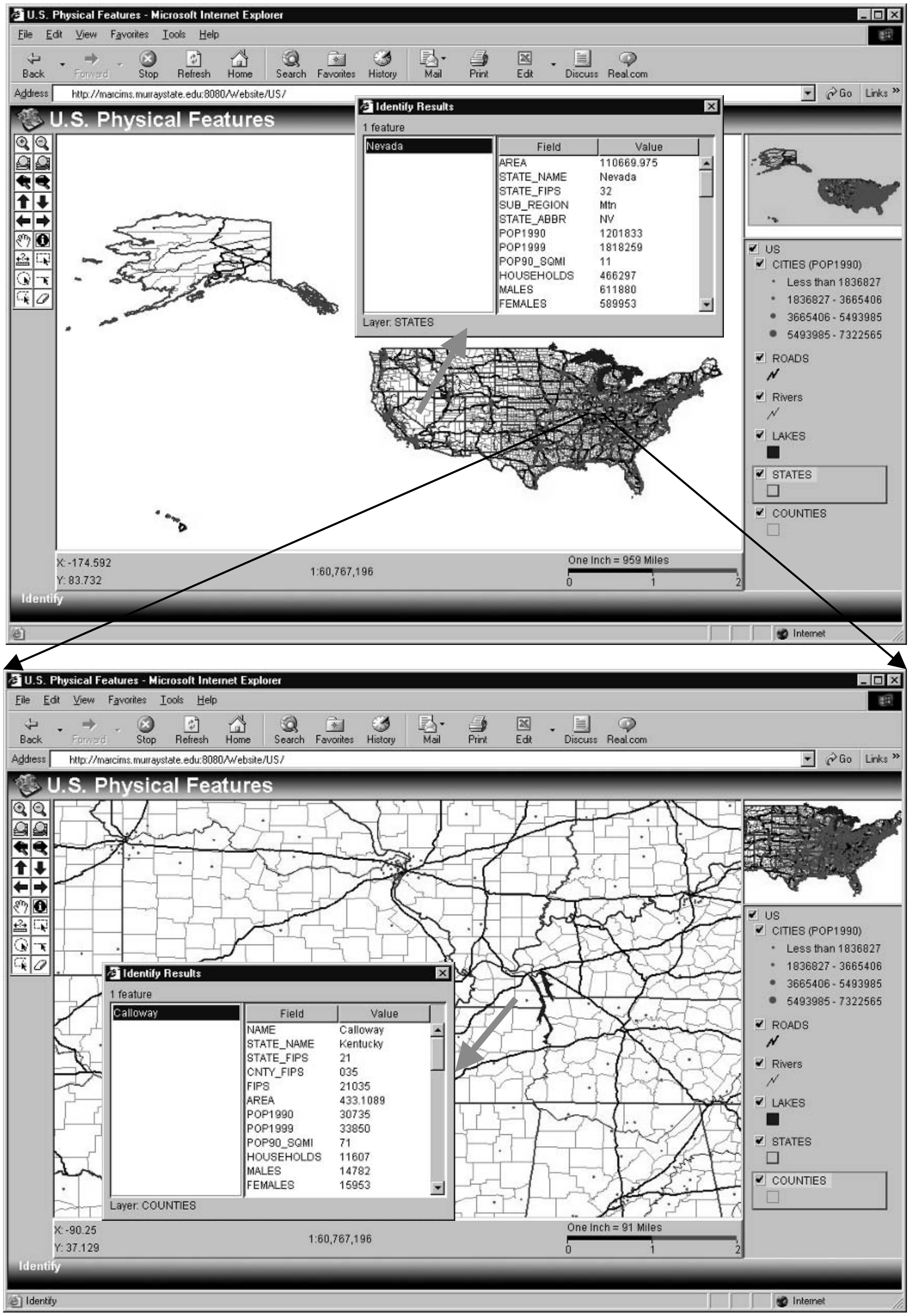
Today there are many commercial Web-based IMS systems such as AltaMap (GeoMicro), ArcIMS, ArcView Internet Server, Autodesk MapGuide, BeyondGeo, Demis Map Server, EarthKey, FreshMaps, GEO-Data Explorer (GEODE), GenaWare, GeoServ.org, InterroMAP, MapInfo MapXtreme, MapObjects IMS, Maptitude for the Web, Map-TV, Oracle, VectorEyes, Web Mapper, and WebView. Some IMSs are available free of charge: ALOV Map, GeoTools, Imapper, Jshape, MapCiti, MapIt!, and Mapserver (see their URLs in the Further Reading section). Currently, most of these IMSs have some technological limitations. In the future, new IMSs will be able to overcome some or most limitations to facilitate analysis of complex geographic databases remotely via the Internet.

THE FUTURE OF INTERNET GIS (I-GIS)

One of the difficulties of offering distance-learning courses, particularly IT courses via the Internet, is the availability of software used over the Internet (e.g., the ability to use software like Excel or Wordperfect on a different (remote) computer using a Web browser). Much research is being conducted to integrate desktop and Internet-based software packages. Geographic information science (GISc) is one of the fields that will benefit the most.

Integration of desktop GIS and i-GIS is still an evolving issue. More and more companies have offered i-GIS packages over the last several years, and shareware software packages are being circulated. These new developments will bring several other issues including data and computer security, meta-data, and data format issues to the forefront.

A standard language for i-GIS is necessary to have systems compatible across the Internet. The Open GIS Consortium (OGC), an international industry consortium of several hundred companies, government agencies, and universities participating in a consensus process, has been established to develop publicly available geoprocessing specifications and to deliver spatial interface specifications that are openly available for global use (http://www.opengis.org/). OGC's OpenGIS specifications for interfaces and protocols, the foundation for "interoperable" geoprocessing, have become widely used. Development of XML in 1996, which has been a W3C standard since February 1998, provided a tool for OGC to develop a geography markup language (GML), a new common geospatial data format. OGC specified GML, an XML extension for encoding the transport and storage of geographic information, including both the geometry and properties of geographic features in 1999. GML, which is based on OGC's abstract model of geography, is becoming a standard for i-GIS. The GML 2.5 version, which was approved in September 2001, incorporated various schema revisions. The latest version, GML 3.0, is currently in progress.



Identify Results

1 feature

Nevada

Field	Value
AREA	110669.975
STATE_NAME	Nevada
STATE_FIPS	32
SUB_REGION	Mtn
STATE_ABBR	NV
POP1990	1201833
POP1999	1818259
POP90_SQMI	11
HOUSEHOLDS	466297
MALES	611880
FEMALES	589953

Layer: STATES

Identify Results

1 feature

Calloway

Field	Value
NAME	Calloway
STATE_NAME	Kentucky
STATE_FIPS	21
CNTY_FIPS	035
FIPS	21035
AREA	433.1089
POP1990	30735
POP1999	33850
POP90_SQMI	71
HOUSEHOLDS	11607
MALES	14782
FEMALES	15953

Layer: COUNTIES

Figure 6: An Internet map server (http://marcIMS.Murraystate.edu/Website/US/).

Table 1 Example Internet Map Servers (IMSs)

The Alexander County IMS	http://www.co.alexander.nc.us/gis/imswebpage/imspage.htm
California climate	http://maps.esri.com/climo/climograph.html
The CARES IMS	http://www.cares.missouri.edu/
The Cincinnati Area Geographic Information System (CAGIS)	http://cagis.hamilton-co.org/map/cagis.htm
The City of Lincoln/Lancaster County (Nebraska) IMS	http://ims.ci.lincoln.ne.us/
Digital Environmental Atlas of Georgia	http://csatims.gis.gatech.edu/Web site/atlas2/info.html
Geography Network	http://www.geographynetwork.com/
GISCafe	http://www.giscafe.com/GISCafe/JAVA/ESRIArc.html
The Indiana Geological Survey IMS	http://igs.indiana.edu/arcims/index.cfm
Iowa IMS	http://igic.gis.iastate.edu/iowaims/ims.html
The Jefferson County, Washington, IMS	http://www.co.jefferson.wa.us/idms/devtest/mapserver_saveme.shtml
The Mid-America Remote Sensing Center (MARC) IMS	http://marcIMS.Murraystate.edu
The Montgomery County, Maryland IMS	http://gis.co.mo.md.us/gis/maproom.asp
The New Mexico IMS	http://geoinfo.nmt.edu/data/ims/home.html
The Pinellas County Government IMS	http://pubgis.co.pinellas.fl.us/maphelp.cfm
The Rhode Island Commercial and Recreational Fisheries interactive maps IMS	http://www.edc.uri.edu/fish/imsmaps.html
State and local government agencies using IMSs in Kentucky	http://kygeonet.state.ky.us/ims.htm
The State of Arizona IMS	http://rangeview.arizona.edu/tools/inet_map_server.html
The USGS Western Earth Surface Processes	http://kaibab.wr.usgs.gov/
The WOGRA Mapper	http://wogra.wygisc.uwyo.edu/misc/wograindex.html

Note. All sites retrieved October 22, 2002.

Internet-GIS is becoming one of the most rapidly evolving fields in e-commerce. GIS applications in wireless e-commerce integrating GPS and GIS technologies to direct customers to the nearest locations such as restaurants, gas stations, and hotels and to track vehicles for automatic vehicle location are now available. For institutions, i-GIS will provide easy access to GIS data without duplication (decentralized database). For developers, i-GIS will facilitate a new challenge as well as an opportunity to increase their market share. Users will become proficient at bringing GIS to the Web and become experts in data import and export, query, manipulation, analysis, output, and synthesis. More important, users may become project managers who will realize the value of i-GIS for quick decision making.

Existing i-GIS map servers are presently not able to fully analyze complex geographic databases. In the near future, we will be able to use a GIS from wherever we are, at home or in a different city. Still, it is likely that a Web-based i-GIS will never be "complete." Upgrades, patches, updates, and changes will always be necessary. As our collective experience and understanding of the technology grows daily, we will never be satisfied with the current technology and will look for better alternatives.

CONCLUSION

Today GIS is being used in almost every field that emphasizes spatial data collection, analysis, manipulation and integration, analysis, and modeling. Desktop GIS has been the system of choice for many institutions requiring analysis of complex geographic databases. Nonetheless, it has many limitations including startup cost, learning curve, management difficulty, and public access issues. On the other hand, i-GIS, using decentralized data storage, provides easy access to GIS data without duplication. With the growth of the Internet, distribution and viewing of data are now an integral part of many projects. The integration of GIS with the Internet is still an evolving issue and an inevitable trend that is becoming a reality. Although today's i-GISs do not offer the full capabilities of a desktop GIS, partly because of software limitations and the bandwith and speed available to most Web users, future i-GISs will include automated GIS processing and will be a part of our daily lives.

ACKNOWLEDGMENTS

The author thanks Drs. Thomas C. Kind, Burl I. Naugle, George W. Kipphut, Peter W. Whaley, and Ms. Lynn V. Leasure for reviewing this manuscript. The U.S. physical map (Figure 6) was created by Mr. Kevin Sturgill.

GLOSSARY

ActiveX ActiveX, which includes both client and server technologies, is a set of technologies enabling interactive content for the World Wide Web.

Attribute Nonspatial information associated with a geospatial feature.

Digital line graph (DLG) A vector data format defined by the U.S. Geological Survey.

Digital Raster Graphics (DRG) A raster data format defined by the U.S. Geological Survey.

Geospatial data Information (derived from mapping, surveying, or remote sensing technologies) that identifies the geographic location and characteristics of natural or artificial features and boundaries on Earth.

Graphic interchange format (GIF) An image format.

Geography markup language (GML) A new common geospatial data format.

Global positioning system (GPS) A satellite navigation system providing specially coded satellite signals that can be processed using a GPS receiver to compute position. At least three GPS satellite signals are used to compute positions in two dimensions and four GPS satellite signals are needed to compute positions in three dimensions. In the United States, GPS is funded and controlled by the U.S. Department of Defense.

Hypertext markup language (HTML) A language used to create Web documents.

Internet map server (IMS) Internet-based mapping system providing maps online and a technology integrating GIS and the Internet.

Joint Photographic Experts Group (JPEG) An image format.

Pixel A two-dimensional picture element (cell). The smallest element of a digital image.

Raster An equally spaced, two-dimensional rectangular grid (a regular or nearly regular, tessellation of a surface) pattern of quantized sample values. It is often called an image-based data structure.

Spatial data transfer standard (SDTS) A spatial data format defined by the Department of Commerce in 1992 (Federal Information Processing Standard 173).

Tagged image file format (TIFF) An image format.

Uniform resource locator (URL) A unique address on the Internet.

Vector A mathematical term for a directed line segment. A data model using sets of coordinates and attributes to define discrete features, points, lines, or polygons.

Extensible markup language (XML) It describes a class of data objects called XML documents and the behavior of computer programs processing them. XML is designed to describe data, whereas HTML is designed to display data.

CROSS REFERENCES

See *ActiveX; File Types; HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); Internet Literacy; Intranets.*

REFERENCES

Aronoff, S. (1995). *Geographic information systems: A management perspective.* Ottawa, Canada: WDL.

Bolstad, P. (2002). GIS fundamentals: A first text on geographic information systems. White Bear Lake, MN: Eider Press.

Brown, I. (1999). Developing a virtual reality user interface (VRUI) for geographic information retrieval on the Internet. *Transactions in GIS, 3,* 207–220.

Chrisman, N. (1997). *Exploring geographic information systems.* New York: Wiley.

Daniel, L. (1992, December). SDSS for location planning, or the seat of the pants is out. *GeoInfo Systems.* Retrieved January 24, 2003, from http://www.colorado.edu/geography/gcraft/notes/gisapps/sdss.html

DeMers, M. N. (2000). *Fundamentals of geographic information systems* (2nd ed.). New York: Wiley.

FGDC Content Standard for Digital Geospatial Metadata (FGDC-STD-001—1998). Retrieved October 22, 2002, from http://www.fgdc.gov/metadata/metadata.html

The FGDC Metadata Entry System. Retrieved October 22, 2002, from http://130.11.52.178/metaover.html

Goodchild, M. F., & Kemp, K. K. (1992). Introduction to GIS. In M. F. Goodchild & K. K. Kemp (Eds.), *NCGIA Core Curriculum* (Units 4 and 13). Santa Barbara, CA: National Center for Geographic Information and Analysis, University of California.

Huang, B., Jiang, B., & Li H. (2001). An integration of GIS, virtual reality and the Internet for visualization, analysis and exploration of spatial data. *International Journal of Geographical Information Science, 15,* 439–456.

Huang, B., & Worboys, M. (2001). Dynamic modelling and visualization on the Internet. *Transactions in GIS, 5,* 131–139.

Jankowski, P., Stasik M., & Jankowska, M. A. (2001). A map browser for an Internet-based GIS data repository. *Transactions in GIS, 5,* 5–18.

Limp, W. F. (2001, Feburary). User needs drive web mapping product selection. *GEOWorld,* 8–16.

Limp, W. F. (2002, March). Web 2002: Commercial improvements, new Web services and interoperability initiatives make for interesting times. *GEOWorld,* 30–32.

Mn/Model (1997, July). Retrieved January 24, 2003, from http://gis.esri.com/library/userconf/proc97/proc97/to200/pap151/p151.htm

The NSDI/FGDC National Geospatial Data Clearinghouse Information Resource Page (1998). Retrieved January 24, 2003, from http://www.fgdc.gov/clearinghouse/clearinghouse.html

Ozkirim, F., & Cetin, H. (2001, April). *A study of spatial relationships between cancer mortality rates and atmospheric emissions using geographic information systems.* Presented at the American Society for Photogrammetry and Remote Sensing Conference, St. Louis, Missouri.

Plewe, B. (1997). *GIS online: Information retrieval, mapping, and the Internet.* Santa Fe, New Mexico: OnWord Press.

SWAT (2001, July). Retrieved January 24, 2003, from http://www.brc.tamus.edu/swat/

Tsou, M. H., & Buttenfield, B. P. (2002). A dynamic architecture for distributed geographic information services. *Transactions in GIS, 6,* 355–381.

Weiner, D., Harris, & Craig, W. J. (2001, December). Community participation and GIS. Workshop on Access and Participatory Approaches in Using Geographic information, Spoleto, Italy. Retrieved October 26, 2002, from http://www.spatial.maine.edu/~onsrud/Spoleto/WeinerEtAl.pdf

Xue, Y., Cracknell, A. P., & Guo, H. D. (2002). Telegeoprocessing: The integration of remote sensing, geographic information system (GIS), global positioning system (GPS), and telecommunication. *International Journal of Remote Sensing, 23,* 1851–1893.

FURTHER READING

ALOV Map. Retrieved October 22, 2002, from http://alov.org/index.html

AltaMap (GeoMicro). Retrieved October 22, 2002, from http://www.geomicro.com/index.htm

ArcView Internet Server, ArcIMS, and MapObjects Internet Map Server. Retrieved March 17, 2002, from http://www.esri.com

Autodesk MapGuide. Retrieved March 17, 2002, from http://usa.autodesk.com/

BeyondGeo. Retrieved October 22, 2002, from http://www.beyondgeo.com/

Demis Map Server; Retrieved October 22, 2002, from http://www.demis.nl/DEMIS_UK/Products/Demis%20Map%20Server.htm

EarthKey. Retrieved October 22, 2002, from http://www.softreality.com/EarthKey/Overview.shtml

FreshMaps. Retrieved October 22, 2002, from http://freshmaps.net/

GEO-Data Explorer (GEODE). Retrieved October 22, 2002, from http://geode.usgs.gov/

GenaWare. Retrieved October 22, 2002 from http://www.genaware.com/products/genaserver/

GeoServ.org. Retrieved from http://www.geoserv.org/

GeoTools. Retrieved October 22, 2002, from http://www.ccg.leeds.ac.uk/geotools/underConstruction/

IMapper. Retrieved October 22, 2002, from http://www.imapper.com/

InterroMAP. Retrieved October 22, 2002, from http://www.interromap.com/

Jshape. Retrieved October 22, 2002, from http://www.jshape.com/

MapCiti. Retrieved October 22, 2002, from http://www.rockware.com/catalog/pages/mapciti.html

MapGuide. Retrieved October 22, 2002, from http://www3.autodesk.com/

MapInfo MapXtreme. Retrieved March 17, 2002, from http://dynamo.mapinfo.com/

MapIt! Retrieved October 22, 2002, from http://www.mapit.de/index.en.html

Mapserver. Retrieved October 22, 2002, from http://mapserver.gis.umn.edu/

Maptitude for the Web. Retrieved October 22, 2002, from http://www.caliper.com/ovuwebpg.htm

Map-TV. Retrieved October 22, 2002, from http://www.spatialmedia.com/

Oracle. Retrieved March 17, 2002, from http://www.oracle.com

VectorEyes. Retrieved October 22, 2002, from http://www.fullcircletech.com/

Web Mapper. Retrieved October 22, 2002, from http://www.web-mapper.com/main.cfm

WebView. Retrieved October 22, 2002, from http://www.zebris.com/english/main_produkte.htm

Global Diffusion of the Internet

Nikhilesh Dholakia, *University of Rhode Island*
Ruby Roy Dholakia, *University of Rhode Island*
Nir Kshetri, *University of North Carolina*

INTRODUCTION

Since the 1970s, the number of countries connected to the Internet has increased steeply from 60 in 1993 to 214 in 2000 (see Figure 1). Despite this growth, the Internet has a highly asymmetric global distribution (see, e.g., Kshetri, 2001). There were 40 million people worldwide using the Internet regularly in 1995 (Media Metrix, 2000). This jumped to 131 million by the end of 1999 (Pastore, 2000) and to 606 million by September 2002. An estimate by the Angus Reid Group (2000) suggests that that there will be one billion Internet users by 2005.

The Internet is the fastest diffusing information and communication technology (ICT) innovation to date. For instance, it took just 10 years for the Web-based Internet to reach 50% of American homes, compared to 52 years taken by electricity and 71 years by the telephone (Thierer, 2000). It took only 3 years for the Internet to reach 50 million users. By contrast, it took 38 years for radio and 13 years for television to have 50 million users (Bell & Tang, 1998). During 1999, the number of Internet users increased by 1 million every month (McLaren, 1999).

The global distribution of Internet penetration and use, however, is far from uniform (see Figures 2 and 3). Developed countries account for a disproportionately high number of Internet users worldwide: 61.7% of the world's Internet users live in North America and Europe, for instance, but they account for less than 20% of the world's population. As of May 2002, less than 0.1% of Africans had access to the Internet compared to 52.2% of North Americans. Similarly, as of 2000, developing countries with 84% of the world's population had less than 6% of world's Internet users (Cyber citizenship gains in developing world, 2000). The discrepancy is even steeper for the number of Internet hosts worldwide: North America and Europe account for almost 90% of Internet hosts, whereas Asia and the Pacific, with over 60% of the world's population and 31% of the world's Internet users, contribute only 8% of the total number of Internet hosts worldwide.

There are, however, some encouraging signs. For instance, Internet users worldwide grew by 30% in 2001 and one-third of the new users were from developing countries (United Nations Conference on Trade and Development [UNCTAD], 2002).

Because the Internet is a "new product," an analysis of the pattern of its spread worldwide and within a given country from the perspective of the innovation diffusion and adoption literature could provide valuable insights into the factors driving the diffusion dynamics (Takacs & Freiden, 1998). The diffusion and adoption patterns of an innovation are functions of several elements including characteristics of the *innovation* itself, the *channel of communication*, the nature of the *social system*, and *time* (Rogers, 1983).

Because social systems around the world differ in several dimensions, diffusion patterns of the Internet vary widely across countries. In this chapter, some of the key differences in Internet diffusion patterns around the world are highlighted.

We examine the global diffusion of the Internet and the factors influencing its diffusion dynamics. The remainder of the chapter (a) provides a brief historical overview of global Internet diffusion, (b) discusses the global diffusion patterns of the Internet, (c) integrates theories from diverse perspectives that help explain the diffusion and adoption phenomena, (d) discusses the factors that are likely to shape the global diffusion of the Internet, and (e) provides some conclusions.

INTERNET DIFFUSION: HISTORICAL OVERVIEW

The Internet passed through various stages to arrive at the present situation. The first wide area network (WAN)

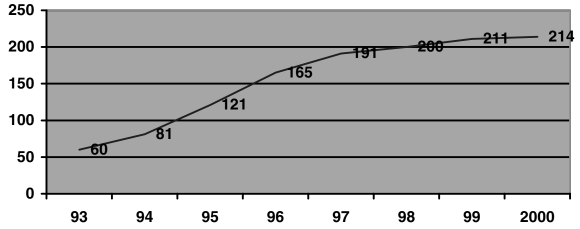


Figure 1: Number of countries connected to the Internet. Source: International Telecommunications Union (2001b).

was developed in 1965. It took another 4 years for the first two hosts in the ARPANET to be connected. The graphical format of the Internet as we know it now emerged in the 1990s.

Internet Diffusion Prior to 1980

Drawing primarily from Leiner et al. (2002), we trace the early history of Internet diffusion in this section. President Eisenhower's request for funds to create the Advanced Research Projects Agency (ARPA) within the U.S. Department of Defense in 1958 laid the foundation for the Internet. The packet switching theory first published by Leonard Kleinrock of MIT in 1961 was a major step toward computer networking. In August 1962, J.C.R. Licklider, also of MIT, proposed the "Galactic Network" concept. The Galactic Network concept provided a remarkably prescient view of the contemporary Internet—it envisioned a set of globally interconnected computers through which data and programs could be accessed from any site. In October 1962, Licklider became the head of the computer research program at the Defense Advanced Research Projects Agency (DARPA). He also convinced his successors at DARPA of the importance of the networking concept. Under a military contract, Paul Baran of Rand Corporation wrote key memoranda in the early 1960s outlining the "survivability" of a distributed, packet-switched network, even under conditions of a major nuclear attack. As Rand Corporation states in its historical notes

> All of the nodes in this unusual network would have equal status; be autonomous; and be capable of receiving, routing, and transmitting information. Under Baran's concept of distributed communications—now called packet switching—each message would be broken into a series of short, fixed-length pieces, and each would be sent as an individually addressed packet that would find its own way through the network by whatever route happened to become available, jumping from node to node until it reached the final destination. If parts of the network were destroyed, the self-sufficiency of each node plus the data within the packet allowed the node to seek alternative ways of moving the packet along. (RAND's History, n.d.)

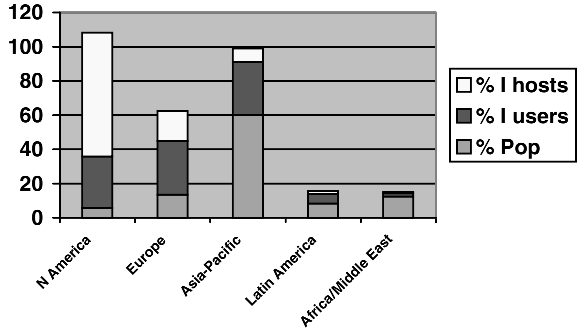


Figure 2: Uneven distribution of the Internet worldwide. Sources: Kshetri and Dholakia (2002), Nua Internet Surveys (2002), and authors' research.

Next, it was necessary to make the computers "talk" to each other. In 1965, MIT researchers Thomas Merrill and Lawrence G. Roberts were able to connect the TX-2 computer in Massachusetts to the Q-32 in California. This was the first wide area computer network. In 1966 Lawrence G. Roberts, who had joined DARPA, put together his plan for the ARPANET. Kleinrock's Network Measurement Center at UCLA was selected to be the first node on the ARPANET in September 1969. In October 1969, the Stanford Research Institute (SRI), about 350 miles to the north of UCLA, provided a second node and the first host-to-host message was sent from Kleinrock's laboratory to SRI. Two more nodes were added at the University of California at Santa Barbara and the University of Utah. Four host computers were connected together into the initial ARPANET, the forerunner of today's Internet, by the end of 1969. It is interesting to note that independent of the U.S. efforts, Donald Davies of the National Physical Laboratory (NPL) in U.K. proposed development of a nationwide packet communications network, based on a packet size of 1,024 bits—the same packet size that Paul Baran had proposed in the U.S.

The number of computers connected to the ARPANET grew rapidly after 1970 (see Table 1). In December 1970, the Network Working Group (NWG) finished the initial ARPANET host-to-host protocol, called the network control protocol (NCP). As the ARPANET sites completed implementing NCP during the period 1971–1972, the network users finally could begin to develop applications. In 1972, electronic mail was introduced. Bolt, Beranek, and Newman (BBN), a consulting firm started by MIT professors Richard Bolt and Leo Beranek and their student Robert Newman, sent the first person-to-person email using the @ symbol in the address. To meet the needs of an open-architecture network environment, Bob Kahn and Vint Cerf developed the transmission control protocol/Internet protocol (TCP/IP) in 1973. TCP/IP proved remarkably resilient and enduring as a means of managing data communication networks.

Starting from the mid-1970s, computer networks grew rapidly. ARPANET expanded internationally by connecting to the University College of London (England) and NORSAR (Norway) in 1973. Most of the early networks, however, were closed communities of scholars. There was virtually no pressure for the individual networks to be compatible and hence they were not.

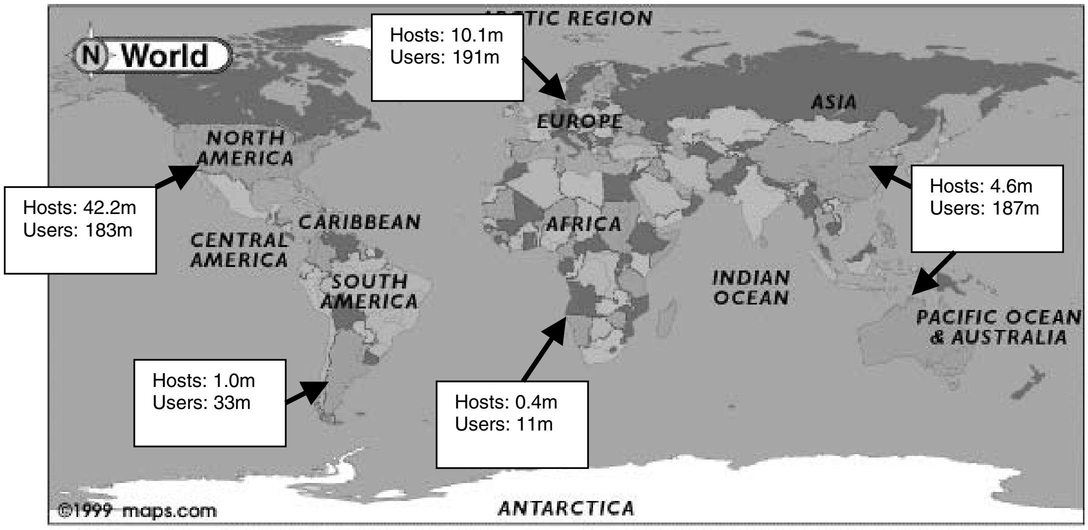


Figure 3: Global distribution of Internet hosts and Internet users. Note: Internet user data are for September 2002 and Internet host data are for the year 1999.

Internet Diffusion During 1980–1990

Various developments in the 1980s facilitated the diffusion of the Internet. The number of Internet hosts grew 500-fold in this period, from 200 in 1980 to 100,000 in 1989 (Table 1). In 1980, Tim Berners-Lee wrote the program known as "Enquire Within," which was a predecessor to the World Wide Web (WWW). Other developments include IBM's announcement of the first personal computer in 1981 and the foundation of Cisco Systems in 1983.

The ARPANET host protocol changed from NCP to TCP/IP on January 1, 1983, to meet the needs of an open-architecture network. In November of 1983, the domain name systems (DNS), such as .edu, .gov, .com, .mil, .org, .net, and .int were created. The first "dot-com" in the world was Symbolic.com. It became the first registered domain on March 1, 1985. In 1985, the National Science Foundation (NSF) also decided that TCP/IP would be mandatory for the NSFNET backbone.

Table 1 Growth in the Number of Internet Hosts

Year	No. of Hosts	No. of Hosts (in millions)	Remarks
1965	2	—	TX-2 and Q-32 connection.
1970	4	—	ARPANET project had a backbone of four computers.
1972		—	E-mail introduced
1973		—	TCP/IP developed
1975	100	—	
1980	200	—	"Enquire Within," predecessor of WWW, developed
1985	2,000	—	TCP/IP mandatory for NSFNET
1986	5,000	—	
1987	10,000	—	
1989	100,000	0.1	
1990	200,000	0.2	NSFNET succeeded ARPANET
1993	2,700,000	2.7	Mosaic Web Browser developed
1994	5,800,000	5.8	
1995	14,000,000	14.0	Windows 95
1996	22,000,000	22.0	
1997	30,000,000	30.0	
1998	43,000,000	43.0	
1999	72,000,000	72.0	
2000	104,000,000	104.0	

Source: International Telecommunications Union (2001b), Schwimmer (2002), Webopedia.com (2002), and authors' research.

Internet Diffusion After 1990

The Internet was opened to the public in the early 1990s. Major technological developments such as Tim Berners-Lee's creation of the World Wide Web (WWW), the development of the Mosaic Web browser in 1993, Sun Microsystems' release of Java, and the release of Windows 1995 further facilitated the diffusion of the Internet (Webopedia 2002). As a result, the Internet grew significantly in the 1990s, in terms of the number of hosts, number of users, and global coverage. The number of Internet hosts increased from 200,000 in 1990 to 104 million in 2000. By 1990, major countries across the world such as Australia (AU), Germany (DE), Israel (IL), Italy (IT), Japan (JP), Mexico (MX), the Netherlands (NL), New Zealand (NZ), Puerto Rico (PR), the United Kingdom (UK), Argentina (AR), Austria (AT), Belgium (BE), Brazil (BR), Chile (CL), Greece (GR), India (IN), Ireland (IE), Korea (KR), Spain (ES), and Switzerland (CH) had country-specific domains and were already connected to the NSFNET (Goldstein, 2000). Several smaller and less developed countries were gradually connected to the Internet.

Future Scenarios

Along with the exponential growth in the number of Internet users worldwide, newer means to access the Internet keep appearing. In particular, Internet access by mobile and broadband technologies is experiencing rapid growth worldwide. For instance, by 2009, cellular phone subscribers in the world are expected to outnumber fixed line subscribers (International Telecommunications Union, 2000). Furthermore, over 25% of e-commerce is predicted to take place over handheld sets by 2005 (Shaffer, 2000).

The number of broadband subscribers worldwide is estimated to exceed 46 million by the end of 2002 (Reports about the death of broadband are premature, 2002). The annual worldwide broadband growth rate for the period 2002–2004 is projected to be in the range of 61% to 150% (Lammers, 2001). Whereas cable modem service is gaining popularity in the U.S., digital subscriber line (DSL) has become the main broadband access technology outside the U.S. Other broadband access technologies such as satellite broadband, fiber-to-the-home, and fixed wireless service accounted for 5% of the worldwide broadband market in 2002 (Study says broadband growth remains robust, 2002).

GLOBAL DIFFUSION OF THE INTERNET

Starting with just a few countries in 1990, the number of countries connected to the Internet crossed 200 by mid-1998 (World Intellectual Property Organization, 2000) and 214 in 2000. The diffusion pattern of the Internet, however, varies widely across the world. Table 2 details the global distribution of the number of Internet users in September 2002. North America, where the Internet originated, has over 182 million users, accounting for 30% of the world's users but only 5.6% of the world's population. Europe, with its early connection to the ARPANET, currently has the highest number of Internet users (more than 190 million). Asia and the Pacific, with over 60% of the world's population, are experiencing some of the most rapid growth in recent years and are expected to double their number of Internet users by 2005.

Internet Diffusion in Africa and the Middle East

In the Middle East region, Israel was a pioneer in connecting to the Internet—in 1989, along with selected key European and Asian nations. African and other Middle Eastern nations came to the Net much later. By

Table 2 Geographical Distribution of Internet Users Worldwide (September 2002)

World Region (% of World Population)	Number of Internet Users (% of World Users)	Remarks
North America (5.6% of world population)	182.67 million (30.2% of the total Internet users in the world)	Enjoys many advantages. The place of origin of most Internet technologies.
Europe (13.5% of world population)	190.91 million (31.5% of the total Internet users in the world)	Early connection to ARPANET proved to be of great value.
Asia-Pacific (60.2% of world population)	187.24 million (30.9% of the total Internet users in the world)	Some of the most rapidly growing Internet markets located here.
Latin America (8.3% of world population)	33.35 million (5.5% of the total Internet users in the world)	Except Mexico, Internet entered late, but is growing very fast.
Africa/Middle East (12.4% of world population)	11.43 million (1.9% of the total Internet users in the world)	Israel was a pioneer. Sub-Saharan Africa was very late. South Africa dominates.
World (Total)	605.60 million	Expected to reach a billion or more by the end of the decade.

Source: Kshetri and Dholakia (2002), Nua Internet Surveys (2002), and authors' research.

Table 3 Internet Users (IU) and Internet Hosts (IH) in Africa and the Middle East (1999)

Country	IU (1999)	IU per 1000 People (1999)	IH (1999)	IH per 1000 People (1999)
Algeria	4,000	4.0	158	0.005
Angola	15,400	1.2	12	0.0009
Bahrain	36,000	59.4	866	1.4
Botswana	30,000	18.8	790	0.5
Burkina Faso	3,200	0.3	211	0.02
Burundi	700	0.3	54	0.008
Cameroon	8,000	1.3	6	0.0004
Cape Verde	250	12.5		
Central African Republic	900	0.3	8	0.002
Cote d'Ivoire	15,000	1.3	254	0.02
Djibouti	1,300	2.2	6	0.009
Egypt	350,000	4.4	3,025	0.04
Ethiopia	10,000	0.1	78	0.001
Ghana	18,000	1.0	211	0.01
Iran	250,000	1.5	293	0.004
Israel	946,000	171.1	143,230	23.4
Jordan	85,300	17.5	518	0.07
Kenya	52,000	1.5	926	0.03
Kuwait	95,000	50.7	8,536	4.5
Lebanon	210,000	63.6	4,291	1.3
Lesotho	1,000	0.5	19	0.009
Liberia	400	0.2	1,350	0.5
Libya	6,800	1.2		
Madagascar	5,300	0.3	122	0.007
Malawi	7,000	0.7	1	0.00009
Maldives	3,000	10.0	160	0.5
Mali	1,900	0.2	1	0.00009
Mauritania	160,000	61.6	21	0.008
Mauritius	45,000	39.1	776	0.7
Morocco	160,000	5.7	2,719	0.09
Mozambique	17,000	0.9	254	0.01
Namibia	14,000	8.3	5,175	3.05
Niger	1,400	0.1	36	0.003
Nigeria	12,000	0.1	820	0.007
Oman	35,000	14.2	668	0.27
Qatar	40,000	67.9	20	0.03
Rwanda	450	0.1		
Saudi Arabia	230,000	11.0	638	0.03
Senegal	12,750	1.4	320	0.03
Seychelles	4,500	57.0	7	0.09
South Africa	1,899,000	47.6	165,600	4.1
Sudan	3,400	0.1	1	0.00003
Swaziland	4,200	4.3	300	0.3
Syrian Arab Rep	17,000	1.1	1	0.00006
Tanzania	9,500	0.3	250	0.007
Tunisia	78,000	8.2	30	0.003
Uganda	35,000	1.7	776	0.03
United Arab Emirates	350,000	146.0	35,000	14.6
Yemen Republic	10,000	0.6	30	0.002
Zambia	12,000	1.3	400	0.04
Zimbabwe	35,000	3.0	1,585	0.1

Source: Euromonitor (2001a, 2001b).

connecting with the global network in 1993, South Africa and Tunisia became the first African countries to join the Internet (Tripod.com, 2002). By 1998, all African countries except Congo had Internet connection (the State of the Internet in Africa, 2000; Tripod.com, 2002). South Africa dominates this region's Internet market, followed by Israel, Egypt, the United Arab Emirates, and Saudi Arabia (Table 3). In discussing ICTs in Africa, Ya'u (2002) tellingly points out, "Of the 157,325 Internet hosts in Africa, 144,445 are in South Africa, leaving less than 10,000 for the rest of Africa."

According to the State of the Internet Report 2000 of the U.S. Internet Council, Internet diffusion in Africa has been hampered by factors such as poverty, low computer penetration, illiteracy, lack of trained personnel, lack of interest, and a failure to understand the benefits of Internet access (CITI, 2000).

Internet Diffusion in Asia and the Pacific

Internet is growing very rapidly in the Asian–Pacific region. The number of Internet users in this region is expected to increase to 374 million by the end of 2005 (Rao, 1999). Japan, South Korea, Taiwan, and New Zealand dominate the region's Internet market so far (Table 4). China and India, however, are the fastest growing Internet markets in this region (Javalgi & Ramsey, 2001; Kshetri, 2002). Each is expected to have more Internet users than the U.S. by 2010 (Nua Internet Surveys, 1999b).

Internet Diffusion in Europe

Most European countries introduced the Internet much earlier than their counterparts in Asia, Africa, and Latin America. For instance, the University College of London (England) and NORSAR (Norway) were connected to ARPANET as early as in 1973. Similarly, Germany (DE), Italy (IT), the Netherlands (NL), and the United Kingdom (UK) were connected to the NSFNET in 1989 and Austria (AT), Belgium (BE), Greece (GR), Ireland (IE), Spain (ES), and Switzerland (CH) were connected in 1990 (Goldstein, 2000). Germany and the U.K. dominate the Internet market of this region (Table 5) .

Internet Diffusion in Latin America and the Caribbean

Compared to the North American and European economies, the Internet is a relatively new phenomenon

Table 4 Internet Users (IU) and Internet Hosts (IH) in Asia and the Pacific (1999)

Country	IU (1999)	IU per 1000 People (1999)	IH (1999)	IH per 1000 People (1999)
Armenia	4,500	1.3	1,788	0.5
Australia	5,000,000	267.3	950,400	50.8
Azerbaijan	1,378	0.2	522	0.1
Bangladesh	2,400	0.02		
Bhutan	25,500	0.2	54	0.02
Brunei	20,000	62.1	2,390	7.4
Cambodia	2,450	0.2	90	0.008
China	7,000,000	5.5	25,882	0.02
Hong Kong	1,500,000	220.5	99,600	14.6
India	2,500,000	2.5	24,518	0.02
Indonesia	1,360,000	6.5	22,399	0.1
Japan	20,458,000	162.1	2,337,880	18.5
Kazakhstan	40,000	2.4	1,776	0.1
Kyrgyz Rep	5,024	1.1	2,500	0.5
Lao PDR	26,300	5.0		
Macao	75,000	160.6	150	0.3
Malaysia	1,200,000	55.0	57,422	2.6
Mongolia	3,600	1.4	25	0.009
Nepal	30,000	1.3	168	0.007
New Zealand	800,000	209.0	145,000	37.9
Pakistan	100,000	0.7	6,192	0.04
Philippines	380,000	5.1	16,330	0.2
Singapore	1,200,000	340.7	73,810	21.0
South Korea	10,106,000	215.5	276,468	5.9
Sri Lanka	40,000	2.1	550	0.03
Taiwan	4,790,000	216.6	509,850	23.1
Tajikistan			148	0.02
Thailand	650,000	10.7	28,860	0.5
Turkmenistan			300	0.1
Uzbekistan	25,000	1.1	472	0.02
Vietnam	25,000	0.3	34	0.0004

Source: Euromonitor (2001a, 2001b).

Table 5 Internet Users (IU) and Internet Hosts (IH) in Europe (1999)

Country	IU (1999)	IU per 1000 People (1999)	IH (1999)	IH per 1000 People (1999)
Austria	1,650,000	201.8	259,500	31.7
Belarus	10,000	1	15,78	0.2
Belgium	1,840,000	181.2	365,155	36.0
Bulgaria	210,000	25.4	14,935	1.8
Croatia	400,000	89.3	10,457	2.3
Cyprus	60,000	77.1	8,334	10.7
Czech Republic	550,000	53.5	116,750	11.4
Denmark	2,100,000	397.6	417,200	79.0
Estonia	250,000	177	34,062	24.1
Finland	2,088,550	404.3	483,000	93.5
France	7,200,000	122.2	702,625	11.9
Georgia	8,750	1.7	1,107	0.2
Germany	12,300,000	149.7	1,721,150	20.9
Gibraltar	1,801		379	0.0
Greece	1,480,000	139.3	34,515	3.2
Hungary	450,000	44.7	139,000	13.8
Iceland	135,000	135	31,250	31.3
Ireland	500,000	155.1	76,526	23.7
Italy	9,100,000	158.7	579,900	10.1
Latvia	150,000	62.8	25,081	10.5
Lithuania	140,000	38	18,819	5.1
Luxembourg	750,00	176.1	10,831	25.4
Macedonia	45,000	22.4	1,853	0.9
Malta	40,000	103.6	2,849	7.4
Moldova	25,000	5.7	2,040	0.5
Netherlands	4,100,000	260.6	871,000	55.4
Norway	2,000,000	450.2	342,925	77.2
Poland	1,950,000	50.3	178,160	4.6
Portugal	800,000	81	70,000	7.1
Romania	385,000	17.2	38,791	1.7
Russia	5,400,000	36.7	230,176	1.6
Slovakia	1,000,000	185.8	32,045	6.0
Slovenia	490,000	246.4	26,335	13.2
Spain	3,625,000	91.5	460,500	11.6
Sweden	3,950,000	444.2	416,900	46.9
Switzerland	1,700,000	231.5	301,350	41.0
Turkey	1,200,000	18.3	70,865	1.1
Ukraine	262,500	5.2	29,663	0.6
United Kingdom	15,200,000	258.8	1,956,150	33.3

Source: Euromonitor (2001a, 2001b).

in Latin America and the Caribbean. The region, however, is experiencing phenomenal growth in the number of Internet users. A survey conducted by Nazca indicated that use of the Internet tripled in this region between 1995 and 1997 (Tudor, 1999). The networks in most of the Latin American economies were established during 1995–99 (Hahn, 1999). By connecting to NSFNet in 1989, Mexico became the first country in Latin America to connect to the Internet (Goldstein, 2000). As Table 6 indicates, big Latin American economies such as Argentina, Brazil, and Mexico dominate the Internet population in the region.

Internet Diffusion in North America

North America, the original locus of the Internet, accounts for numbers of Internet hosts and Internet users disproportionately higher than in the rest of the world. By 1999, the U.S. had over 98 million Internet users and 40 million hosts (Euromonitor, 2001b). Canada, similarly, had over 11 million Internet users and 1.4 million Internet hosts by 1999 (Euromonitor, 2001b). These numbers have been growing rapidly and reaching saturation levels.

Unconnected Pockets

Most of the countries that connected to the Internet very late, connected in very limited ways, or did not connect at all did so not because of economic reasons but because of *political* or *religious* reasons. Commenting on Internet-free countries in 1996, Maslen (1996) made the following observations:

Table 6 Internet Users (IU) and Internet Hosts (IH) in Latin America and the Caribbean (1999)

Country	IU (1999)	IU per 1000 People (1999)	IH (1999)	IH per 1000 People (1999)
Argentina	520,000	14.2	132,000	3.6
Barbados	12,000	44.6	88	0.3
Belize	20,000	85.8	260	1.1
Bolivia	42,000	5.3	689	0.09
Brazil	6,800,000	40.3	376,425	2.2
Chile	425,000	28.5	451,510	30.3
Colombia	450,000	11.6	25,110	0.6
Costa Rica	150,000	38.6	3,587	0.9
Cuba	50,000	4.5	120	0.01
Dominica	48,000	738.5		
Ecuador	16,500	1.3	1,935	0.2
El Salvador	50000	8.2	1,250	0.2
Guatemala	10,0000	9.1	1233	0.1
Guyana	3,500	5.0	72	0.1
Haiti	4,000	0.5		
Honduras	27,000	4.3	129	0.02
Jamaica	100,000	39.2	386	0.2
Mexico	2,100,000	21.6	197,750	2.0
Nicaragua	25,000	5.1	865	0.2
Panama	52,500	18.8	816	0.3
Paraguay	20,000	3.8	2,121	0.4
Peru	400,000	15.0	5,987	0.2
Surinam	10,854	25.2	1	0.002
Trinidad	40,000	31.1	3,013	2.3
Uruguay	320,000	96.9	22,706	6.9
Venezuela	500,000	20.9	13,842	0.6

Source: Euromonitor (2001a, 2001b).

In June 1996, just twenty countries remained with absolutely no e-mail or Internet connection: Afghanistan, Bhutan, Burma, Burundi, Congo, Gabon, Guinea Bissau, Iraq, Liberia, Libya, Mauritania, North Korea, Oman, Rio Muni, Rwanda, Somalia, Syria, Western Sahara, Yemen, and Zaire. . . . Of these countries, more than half are in Africa, in most cases only recently (fifty years or less) free from colonial rule. . . . Many of these countries are not wealthy, but there are surprising exceptions. Libya, Western Sahara, Oman and Gabon are relatively affluent, so their abstention from the Internet cannot be explained in purely economic terms.

Politics, economics and international diplomacy have always been involved in the spread of the Internet. Sixteen of these twenty "unwired" countries are Islamic, or have large Islamic populations. . . . At least six Internet-free countries have been involved in destructive internal or external conflicts in the last ten years. . . . Iraq, Libya, North Korea and Somalia (a fifth of the no-Internet countries) have all been on poor (or hostile) diplomatic terms with the United States. . . . [There is also] a definite (two-way) correlation between countries' levels of democracy and Internet connectivity. Sixteen of the twenty unconnected nations possessed very low levels of democracy.

THEORIES OF DIFFUSION AND ADOPTION OF INNOVATIONS

A deeper and richer understanding of the multifaceted and complex processes of diffusion and adoption of the Internet requires an integration of theories from diverse perspectives such as political science and international relations, sociology, marketing, communications, information systems, and geography.

Theoretical and empirical evidence suggests that politics and government policy play important roles in the diffusion of modern ICTs. Diffusion theories that relate diffusion pattern with political structure in a country argue that a technology is not equally compatible with all types of political structures. These theories, for instance, predict that authoritarian governments have relatively less favorable attitudes toward developing interpersonal means of communications such as the telephone and the Internet (e.g., Groth & Hunt, 1985; Kshetri, 2001; Kshetri & Dholakia, 2001). Groth and Hunt (1985), for instance, postulate that Marxist governments allocate a relatively smaller proportion of resources for the development of interpersonal communications. The unfavorable attitude towards the Internet has hampered the growth of Internet in countries with authoritarian regimes such as China, Cuba, and Syria.

Diffusion of innovations also entails "interaction between life-experience and history" (Erben, 1993) and hence theories from sociology may help explain some aspects of the diffusion dynamics. For instance, the cultural imperialism hypothesis (Tomlinson 1991)—western control of mass media combined with human desire to improve their lives leads consumers from developing countries to imitate the developed ones—could help explain the role of global media in the diffusion of modern ICTs in developing countries. Similarly, technology–society compatibility theories (Gatignon & Robertson, 1985; Rogers, 1983) could help explain why a particular technology is more compatible with certain societies than others. Past research has found that "country-level effects" or "societal effects" (Kshetri & Dholakia, 2002; Zaheer & Zaheer, 1997) have a strong influence on the technology adoption behavior of firms and individuals in a country.

Because diffusion of innovation requires an understanding of "people, hardware, software, communication networks and data resources" that collect, transform, and disseminate information (O'Brien, 1996), theories from information systems are well suited for explaining such aspects of the diffusion and adoption phenomena.

Diffusion and adoption of ICTs entail decisions at various levels (individual, organization, national, international) about allocation of scarce resources (Andrews, 2002). Economic theories help explain how individuals, organizations, and governments make decisions regarding the allocation of resources in ICT investment. Basic demand and supply theories are helpful in explaining the varied distribution patterns of the Internet across countries. For instance, income can be expected to be positively associated with the demand for modern ICT products (Gatignon & Robertson, 1985; Rogers, 1983).

Geography and geopolitics are important aspects of international business in general (As-Saber, Liesch, & Dowling, 2000) and global diffusion of innovations in particular. Geopolitical theories help in answering questions such as, "Why are things located where they are? How do different places relate to each other? How have geographic patterns and relationships changed over time?" (Baerwald, 1996, p. 23). Geopolitical variables include population (Cohen, 1963) as well as characteristics of populations such as skills, educational qualifications, productivity, and the cost of labor (As-Saber et al., 2000; Baerwald, 1996). Literacy rates and English language skills are other geopolitical factors likely to influence the diffusion of innovations. Such skills are associated with the availability of information and the proficiency required to use a new technology or the "interrelatedness" among users and producers of the technology (Cassiolato & Baptista, 1996). The diffusion of a technology is negatively related to the disparity between the skills needed to use the technology and the consumer's existing knowledge (Gatignon & Robertson, 1985). High illiteracy rates have hampered the adoption of the Internet in less developed countries. Geopolitical variables such as geographical distance between adopting units (Gatignon & Robertson, 1985) influence the level of terrestrial barriers and hence the diffusion of the Internet. Origination of the Internet in the U.S. is one of the major factors responsible for the disproportionately higher number of Internet users in North America. Furthermore, high-bandwidth backbones were created first in North America, followed by Europe, and the data traffic to other regions was constrained by low-bandwidth connections.

FACTORS IMPACTING THE DIFFUSION OF THE INTERNET

Economic Factors

Economic factors such as income level, availability and price structures of ICT products and services, and bandwidth and supporting infrastructures influence the diffusion of the Internet (Kshetri, 2001). The cost of a PC as a proportion of per capita GDP, for instance, is 5% in high-income countries compared to about 300% in low-income countries (International Telecommunications Union, 2001a). In January 2001, the price of the cheapest Pentium III computer was US $700 (United Nations Development Program [UNDP], 2001b), an amount much higher than the average per capita GDP of most developing countries. Similarly, monthly Internet access charge as a proportion of per capita GDP in 2001 varied from 1.2% in the United States to 118% in Sierra Leone (international Telecommunications Union, 2001a).

Economic factors influence the means used to access the Internet. Many people in developing countries who cannot afford PCs can access the Internet through public kiosks and cafes at much lower rates (e.g., Kirby, 2002). Paging networks are popular in China because of the higher costs of computer and the fixed Internet access (Ebusinessforum.com 2000). In Japan, the popularity of NTT DoCoMo's i-Mode services and the government's effort to lower the cost of data flow over the mobile network accelerated the growth of the mobile Internet (Stout, 2001).

Finally, bandwidth availability is a determinant of Internet adoption and diffusion. In general, it is very low in developing countries. For instance, 50% of the worldwide bandwidth capacity is in North America compared to 3% in the Middle East and Africa (Frontline.net, 2001). Lower bandwidth results in longer times to transfer data and hence low relative advantage of Internet use. Moreover, the lack of intraregional infrastructures in developing nations of Asia, Africa, and Latin America means that even Internet communications with neighboring countries have to be routed through the U.S. or other industrialized countries in Europe, further increasing the costs. When high bandwidth is available and reasonably priced, as in South Korea, it becomes a driver of rapid Internet diffusion.

Sociocultural Factors

The degree of compatibility of the Internet and its various uses with the values and norms of a social system (Rogers, 1983) influences its diffusion patterns in that social system. An examination of the "values" and "culture" inherent in the Internet thus helps predict the degree of acceptance or rejection of the Internet in a society. An important component of this value system is related to skills required to use the Internet. Literacy and computer skills are almost prerequisites for Internet use. A large proportion of the population in developing countries is illiterate and a still higher proportion lacks computer skills.

Some attempts to overcome the literacy and language barriers are under way. One such attempt is the creation of the Simputer (Sterling, 2001), "a small handheld device designed for the rough conditions of rural India. It operates—without a keyboard—through touch, sound, and simple visual icons. It translates English-language Web sites into local Indian languages, reading the content aloud to illiterate users."

Moreover, the Internet tends to favor the English-speaking population because most of the software and interfaces used in the Internet are in English. Also, a large proportion of the WWW content is in the English language. For instance, a survey conducted in 1998 found that about 85% of the texts on the WWW were in English (Nunberg, 2000), which decreased to about 80% in 1999 (Nua Internet Surveys, 1999a).

Internet's asynchronous nature and impersonal style of communication make it incompatible with the cultures of some societies. For instance, in Japan personal correspondence is normally handwritten to show respect and courtesy (James, 1998).

Another component of the Internet's value system has to do with the place of origin of the core technology as well as the bulk of the content. The Internet originated in the U.S. and most of the content available on the WWW originates in the Western countries. Many people in the East tend to doubt the integrity of information originating from the Western world and view the use of English as a vehicle for executing an electronic "Pax Americana" (Shabazz, 1999).

Geopolitical Factors

National institutions can take measures to influence an innovation as well as its diffusion. As discussed in an earlier section, then U.S. President Eisenhower took initiatives to create the Advanced Research Projects Agency (ARPA) within the U.S. Department of Defense in 1958, thereby laying the foundation for the Internet. Similarly, measures taken by the National Science Foundation (NSF) in 1985 to make TCP/IP the mandatory protocol accelerated the diffusion of the Internet.

Political support strongly influences the diffusion of the Internet. Investments in modern ICTs alone will not bridge the digital divide in the absence of such political support (Koss, 2001). Internet diffusion in several countries, especially the developing ones, has been hampered by political factors such as authoritarian governments' concern about the flow of information on the Internet, tariff/nontariff barriers to ICT products, and unfavorable regulatory environments that negatively influence the telecom markets (Kshetri, 2001). The Internet has been described as "the greatest democratizer the world has ever seen" (Pitroda, 1993, p. 66). It is, thus, incompatible with most authoritarian regimes. Threatened by free flow of information on the Internet and the possible negative impact on their "right to rule," authoritarian regimes such as those of Malaysia, China, Singapore, Syria, and Cuba have opted for several mechanisms to control the Internet. For instance, the Chinese government closed 150,000 unlicensed Internet cafes following a deadly fire in a Beijing Internet cafe and the remaining cafes are required to install software that prevents access to up to 500,000 banned pornographic sites or sites with "subversive content" (Behind China's Internet red firewall, 2002). Similarly, a recent study conducted by Harvard Law School found that a central array of proxy servers in Saudi Arabia filters and blocks "sexually explicit" content (Hermida, 2002).

Authoritarian regimes are also slow to enact laws to recognize digital and electronic signatures (DES) (Kshetri and Dholakia, 2001). Many democratic governments lack DES laws as well. By the end of 2000, for instance, only about 45 nations in the world had laws recognizing DES (Stephens, 2001). Lack of DES laws has hampered the use of the Internet for commercial and government services in several countries.

Tariff and nontariff barriers and regulations in telecom markets are hindering e-commerce development in some developing markets. Many developing countries tend to treat ICT products as luxury items and impose import duty, surtax, value added tax, sales tax, etc., making these products expensive and unobtainable (UNCTAD, 2000).

Measures Taken by International Agencies

Several international institutions have launched ICT-led initiatives to accelerate the diffusion of the Internet (Kshetri, 2001; Shadrach, 2002). First, international agencies have helped *introduce* the Internet for the first time in developing countries that lacked the infrastructure for the Internet (Brown, Malecki, & Spector, 1976). In this way, the international agencies are breaking the "hierarchical pattern" (Gatignon & Robertson, 1985, p. 858) of Internet diffusion—a pattern of slow trickle-down of innovations from the rich to the poor. The United Nations Development Program (UNDP) introduced the Internet in more than 15 countries by connecting them to the global network (UNDP, 2001a). Similarly, the World Bank initiated the Information for Development Program (InfoDev) in 1995 to promote projects emphasizing the use of ICTs for economic and social development. This program has a special emphasis on the needs of the poorest in developing countries.

Second, some international agencies are helping to *reduce the gap between skills required for Internet use and existing skill*s of potential users in developing countries. By early 2001, for instance, UNDP had trained over 25,000 organizations and helped create over 40,000 Web sites for governments and civil society stakeholders (UNDP, 2001a).

Third, international institutions are influencing national governments to *increase the level of competition* in the telecom sector, which has resulted in the availability of higher quality ICT products and services at lower prices. In 1997, 60 developing countries made commitments to the World Trade Organization (WTO) to introduce competition in the telecom sector. Similarly, 13 developing countries also signed the Information Technology Agreements (ITA) under the WTO to eliminate customs duties on seven broad categories of ICT products.

Table 7 Internet Penetration and Related Indicators for Selected Economies in the World

	Internet Penetration			Factors Influencing Internet Penetration			
Country	No. of Internet users per 1000 (1999)	No. of Internet hosts per 1000 (1999)	Rank order of Internet penetration (within region)	GNP per capita ($, 2000)	Population ('000), (2000)	Civil Liberty Index (2000)	Literacy rate (%, 2000)
Africa and Middle East							
Israel	155.1	23.5	1	16,310	5,842	2	95.7
United Arab Emirates	146.0	14.6	2	17,276	2,369	5	74.6
South Africa	47.6	4.2	3	3,020	43,421	2	84.6
Sierra Leone	1.5	0.0	4	130	5,233	5	31
Asia-Pacific							
Singapore	340.7	21.0	1	24,740	4,152	5	91.8
Australia	267.3	50.8	2	20,530	19,165	1	99
Japan	162.1	18.5	3	34,210	126,550	2	99
China	5.5	0.0	4	840	1,261,832	6	82.8
Pakistan	0.7	0.0	5	470	141,554	5	44
Europe							
Finland	404.4	93.5	1	24,900	5,167	1	99
Slovenia	246.4	13.2	2	10,070	1,928	2	99.6
Portugal	81.0	7.1	3	11,060	10,048	1	91.4
Romania	17.2	1.7	4	1,670	22,411	2	97.9
Latin America and Caribbean							
Uruguay	96.9	6.9	1	6,090	3,334	2	97.6
Argentina	14.2	3.6	2	7,440	36,955	3	96.7
Bolivia	5.3	0.1	3	1,000	8,153	3	84.4
Cuba	4.5	0.0	4	1,700 (ppp)	11,142	7	96.4
North America							
USA	351.7	145.6	1	34,260	275,563	1	99
Mexico	21.6	2.0	2	5,080	100,350	4	90.8

Notes: 1. ppp stands for purchasing power parity. 2. Civil liberty index: 1 represents most free and 7 represents least free.
Sources: http://www.lyd.org/english/economic_freedom/percapita_income_troughout.pdf, http://www.odci.gov/cia/publications/factbook/geos/cu.html#Econ, European Marketing Data and Statistics, Freedom House, International Marketing Data and Statistics.

Fourth, international institutions are facilitating Internet *adoption by small and medium-sized enterprises* (SMEs), which otherwise may be reluctant to adopt the Internet. The United Nations Conference on Trade and Development (UNCTAD) launched the Global Trade Point Network (GTPN) in 1992. Its objective is to facilitate Internet adoption by SMEs, especially for accessing global markets. As of 2000, its electronic trading opportunity (ETO) system connected more than 20,000 trade organizations worldwide.

Fifth, international institutions are *influencing national laws, regulations and policies*, and are making them more conducive to the use of the Internet for various purposes. The UN Commission on International Trade Law (UNCITRAL) undertook a major initiative leading to the adoption of the Model Law on E-Commerce. Many countries around the world have enacted new Internet laws by taking the UNCITRAL model law as the guideline. The World Intellectual Property Organization (WIPO) member states also approved the establishment of WIPOnet, which provides basic, secure Internet connectivity and services to intellectual property offices. Similarly, the International Chamber of Commerce (ICC) has developed a model contract for privacy and transborder data flows. The Organization for Economic Cooperation and Development (OECD), in a similar vein, has developed action plans to address issues related to authentication, certification, consumer protection, and privacy in the use of the Internet (Kshetri, 2001).

The differences in social systems across the world have been selectively summarized in Table 7 to show the influence of several factors shaping the diffusion dynamics of the Internet. It is quickly evident that countries with the lowest Internet penetration levels also have the highest restrictions on civil liberties (e.g., Sierra Leone, Pakistan, and Cuba). It is possible to overcome some of the barriers to Internet penetration created by restricted civil liberties through high literacy levels, as in the case of China and Singapore. In addition to an open society and high literacy, economic resources are also necessary. Countries with the highest Internet penetration levels also have the highest levels of per capita income (e.g., USA, Finland). Finally, sociocultural factors such as English language skills explain some of the differences in Internet penetration between Australia and Japan.

CONCLUSION

In this chapter, we provided an overview of the global diffusion of the Internet and examined the factors that have shaped and are continuing to shape the diffusion dynamics of the Internet. The diffusion pattern of an innovation such as the Internet is a function of economic, political, cultural, and geographical factors. Because social systems across the world differ significantly in terms of these factors, diffusion patterns of the Internet also vary widely across nations. Low levels of income, authoritarian governments' distaste toward the Internet, sociocultural environments that are incompatible with the Internet, and terrain barriers have hampered the rapid diffusion of the Internet in developing countries. As a result, developing countries account for a disproportionately low number of Internet users and Internet hosts worldwide. Enlightened public policies, forward-looking corporate strategies, and concerted efforts by international agencies are of value in extending the reach of the Internet to all parts and all citizens of the globe.

ACKNOWLEDGMENT

The authors are grateful to multiple reviewers of *The Internet Encyclopedia* for detailed and valuable comments on earlier versions of this chapter.

GLOSSARY

Adoption of an innovation A micro process that focuses on the stages through which an individual passes when deciding to accept or reject the innovation.

ARPANET (Advanced Research Projects Agency Network) Developed by ARPA with support from the United States Department of Defense, this can be considered as the origin of all networks, including the Internet. Arpanet lines were considered superfast for their time, *56 kilobits per second.* ARPA and DARPA are used interchangeably, as the agency has used both names over the course of its existence.

Diffusion of an innovation A macro process concerned with the spread of an innovation from its source to the public.

Interface Message Processors (IMPs) The packet switches developed by the Defense Advance Research Project Agency (DARPA) in 1968.

Information and communications technologies (ICTs) Technologies that facilitate the capture, processing, storage, transfer, and presentation of information.

Internet host A computer system connected to the Internet—either a single terminal directly connected or a computer system that allows multiple users to access network services through it.

Network control protocol (NCP) The initial ARPANET host-to-host protocol developed by the Network Working Group (NWG) in 1970.

NSFNET (National Science Foundation Network) The first backbone for the U.S. portion of the Internet. It was originally conceived as a way for researchers to submit jobs to supercomputers located at various universities around the U.S.

Transmission control protocol/Internet protocol (TCP/IP) A protocol designed to meet the needs of an open-architecture network environment. TCP verifies the correct delivery of data from client to server. TCP adds support to detect errors or lost data and to trigger retransmission until the data is correctly and completely received. IP is responsible for moving packets of data from node to node. IP forwards each packet based on a four-octet destination address known as the IP number. IP operates on gateway machines that move data from department to organization to region and then around the world. In the future, IP is projected to move to six-octet addressing.

CROSS REFERENCES

See *Feasibility of Global E-business Projects; Gender and Internet Usage; Global Issues; Internet Literacy; Internet Navigation (Basics, Services, and Portals).*

REFERENCES

Angus Reid Group (2000). New benchmark study pegs global Internet population at more than 300 million—Wireless devices, not PC's, critical to next generation growth. Retrieved March 22, 2000, from http://www.immedia.it/published/20000322/200003225936.shtml

Andrews, D. (2002). The economy of attention. *Business Communication Quarterly, 65*(1), 7–8.

As-Saber, S. N., Liesch, P. W., & Dowling, P. J. (2000, November 17–20). *Geopolitics and its impacts on international business decisions: A framework for a geopolitical paradigm of international business.* Paper presented at the Annual Meeting of the Academy of International Business, Phoenix, AZ.

Baerwald, T. J. (1996). Geographical perspectives on international business. In M. R. Czinkota, I. A. Ronkainen, & M. H. Moffett (Eds.), *International Business* (4th ed.). Orlando, FL: Harcourt.

Behind China's Internet red firewall (2002, September 3). *BBC News*. Retrieved September 3, 2002, from http://news.bbc.co.uk/2/hi/technology/2234154.stm

Bell, H., & Tang, N. K. H. (1998). The effectiveness of commercial Internet web sites: A user's perspective. *Internet Research: Electronic Networking Applications and Policy, 8*(3), 219–28.

Brown, L., Malecki, E., & Spector, A. (1976). Adopter categories in a spatial context: Alternative explanations for an empirical regularity. *Rural Sociology, 41*, 99–118.

Cassiolato, Jose E., & Baptista, M. A. C. (1996). The effects of the Brazilian liberalization of the IT industry on technological capabilities of local firms. *Information Technology for Development, 7*(2), 53–73.

Cohen, S. B. (1963). *Geography and politics in a world divided.* New York: Random House.

Cyber citizenship gains in developing world. (2000). *Futurist 34*(5), 19–21.

Erben, M. (1993). The problem of other lives: Social perspectives on written biography. *Sociology: The Journal of the British Sociological Association, 27*(1), 15–25.

Euromonitor (2001a). *European Marketing Data and Statistics*. London: Euromonitor.

Euromonitor (2001b). *International Marketing Data and Statistics*. London: Euromonitor.

Frontline.net (2001). Broadband in the developing world. Retrieved April 21, 2001, from http://www.pressroom.com/~screenager/broadband/Intro.html

Gatignon, H., & Robertson, T. S. (1985). A propositional inventory for new diffusion research. *Journal of Consumer Research, 11*, 849–867.

Goldstein, D. (2000). History of the Internet. Retrieved April 21, 2001, from http://www.nic.at/english/geschichte.html

Groth, A., & Hunt, W. R. (1985). Marxist–Leninist communications systems in comparative perspective. *Coexistence, 22*, 123–36.

Hahn, S. (1999). Case studies on developments of the Internet in Latin America: Unexpected results. *Bulletin of the American Society for Information Science 25*(5), 15–17.

Hermida, A. (2002, July 31). Saudis block 2,000 Websites. *BBC News*. Retrieved September 3, 2002, from http://news.bbc.co.uk/2/hi/technology/2153312.stm

International Telecommunications Union (2000). Digital mobile growth. Geneva, Switzerland: International Telecommunications Union.

International Telecommunications Union (2001a, May 17), *The Internet: Challenges, opportunities and prospects*, 17 May—World Telecommunication Day. Retrieved July 1 2001, from http://www.itu.int/newsroom/wtd/2001/ExecutiveSummary.html

International Telecommunications Union (2001b) ITU telecommunications indicators update. Geneva, Switzerland: International Telecommunications Union.

James, D. (1998). No tsunami yet. *Upside, 1*(3), 72–74.

Javalgi, R., & Ramsey, R. (2001). Strategic issues of e-commerce as an alternative global distribution system. *International Marketing Review, 18*(4), 376–91.

Kirby, A. (2002, October). Doing business in Asia. *Credit Management*, 24–25.

Koss, F. A. (2001). Children falling into the digital divide. *Journal of International Affairs, 55*(1), 75–90.

Kshetri, N. (2001). Determinants of the locus of global e-commerce. *Electronic Markets, 11*(4), 250–257.

Kshetri, N. (2002). What determines Internet diffusion loci in developing countries: Evidence from China and India. *Pacific Telecommunications Review, 23*(3), 25–34.

Kshetri, N. (2003). How international institutions influence ICT diffusion loci in developing countries: The case of Asia. In C. Shultz, M. Speece, and D. Rahtz (Eds.), *Proceedings of the 8th International Conference on Marketing and Development: New Visions of Marketing and Development: Globalization, Transformation and Quality of Life* (pp. 314–326). Muncie, IN: International Society for Marketing and Development.

Kshetri, N., & Dholakia, N. (2001). Impact of cultural and political factors on the adoption of digital signatures in Asia. In *Proceedings of the Americas' Conference on Information Systems* (pp. 1666–1673). Atlanta, GA: Association of Information Systems.

Kshetri, N., & Dholakia, N. (2002). Determinants of the global diffusion of B2B e-commerce. *Electronic Markets, 12*(2), 120–129.

Lammers, D. (2001, May 14). Mixing up broadband. *Electronic Engineering Times*.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2002). A brief history of the Internet. *Internet Society*. Retrieved September 3, 2002, from http://www.isoc.org/Internet/history/brief.shtml

Maslen, P. (1996). Control, Change and the Internet (Chapter 4) Retrieved from http:/home.vicnet.net.au/~qbird/cci-ch4.htm

McLaren, B. J. (1999). Understanding & using the Internet. Cincinnati, OH: South-Western Educational.

Media Metrix (2000). Top rankings. Retrieved April 12, 2000, from http://www.mediametrix.com/TopRankins/TopRankings.html

Nua Internet Surveys (1999a, May 29). Internationalization of the Web. Retrieved from http://www.nua.ie/surveys/

Nua Internet Surveys (1999b, November 9). Chinese users to outnumber US users by 2010. Retrieved from http://www.nua.ie/surveys/

Nua Internet Surveys (2002). How many online. Retrieved December 3, 2002, from http://www.nua.ie/surveys/how_many_online/index.html

Nunberg, G. (2000, March 27–April 10). Will the Internet always speak English? *The American Prospect*, 40–43.

O'Brien, J. A. (1996). *Introduction to information systems* (8th ed.). Chicago: Irwin/McGraw Hill.

Pastore, M. (2000). U.S. Internet dominance slipping. *InternetNews—International News*. Retrieved September 3, 2002, from http://cyberatlas.internet.com/big_picture/geographics/article/0,1323,5911_330201,00.html

Pitroda, S. (1993, November/December). Development, democracy, and the village telephone. *Harvard Business Review, 71*, 66.

Rao, M. (1999). Asia.com: Will steam ahead despite bumps on the information superhighway. *Asia Links*. Retrieved October 30, 2002, from http://www.asia-links.com/scripts/articles/article.asp?articleid=92&folder id=8

RAND's Hisotry (n.d.). RAND and the origins of the Internet. Retrieved April 29, 2003, from http://www.rand.org/history/baran.html

Reports about the death of broadband are premature. (2002). *High-Speed Internet Access, 18*(8), 14–15.

Rogers, E. M. (1983). *The Diffusion of Innovations* (3rd ed.). New York: Free Press.

Schwimmer, B. E. (2002). The anthropology of cyberspace, Unit 3: History and expansion of the Internet. Retrieved September 30, 2002, from http://www.umanitoba.ca/faculties/arts/anthropology/courses/478/unit3.htm

Shabazz, D. (1999). International politics and the creation of a virtual world. *International Journal on World Peace, XVI*(3), 27–44.

Shadrach, B. (2002). India's development information network: Lessons learned. *Bulletin of the American Society for Information Science, 28*(2), 23–27

Shaffer, R. (2000, July 10). M-commerce: Online selling's wireless future. *Fortune*.

Stephens, D. O. (2001). Digital signatures and global e-commerce. I. U.S. initiatives. *Information Management Journal, 3*(1), 68.

Sterling, B. (2001, December). Simputer. *The New York Times Magazine*. Retrieved July 20, 2002, from http://

www.nytimes.com/2001/12/09/magazine/09SIMPUTER.html

Stout, K. L. (2001, April 24). Japan Internet users up 74 percent. *CNN.com*. Retrieved September 28, 2002, from http://edition.cnn.com/2001/BUSINESS/asia/04/24/tokyo.netuserup/

Study says broadband growth remains robust. (2002, July 23). *Communications Today*.

Takacs, S. J., & Freiden, J. B. (1998). Changes on the electronic frontier: Growth and opportunity of the World-Wide Web. *Journal of Marketing Theory and Practice. 6*(3), 24–37.

The state of the Internet in Africa (2000). CITI. Retrieved September 3, 2002, from http://www.citi.org.za/cgi-bin/read.pl?DBNAME = NEWS&ID = 31

Thierer, A. D. (2000). Is the "digital divide" a virtual reality? *Consumers' Research Magazine, 83*(7), 16–20.

Tripod.com (2002). Internet in Africa. Retrieved November 9, 2002, from http://tjj1.tripod.com/africa.htm

Tudor, J. D. (1999). Latin American information environment. *Database, 22*(2), 66–69.

United Nations Conference on Trade and Development (2000). *Building confidence: Electronic commerce and development*. Geneva: United Nations Conference on Trade and Development.

United Nations Conference on Trade and Development (2002). E-commerce and development report 2002. Retrieved November 21, 2002, from http://r0.unctad.org/ecommerce/docs/edr02_en/ecdr02ch3.pdf

United Nations Development Program (2001a). *Human development report 2000*. New York: United Nations Development Program. Retrieved July 22, 2001, from Retrieved, from http://www.undp.org/hdr2001/completenew.pdf

United Nations Development Program (2001b). *Human development repor*t. New York: United Nations Development Program. Retrieved June 26, 2001, from http://www.undp.org/hdr2001/completenew.pdf

Webopedia.com (2002). Brier timeline of the Internet. Retrieved November 10, 2002, from http://www.webopedia.com/quick_ref/timeline.asp

World Intellectual Property Organization (2000). *Primer on electronic commerce and intellectual property issues*. Geneva: World Intellectual Property Organization. Retrieved June 26, 2001, from http://ecommerce.wipo.int/primer/section1.html

Ya'u, Y. Z. (2002, April 26). Confronting the digital divide: An interrogation of the African initiatives to bridge the gap. TWN-Africa. Retrieved July 22, 2002, from http://twnafrica.org/ISSUES/nepad/nepad_eventdetail.asp?twnID = 216

Zaheer, S., & Zaheer, A. (1997). Country effects on information seeking in global electronic networks. *Journal of International Business Studies; 28*(1), 77–100.

FURTHER READING

McNulty, L. (1999). Image display size. Retrieved September 17, 2000, from http://www.envisiondev.com/idug/oct96_w/WG_FAST.htm

United Nations Development Program (2001). *Driving information and communications technology for development*: *A UNDP agenda for action 2000–2001*. New York: United Nations Development Program. Retrieved June 26, 2001, from http://www.sdnp.undp.org/it4dev/ffICTe.pdf

Global Issues

Babita Gupta, *California State University, Monterey Bay*

INTRODUCTION TO GLOBAL ELECTRONIC COMMERCE

Globalization as a term evokes strong emotions—either as an inevitable force that will transform lives of people worldwide for the better or as a force that will increase existing inequities and dilute local cultures. Globalization refers to the process of diffusion of people and their knowledge across nations, cultures, and economies. From an economic viewpoint, globalization is the process of many distinct national economies integrating into one global economy through free trade and free capital mobility, but also by easy or uncontrolled migration (Daly, 1998).

Global trade has existed for hundreds, perhaps thousands, of years as a result of breakthroughs in navigation technologies that allowed people to travel to distant lands and open new markets in exotic goods. In the more recent past, innovations in transportation and communication have allowed global business to expand dramatically at decreasing marginal costs. The Internet has emerged as another enabling technology in the process of globalization and global business expansion, facilitated by collaborations between public and private enterprises across borders to expand and develop new know-how (Iammarino & Mitchie, 1998). However, much of what is usually referred to as "globalization" is more accurately the process of "internationalization," in which a company expands its domestic base globally utilizing technological capabilities such as the Internet and international collaborations strategies such as mergers and acquisitions (Howells & Mitchie, 1998). From organizational perspective, firms that pursue a transnational strategy by locating their organizational activity in a region or country where they can efficiently utilize local assets and are likely to be more globally competitive (Boudreau, Loch, Robey, & Straud, 1998).

Global Electronic Commerce Growth

In the early 1990s, many analysts predicted that the Internet would transform the global economy by making possible a single world economy with insignificant national boundaries (de la Torre & Moxon, 2001). These predictions, however, proved to be unrealistic, as the Internet, like any other technology, operates within the context of the cultural, political, and economic impulses of the people of any country and needs to be considered in a broader context than just the organizational context.

According to one estimate (Forrester Projects $6.8 Trillion for 2004, 2002), by 2004, the United States alone would account for more than 47% of worldwide electronic commerce sales. Despite high variance in electronic commerce volume figures among various sources, worldwide expenditures distribution in the business-to-consumer (B-to-C) and business-to-business (B-to-B) markets among world regions show a clear domination of North American markets (Iyer, Gupta, & Taube, 2002b). Table 1 summarizes the growth of global electronic commerce as reported by various sources. It suggests that the gap between North America and other world regions is smaller for B-to-B than B-to-C electronic commerce (Kshetri & Dholakia, 2002).

The Internet Diffusion

Research suggests that the per capita income or economy of a nation is not the only explanation for the uneven growth of electronic commerce. Diffusion of new technologies such as the Internet differs among countries and is a function of a country's unique cultural,

Table 1 B-to-C and B-to-B Electronic Commerce Revenue Estimates Across World Regions

World Regions	B-to-C E-commerce Revenue Estimates	B-to-B E-commerce Revenue Estimates
North America (5.6% of world population)	*Y2000:* $38 billion 47.5 billion* *Y2004 Projections:* $184 billion, 197.9 billion* *Y2006 Projections:* $211 billion**	*Y2000:* $1.2 trillion(59% of global revenues) $159 billion* *Y2004 Projections:* $4.8 trillion $1.6 trillion* *Y2006 Projections:* $7.13 trillion**
Europe (13.5% of world population)	*Y1999:* $3.5 billion *Y2000:* $8.1 billion* *Y2004 Projections:* $182.5 billion* *Y2006 Projections:* $144 billion**	*Y1998:* $3.75 billion *Y2000:* $26.2 billion* *Y2002 Projections:* $174 billions $132.7 billion* *Y2004 Projections:* $797.3 billion* *Y2006 Projections:* $2.4 trillion**
Asia-Pacific Rim (60.2% of world population)	*Y1999:* $2.8 billion *Y2000:* $3.2 billion* *Y2004 Projections:* $38 billion* *Y2006 Projections:* $185 billion**	*Y1999:* $9.2 billion (excluding Japan) *Y2000:* $36.2 billion* *Y2003 Projections:* $430 billion $199.3 billion* *Y2004 Projections:* $1 trillion $300 billion* *Y2006 Projections:* $2.46 trillion**
Latin America (8.3% of world population)	*Y1999:* $77 million *Y2000:* $70 million* *Y2003 Projections:* $3.8 billion $5.5 billion* *Y2006 Projections:* $16 billion**	*Y2000:* $2.9 billion* *Y2004 Projections:* $76 billion $58.4 billion* *Y2006 Projections:* $216 billion**
Africa/Middle East* (12.4% of world population)	*Y2000:* $20 million* *Y2004 Projections:* $1.6 billion* *Y2006 Projections:* $5 billion**	*Y2000:* $1.7 billion* *Y2004 Projections:* $17.7 billion* *Y2006 Projections:* $69 billion**

Sources: Kshetri and Dholakia (2002); * Iyer, Taube, & Raquet (2002a); ** United Nations Conference on Trade and Development Report (2002).

economic, legal, religious, and political forces. The role of governments in providing incentives for IT innovation and diffusion in the form of regulatory and policy frameworks is well documented (King et al., 1994). Wolcott, Press, McHenry, Goodman, and Foster (2001) developed a framework using six dimensions to describe Internet diffusion across nations. *Connectivity infrastructure* describes underlying telecommunication networks necessary for the presence of the Internet; *pervasiveness* refers to the number of Internet users in a country; *sectoral* *absorption* is a measure of Internet use in organizations within a country; *organizational infrastructure* refers to the number and capabilities of organizations that can provide Internet services to users; *geographic dispersion* refers to the extent of geographic distribution of such organization throughout the nation; and *sophistication of use* refers to both how many people use the Internet and how they use it.

Research studies of the role of institutional influences in technology innovation (King et al., 1994; Montealegre, 1999), growth factors in B-to-B electronic commerce (Kshetri & Dholakia, 2002), and global electronic commerce (Farhoomand, Tuunainen, & Yee, 2000; Kshetri, 2001) indicate that adoption and diffusion of electronic commerce within a country and between countries is facilitated (or hindered) by the following:

Geopolitical issues such as effective partnership between government and private sectors, attitudes toward foreign investments, investments in infrastructure, subsidies, taxation, and other regulatory policies.

Cultural and social factors that determine how individuals in a country feel toward technological change and their level of trust in financial and business institutions.

Technology issues such as telecommunication infrastructure, adoption of uniform technical standards, and availability of a skills base to adopt and use new technologies.

Financial and economic issues such as viability of financial institutions, currency exchanges, payment options, availability of skilled human capital, distribution infrastructure, and effective trading policies.

This chapter discusses these issues in the globalization context of the Internet and electronic commerce and the resulting global digital divide. Its sections are organized as follows: a discussion of global electronic commerce and its advantages and disadvantages; role of governments in enacting regulations and laws to address cross-border issues arising due to the Internet; privacy and security issues that impact the growth of electronic commerce as a result of the lack of consensus among sovereign entities; the state of telecommunication and physical infrastructure, which are still nonexistent in many parts of the world; the state of financial services; and the cultural issues that impact the adoption and growth of Internet technologies and electronic commerce. We conclude with strategies that may be useful in spurring the growth of Internet technologies worldwide and some directions for future research.

GLOBAL ELECTRONIC COMMERCE

The modern economy is largely based on the production and distribution of products and services according to well-established economic laws in quantifiable measures such as gross national product and per capita income. These economic concepts determine the relative economic status of nations. As the information technology is transforming the "industrial" economy (i.e., economic growth through maximizing output) to the "informational" economy (i.e., economic growth through efficiencies in information processing and the creation and accumulation of knowledge), it is leading to the emergence of informational capitalism, where companies employ information technologies for profitability and nations are motivated to facilitate infrastructure for creating an information economy and thus to be competitive in the global economy (Venkatesh, 1999).

Current Models for Measuring Global Electronic Commerce Readiness

Before we can begin to comprehend the reasons for the uneven growth of electronic commerce worldwide, it is important to understand the models used to compare electronic commerce growth across countries.

The e-business country opportunity index, one such framework developed by the Gartner Group, is based on the Internet penetration ratio (the percentage of a country's population with access to Internet) against e-commerce transaction values per person achieved by that country (Iyer, Taube, & Raquet, 2002a). Figure 1 includes the United States and some European countries and shows that the United States has a substantial lead over other countries benchmarked in the index (Iyer et al., 2002a).

The gross domestic product (GDP)/Internet penetration matrix, another matrix developed by the Gartner Group, maps total Internet usage in percentage of population against GDP per capita and includes several non-European countries as well (see Table 2, which uses this index to group countries into four basic clusters).

In the past few years, a number of assessment tools using different frameworks have been developed to measure the readiness of a community, country, or region to benefit from information technology and electronic commerce. Although there are a number of such tools and they differ widely, most can be categorized as either *e-economy* or *e-society* assessment tools. E-economy assessment tools look at the ability of information and communications technologies (ICTs) to impact the economy, whereas e-society assessment tools look at the potential impact of

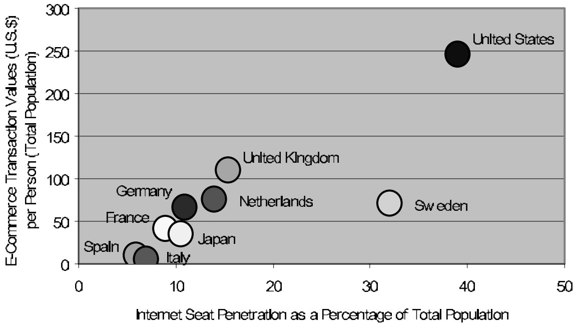


Figure 1: GartnerGroup's e-business country opportunity index. Source: Iyer, Taube, & Raquet (2002a). Reprinted with permission.

Table 2 GDP per Capita/Internet Penetration Ratio for Selected Countries

	High Internet Penetration	Low Internet Penetration
High GDP	U.S., Sweden, Canada, Australia	France, Japan, Germany
Low GDP	S. Korea, Singapore, Netherlands, UK	Brazil, China, Spain

Source: Iyer, Taube, & Raquet (2002a).

ICTs on the wider society (Comparison of E-Readiness Assessment Models, 2001).

Use of ICTs can contribute significantly to the growth of the global Internet by reducing transaction costs (through declining marginal costs) by facilitating transparency in the market processes that offer greater efficiencies. These lead to lowered costs of goods to consumers. Also, higher use of ICTs has a positive correlation with employment growth (Campbell, 2001). A country's population's access to the Internet and participation in its electronic commerce activities reflects its access to telephone networks, computer access, and available educational opportunities and its economic means.

The e-commerce readiness assessment tool created by the Asian Pacific Economic Cooperation (APEC) Electronic Commerce Steering Group in 2000 is geared "to help governments develop their own focussed policies, adapted to their specific environment, for the healthy development of e-commerce" (Comparison of E-Readiness Assessment Models, 2001, APEC's E-Commerce Readiness Assessment). This tool has six categories to measure readiness for e-commerce:

- Basic infrastructure and technology (speed, pricing, access, market competition, industry standards, foreign investment),
- Access to network services (bandwidth, industry diversity, export controls, credit card regulation),
- Use of the Internet (use in business, government, homes),
- Promotion and facilitation (industry led standards),
- Skills and human resources (ICT education, workforce), and
- Positioning for the digital economy (taxes and tariffs, industry self-regulation, government regulations, consumer trust) (Comparison of E-Readiness Assessment Models, 2001).

Based on the result of answers to questions based on these categories, each country is assigned an e-readiness level, which indicates that country's level of free trade, industry self-regulation, ease of exports, and compliance with international standards and trade agreements. A country with higher level of e-readiness exhibits fewer barriers to electronic commerce growth.

Each of these models provides a measure against a specific goal, and as such, would be useful only if used within the right context and with the recognition of its limitations. So far, there is no one tool that combines different frameworks of existing tools into one comprehensive tool to measure various aspects of e-readiness of a region.

Advantages of Global Electronic Commerce

Governments have traditionally encouraged globalization as it provides a means to expand markets, engage in knowledge and cultural exchange, and of course, generate higher revenues in the form of higher sales, profits, and/or taxes from increased trade. With worldwide electronic commerce revenues expected to reach $6.8 trillion by 2004 (Forrester Projects $6.8 Trillion for 2004, 2002), the Internet can provide the means for long-term sustainable competitive advantage. National and multinational companies are looking for ways to claim a share of this growing market. As world economies move toward an information economy from the traditional production-based economy, the Internet influences macroeconomic and microeconomic developments positively (Iyer et al., 2002a).

One enabling factor is the ability of the Internet to boost innovation by accelerating diffusion of ideas, information, and knowledge worldwide. At the macroeconomic level, as the size of potential markets increases with globalization of electronic commerce, entrepreneurs are willing to take more risks to reap rewards for uncovering lucrative new ideas, and thus to boost innovation additionally (Iyer et al., 2002a). At the microeconomic level, this has led to innovations in one-to-one marketing utilizing the immense number of data collected over Internet transactions and using that information to customize products or services to individual customers. For example, companies such as amazon provide dynamically customized views of products (and services) to customers based on those customers' previous online surfing and purchase patterns.

A second factor is the ability of the Internet to increase market efficiency. By increasing the real-time availability of information and the ability to match vast numbers of potential buyers with sellers that best meet their needs at increasingly low marginal costs, the Internet brings in new efficiencies that can be leveraged by a company to scale up at lower costs (Iyer et al., 2002a). Global intra-e-business with well-implemented intranets can allow a company to perform its internal functions and manage its resources with greater efficiency and foster cost-efficient collaborations. Global business-to-business electronic commerce provides another opportunity to leverage the Internet to automate relationships with customers, suppliers, and partners and improve supply-chain and electronic data interchange (EDI) transaction efficiencies.

The third factor is that the Internet has facilitated greater distribution of information and services, thus making services such as education, consulting, banking, and gambling globally accessible (Iyer et al., 2002a). The

Internet has opened up a new low-marginal-cost distribution channel that is shaping supply-chain operations, especially in the business-to-business sector.

Disadvantages/Challenges of Global Electronic Commerce

Although the Internet offers unprecedented opportunities in the global markets, formulating strategies to take advantage of these opportunities is not trivial, either at the organizational or at the policy-making institutional level. Companies are often facing competitive pressures to invest in electronic commerce. In order to determine the appropriate electronic commerce strategy, a company needs to understand how it is positioned compared to its competitors. In addition to complex technology issues, it has to decide which country to invest in and consider the high costs when developing a realistic strategy. There are various direct and indirect costs of expanding business via the Internet including the cost of realigning business processes, both external and internal, training employees and partners, and managing a global workforce. There is still a notion of the Internet as a free channel of distribution and, therefore, a tendency to underestimate the associated costs and overestimate the market size forecasts and ensuing profits. This is compounded by the fact that the models to assess Internet costs and benefits are still evolving.

Electronic commerce globalization also presents unique management challenges. These range from dealing with global customers with widely varying structural, cultural, and behavioral demographics to managing geographically dispersed global employees and global teams and managing compensation issues equitably in workforces spanning national boundaries. Some management issues may also arise due to the reluctance of a team to relinquish local autonomy and share information about markets and customers with other teams on a global scale. Devising policies that find a balance between information sharing across a company to stay competitive in global market space and allowing decision-making ability at local levels is a key issue. The added challenge is the regulatory complexities involved in any international trade. In fact, this has given rise to several companies that provide access to comprehensive databases of import and export regulations of countries and offer consulting in global trade management.

The other challenges are in understanding and managing the structural conditions such as physical, social, and economic arrangements necessary for e-commerce adoption by individuals and organizations (Markus & Soh, 2002). These conditions vary widely among countries and even within countries. For example, for a business-to-consumer site in Asia, a company could encounter issues such as high variance in urban-to-rural population ratio in Asian countries, difficulties in access to the potential market due to poor transportation, and lack of other electronic commerce infrastructure. Also, due to high population densities in urban areas, most businesses usually deliver everything people need or desire to their homes and offices in a convenient and cost-effective manner. This obviates the need for online shopping from the point of view of convenience for a large part of this market. In addition, many in Asia have never used credit cards or done catalog shopping and, therefore, are not drawn to online shopping readily (Markus & Soh, 2002).

Table 3 Internet Usage Across the World

World Region	Internet Usage Ratio
Africa	0.50%
South Asia	0.40%
Arab States	0.60%
East Asia	2.33%
Latin America/Caribbean	3.33%
North America/Europe	50.00%
Global Average	6.67%

Source: Jensen (2002).

At the institutional level, government entities have to balance the policies that push for Internet diffusion for economic development with the loss of governmental controls. Another issue in electronic commerce globalization is the dilemma faced by developing and underdeveloped nations on what priority should be placed on reallocating scarce resources to create the telecommunication infrastructure necessary for electronic commerce growth. These reallocations take resources away from much more profound needs of their citizens, such as education, food, clean water, shelter, and health care.

Global Digital Divide

The term *digital divide* has come into widespread use after appearing in the 1995 U.S. Department of Commerce report *Falling Through the Net: A Survey of 'Have Nots' in Rural and Urban America* (Lu, 2001). The digital divide refers to "great disparities in opportunity to access the Internet and the information and educational/business opportunities tied to this access" (Lu, 2001, p. 1). The digital divide can be measured by inequity in access to the use of information and communication technologies such as number of telephone lines, Internet hosts, or mobile phones per inhabitant. Various aspects such as literacy rate, education facilities, geographic area, income, and other demographic factors affect the degree of digital divide. Studies show that although the digital divide exists even within wealthier nations, the gap between the information-rich and -poor within their societies has been closing rapidly for the past few years. However, the same is not true of developing or underdeveloped nations, where disparities in access to basic telecommunication infrastructures are even greater (see Table 3).

ROLE OF GOVERNMENTS IN GLOBAL ELECTRONIC COMMERCE

The political climate in a country has a direct correlation with the willingness of foreign companies to consider expanding their electronic commerce operations in that country. Companies are understandably reluctant to invest in countries prone to frequent coups and wars, or those that have widespread corruption, internal law and order problems, or fiscal unpredictability. Another factor that influences this decision is the level of government controls and regulations that a business has to deal with for its operations.

The extent of Internet diffusion in a country is a function of the political ideology and will that shape its information technology (IT) policies and the legislative and regulative policy environment, and therefore, the growth of the Internet in that country. It is heavily influenced by what opportunities and threats this diffusion poses for the sociopolitical climate of that country. For example, in China, the spread of electronic commerce presents a dilemma for the Chinese regime, as on one hand it presents huge economic opportunities, whereas on the other hand it presents the threat of knowledge diffusion that is outside of governmental control and may result in diminution of the power of the state. Thus, government policies and decisions can act as a stimulant or as an inhibitor for the growth of electronic commerce. This certainly creates a predicament for firms that acquiesce in such policies. For example, Yahoo!, which has Internet operations in China, recently came under fire from human-rights groups for complying with the Chinese laws and thus filtering out content that was deemed pernicious information jeopardizing state security and stability.

Global Internet Regulations

Global electronic commerce raises unprecedented issues in the areas of jurisprudence, applicable law (any rule of law that could be applied by any national or arbitral tribunal in the event of a dispute arising between a business and a consumer who interact online), consumer protection, and taxation of electronic commerce transactions. Although business-to-business global electronic commerce presents fewer problems with its already existing mature contractual practices among various parties, this is less so with transnational business-to-consumer electronic commerce. Various international organizations are currently engaged in dialog to come up with policy directives to deal with these concerns and achieve a consensus among participating nations. There are inherent problems with drafting regulations that would apply to all participating nations. For example, the protocol on the Cybercrime Convention on Hate Speech drafted by the Council of Europe was not signed by United States, as the First Amendment under the U.S. constitution protected some acts criminalized under the protocol.

Similar concerns exist in other areas involving electronic commerce transactions entered into by non-European Union (non-EU) firms and EU consumers. According to a letter by the U.S. Council for International Business to the European Commission (USCIB Letter, 2002), there are a number of concerns on the proposed imposition of value-added tax (VAT) on non-EU firms engaging in electronic commerce with EU consumers, as it unduly burdens these firms.

As a consequence of these uncertainties, many companies prefer not to incur expenses related to compliance with a myriad of rules, instead limiting themselves to conducting business with partners and consumers of countries that have well-established policies on such matters. This of course inhibits the growth of global electronic commerce, as a large number of consumers worldwide are denied access to online marketplaces.

Web site content regulations and strict governmental controls over Web content also have an adverse effect on electronic commerce. Countries that have seen the most growth are adopting the model of Internet service provider (ISP) self-regulation combined with user-empowering technologies such as Web filtering and monitoring software to provide a balanced and flexible solution to content control.

Consumer trust is an important ingredient for the adoption and success of electronic commerce practices and is influenced by tangible factors such as privacy and security of networks and encryption policies, as well as intangible factors such as branding, consumer empowerment, industry accreditation systems and regulations, industry self-regulation, and consumer protection laws.

Taxation codes in each country, and tax treaties among nations to deal with globalization, had originally been formulated when even the notion of a digital economy with its borderless attributes was not around. Taxation systems in response to electronic commerce are just beginning to evolve. With increasing globalization, there is a critical need for adoption of a worldwide tax base for corporations with operations under different local tax laws to avoid double taxation, and to provide incentives to attract foreign direct investment. An additional issue is the non-taxation of online transactions (which reduces the tax revenues for tax administrators), but the physical delivery of the goods across national or state borders may still have substantial tariffs and import/export restrictions attached to it. Monitoring taxation for online transactions would require sufficient infrastructure on the part of the tax governing bodies, which may not be feasible for most developing countries for some time. The Organization for Economic Co-operation and Development (OECD) has published various reports that highlight challenges posed by global growth of the Internet in issues related to regulating cross-border financial services and taxes.

Legal Issues in Global Electronic Commerce

For global electronic commerce growth, laws are necessary to protect the individual's and company's right to conduct electronic commerce safely and securely. There are specific legal issues regarding the protection of intellectual property and rights to privacy and security.

Intellectual property is defined by the World Intellectual Property Organization (WIPO) as creations of the mind: inventions, literary and artistic works, and symbols, names, images, and designs used in commerce. Intellectual property is divided into two categories: industrial property, which includes inventions (patents), trademarks, and industrial designs; and copyright, which includes literary and artistic works such as novels, poems and plays, films, musical works, artistic works such as drawings, paintings, photographs, and sculptures, and architectural designs (About Intellectual Property, 2002).

Intellectual property laws that pertain to copyrights, patents, and trademarks give companies exclusive rights that can provide competitive advantages. Before the advent of the Internet, intellectual law had sovereign boundaries. However, in the digital world, such territorial boundaries disappear, raising questions regarding jurisdiction on these issues. The Internet, by the very nature of its technology, makes a "temporary" copy every time material is accessed, which stand in direct violation of copyright laws. Governments are beginning to address this issue. The 1996 Trade Related Aspects of Intellectual Property (TRIPS) agreement defines standards of intellectual property rights and enforcement mechanisms for individual countries to follow and incorporates these into the existing dispute settlement mechanism of the World Trade Organization (WTO) (TRIPS, 2002).

To implement TRIPS, the United Stated passed the Digital Millennium Copyright Act (DMCA) in 1998, which also addresses several other areas such as the circumvention of copyright protection systems, fair use in a digital environment, and Internet service provider liability (including details on safe harbors, damages, and "notice and takedown" practices). In Europe, the United Kingdom has passed three laws, the Data Protection Act 1998 (DPA), the Electronic Communications Act 2000 (ECA), and the Regulation of Investigatory Powers Act 2000 (RIPA). These address several issues including legality of digital signatures, rights of enforcement agencies to read encrypted data, and regulations regarding the use of personal data. One of the problems in electronic commerce transactions has been the potential security breach due to authentication problems (i.e., ensuring that the sender of an electronic message is correctly identified and validated). Digital signatures, digital encryption codes attached to transmitted messages that can authenticate the senders, are emerging as a solution to confirm that a digital signature will satisfy the requirements for a signature. The EU adopted a Directive on Electronic Signatures in 1999 and the United Nations Commission on International Trade Law (UNCITRAL) adopted the Uniform Rules on Electronic Signatures in the summer of 2001. Other parts of the world have been slow to adopt such legislation but are beginning to address intellectual property issues.

Trademark rights are another potentially thorny issue, as conflicts may arise when the same or similar trademark may be owned by companies in different countries or different standards that may exist in different countries to determine trademark infringements. Domain names, the Web addresses used to uniquely identify electronic commerce sites, are an example of trademarks that may give rise to potential problems in different countries. Domain name disputes are subject to the Uniform Domain Name Dispute Resolution Policy (UDRP), which was adopted by the Internet Corporation for Assigned Names and Numbers (ICANN) and is administered by the WIPO.

PRIVACY AND SECURITY ISSUES IN GLOBAL ELECTRONIC COMMERCE

As companies rely on global networks to do business, privacy and security of information becomes a critical issue, for consumers as well as companies. It is estimated that half of all small and medium-sized businesses will fall victim to Internet attacks by 2003 (McCusker, 2001). Companies from other countries may comply with their country's laws, but those laws may not conform to a standard that will adequately protect security of information. Countries eager to realize the promises of global electronic commerce are now formulating laws for Internet-security-related crimes. The treaty signed by 30 countries at Budapest in November 2001, the Convention of Cybercrime, is aimed at addressing the lack of common standards. "The treaty has a threefold aim: to lay down common definitions of certain criminal offences relating to the use of the new technologies, to define methods for criminal investigations and prosecution, and to define methods for international communication. The criminal offences concerned are as follows:

Those committed against the confidentiality, integrity and availability of computer data or systems (such as the spreading of viruses);

Computer-related offences (such as virtual fraud and forgery);

Content-related offences (such as the possession and intentional distribution of child pornography);

Offences related to infringements of intellectual property and related rights" (The Convention on Cybercrime, 2002).

The growth of global electronic commerce is also impeded by the lack of global patent and copyright protection and the lack of an international body to enforce the rights of companies without infringing on the freedom of governments to pursue their own goals.

Privacy Issues

Privacy refers to the right of a person to control his or her own personal information (Head & Yuan, 2001). In the United States, widespread industry resistance to government regulation has persuaded the Federal Trade Commission (FTC) to adopt a passive approach to enacting laws to ensure privacy, instead relying on industry to regulate itself around five basic principles of notice, choice, access, security, and redress. One of the most popular self-regulatory mechanisms is TRUSTe, a nonprofit initiative whose main purpose is to ensure that sites are adhering to the privacy notices posted by using random member site checks for compliance with privacy policies. The privacy laws in the 15 member states of the EU are markedly different. Instead of being self-regulated, European privacy laws are government regulated through the Data Protection Directive adopted in 1998. European rules forbid the transfer of personal data to a country that does not provide a level of protection similar to its own.

These opposing schools of thought on privacy threaten to hamper the ability of U.S. companies to engage in e-commerce with the EU countries without the risk of the U.S. businesses incurring penalties (McKenna, 2001). Therefore, the U.S. companies could be denied access to information from their own European subsidiaries or other companies located in Europe. On a global scale, personal data on millions of individuals could be sent through the Internet to companies worldwide where privacy laws may vary from stringent to ambiguous to nonexistent. The current state of these laws raises the specter of serious setbacks to companies located in non-stringent-privacy-law countries trying to engage in electronic commerce activities in countries with rigorous privacy laws.

In July 2000, the EU and the United States compromised on the Safe Harbor Agreement as a way to bridge the policy gap in which U.S. companies can avoid sanctions if they develop a self-regulatory privacy program adhering to the Safe Harbor's requirements. The seven primary principles of the Safe Harbor Agreement are notice, choice, onward transfer, security, data integrity, access, and enforcement (Shimanek, 2001). Recently the Federal Trade Commission (FTC) investigated Microsoft as part of the FTC's enforcement of the Safe Harbor Act. Microsoft came under scrutiny for not taking adequate steps to protect the privacy of Microsoft Passport Service's U.S. customers. In the settlement reached in August 2002, Microsoft consumer privacy laws would be under governmental oversight for the next 20 years. This is signaling a shift in the accommodation of international companies in adhering to the laws of each nation in the direction of global information privacy law standards to be consistent with privacy laws of the EU (Rudraswamy & Vance, 2001). A recent example of a bill under consideration by the California legislature (as of August 2002) to limit the use of a customer's personal data by financial companies may raise the bar for other states within the United States.

Transborder data flows or TDBFs are the exchange of personal and intracorporate data across national boundaries that include operational data, personnel data, electronic transfer of money, and technical and scientific data (Chen, Tu, & Lin, 2002; Rudraswamy & Vance, 2001). Because TDBFs involve personal data, due to the differences among countries in the level of protection accorded to the privacy of individuals and legal entities, TDBFs laws vary across national boundaries, posing serious challenges to formulation of global data standards to ensure global data sharing. Many countries have stringent TDBFs laws prohibiting personnel data transfer, raising difficulties in global human resource management for multinational corporations wishing to expand into new markets (Chen et al., 2002).

Security of Information Issues

Security of information refers to the ability of the *owner* of information to control that information (Head & Yuan, 2001). Some examples of security violations are access control problems, where an unauthorized party gains access to sensitive information (e.g., payroll records); integrity problems, where the contents of a message between two parties are altered without their knowledge; authentication problems, where it is not possible to ascertain a sender's identity; nonrepudiation problems, where the sender or receiver of a message can deny transmittal or receipt; and availability problems, where unauthorized parties can make use of system assets. Lack of security has major repercussions for a company; it may lose the trust of its customers, affecting its brand, lose revenues, and incur costs arising out of security breaches, thus losing strategic advantage.

Worldwide, companies use various security models to keep their systems and Internet transactions secure. These include the use of public key encryption, IP security, Web security based on SSL/SET (Server Sockets Layer/Secure Electronic Transactions), and firewalls.

Encryption/decryption policies are another leading issue in the facilitation of electronic commerce. Secure online transactions are an imperative to ensure authenticity and integrity of information exchange among parties. Whereas businesses want to use encryption that cannot be decrypted by any unauthorized entity, governments worry that their inability to intercept and decipher such messages would result in more malicious activities and threats to national security. This is one of the reasons that governments are reluctant to share these technologies and have specific policies in place to prevent this. In the United States, the Department of Justice controls the export review process and the National Security Agency (NSA) remains the final arbiter of whether to grant encryption products export licenses. Some countries have not formulated any encryption import/export policies, whereas others such as Singapore have stringent controls in place, including controls on domestic use of cryptography.

INFRASTRUCTURE ISSUES IN GLOBAL ELECTRONIC COMMERCE

Global electronic activities such as EDI (electronic data interchange) have existed for over 25 years. However, the supporting technological infrastructure was expensive and therefore usually only affordable by larger corporations. The Internet revolution allowed the emergence and growth of inexpensive and flexible infrastructure that has allowed smaller companies and individual entrepreneurs to enter into this market space and thus accelerate the growth of global electronic commerce.

When the technical infrastructure needs of a country to foster electronic commerce growth are considered, regional policies, political stability, cultural uniqueness, economic viability, and its neighbors and major trading partners all play an important role. This also raises some pertinent questions such as whether, for a developing country, it is truly worth investing in electronic commerce infrastructure at the expense of other, more basic infrastructure needs.

Telecommunication Infrastructure

Telecommunication hardware and software are necessary precursors to any electronic commerce activity and are usually representative of a country's technical competency. The specifics of various communities, regions, cultural beliefs, market practices, or income levels in a nation may affect the level of that nation's telecommunication infrastructure. Local topology plays a major role in determining whether or not laying down the infrastructure is feasible or how evenly the infrastructure is spread out spatially. Weather and climate anomalies affect the successful stringing of telephone wires or laying down of underground fiber optic networks. As telecommunication companies are increasing the available options in Internet services, this is driving down the costs, further spurring local demands for Internet access.

Cost of local calls can sometimes be a prohibitive factor in Internet usage. For example, whereas in the United States, calls to local Internet service providers (ISPs) are free, this is not the case in most other countries. Several countries such as Argentina, Chile, Belize, Colombia, Barbados, and Uruguay have come up with innovative solutions such as public Internet terminals and cybercafés to solve the problems of high costs of owning necessary hardware and software for high-speed Internet access for low-income citizens and those in remote areas. The lack of uniform worldwide technical standards and electronic transmission standards is a key challenge for companies aspiring to engage in global electronic commerce. Overly rigid and bureaucratic regulations can stifle the growth of telecommunication infrastructure and limit the speed of technology diffusion and market expansion. Liberalization leads to increased access and affordability of services, which accrues wider societal benefits.

Uniform technical standards are necessary to ensure interconnectivity and interoperability of information technology systems between countries engaging on electronic commerce. The Global Information Infrastructure Commission (GIIC) has determined that inadequate standards and the lack of interoperability and interconnection will delay the development of the GII and electronic commerce. Several organizations such as the International Telecommunication Union (ITU), the International Standards Organization (ISO), and the European Telecommunications Standardization Institute (ETSI) are cooperating to study this issue from a global perspective and develop technical standards through cooperation.

International domain names are called country-code top level domains (ccTLD Resource Materials, 2002) and they are allocated to each sovereign nation or territory of a nation based on a unique two-character code through the Internet Corporation for Assigned Names and Numbers (ICANN). Although this body has successfully implemented a standardized process for assigning domain names, there is still the problem of the disproportionate numbers of domain names being in English and the limited interoperability between translations of these domain names into the target languages of the countries.

Fulfillment Infrastructure

The development of order fulfillment infrastructure in the cycle of an online transaction is another necessary step to allow delivery of physical goods ordered online. The real-time nature of electronic commerce has raised consumer expectations regarding online purchasing behavior (Ahuja, Gupta, & Raman, forthcoming). An important objective of a successful global electronic commerce business is to deliver products and services across borders as quickly and dependably as possible. Thus rapid customs clearance is critical in the electronic commerce value chain, and the absence of a preclearance system is going to result in the loss of electronic commerce efficiencies. The lack of reliable transportation systems to ensure speedy and safe deliveries is a major reason for slower electronic commerce growth in regions outside of North America and Europe. Growth in other regions is only possible if business and customer clearance services are adequate and if the physical distribution system is reliable, fast, and affordable.

FINANCIAL ISSUES IN GLOBAL ELECTRONIC COMMERCE

Robust financial institutions, particularly the stability of the banking sector, are a prerequisite for the functioning of the modern world economy, which is also increasingly interdependent. Some of the issues that affect the growth of global electronic commerce are the state of online financial institutions and services, policies relating to these financial institutions, and the availability of financial capital funding for global e-commerce. E-commerce requires specialized support in financial services and an advance payment system.

Currency Issues

An online transaction is not complete until the business receives the payments in the correct currency. Given the different currencies that are used throughout the world, companies will need e-banking services that can operate in such an environment. In addition, exchange rate fluctuations affect a corporation's overall financial performance. Banking services that can facilitate these kinds of financial services are attractive to companies, as these can boost their financial performance by diluting currency exchange and funds transfer risks. Some companies such as WorldPay, Inc., provide solutions that can enable a company to conduct online transactions in 130 currencies.

Online Payment Systems

Commercial activity, especially between a business and an individual, is still largely on a cash basis in a large part of the world. Transactions on a credit basis (e.g., using credit cards or checks) may only be available in a few large metropolitan cities. In these countries, people routinely carry huge amounts of cash in their wallets, as that is the only mode of payment for buying a cup of tea or truckloads of material for building a house. Against such a background, it requires a tremendous change in financial processes as well as behavioral patterns to enact successful electronic commerce activity. Given that even in countries where credit transactions have been the norm for decades, consumers are still reluctant to pay online to unseen merchants for a variety of reasons, the problem is exacerbated in countries where the use of credit is still a nascent practice. Another issue in some countries is that consumers are used to paying upon delivery of goods/services and, therefore, online transactions requiring upfront payment for later delivery raise trust issues and have still to gain widespread acceptance.

The three key drivers of online payment system success are merchant acceptance, authentication, and security. In the United States, the recent enactment of the Electronic Signatures in Global and National Commerce Act (E-SIGN) provides an increased impetus for the growth of domestic electronic commerce, as well as incentives for other countries to follow suit. The E-SIGN Act provides the same certainties for conducting online business as exist for doing offline.

CULTURAL ISSUES IN GLOBAL ELECTRONIC COMMERCE

The Internet has offered unparalleled global exposure to companies through Web sites, intranets, and extranets. This exposure, although affording new opportunities and global distribution capabilities, also raises issues of how to cater to local markets by taking into account local cultures, languages, and behavior.

Localization

Localization refers to the unique set of requirements within each country, and sometimes within regions in a country, that combine jurisdictional requirements with local consumer needs, social and cultural patterns, behavior, beliefs, and language(s). This also raises the stakes in Web site design issues on ensuring international and local usability. Cultural issues affect a country's new technology adoption patterns (Palvia, Palvia, & Whitworth, 2002; Straub, Loch, & Hill, 2001). The nature of social networks affects the perception of change that any new technology brings in its wake. For example, is the prevailing culture resistant to new technologies because of perceptions of associated job loss, changes in social dynamics, or loss of a way of life? Localization matters for innovation, and a global company must develop localization strategies to take advantage of the local strengths, including local experts, and to proactively enhance the environment for innovation and commercialization in locations where it operates (Farhoomand et al., 2000; Porter, 2001).

Language, Geography, and Religion

For a company targeting a country for its products and services, language is one of the most obvious concerns. Every year, the number of non-English-speaking Internet users is increasing; in the year 2000, they outnumbered English speaking Internet users for the first time. Until quite recently, one problem with displaying letters and characters of a language had to do with the lack of a standard encoding mechanism in computer systems. Characters other than standard ASCII displayed on Web pages were traditionally encoded in HTML through the use of character entity references, so that an umlauted lower case a (ä), for example, was encoded as &auml or its numerical equivalent (&#228;). Although this system worked for major Western European languages, it did not for non-roman-alphabet-based languages such as Chinese, Cyrillic, and Japanese. With the emergence of the Unicode Standard and the availability of tools supporting it, this is likely to lead to more Web sites being able to render characters in a particular language's character set. Another level of language concern is the display of Web site content in users' native language in a way that takes into account its cultural context and incorporates its nuances. This is much harder to address without significant spending of resources to get it right and avoid linguistic faux pas. The multiplicity of languages in Asia–Pacific Rim and European countries has been a major obstacle to achieving business-to-business operation economies of scale (Kshetri & Dholakia, 2002).

In designing a Web site for the users of a country, another thing to keep in mind is the time zones used there. Different countries write dates differently; for example, 5/10/02 mean completely different things in the United States and in India. Different countries have different religions, and that in turn provides a distinct flavor to the yearly calendar in use there. The use of colors in a Web site is an integral part of Web site design. It is also open to cultural interpretation, as color symbolisms vary across cultures, and therefore requires understanding of the cultural context in which a site is going to be viewed.

Culture

Differences in cultural beliefs of a society are well known to influence the way people react to the use of technology and information system adoption, development, and implementation (Hofstede, 1997; Kaarst-Brown & Evaristo, 2001; Palvia, Palvia, & Whitworth, 2002; Straub et al., 2001). For example, despite technical and regulatory infrastructure in place, and high use of the Internet for communication purposes, Europe is far behind the United States in the use of Internet for electronic commerce. The reluctance of Europeans to shop online and engage in electronic commerce could be of cultural origin. Cultural differences among nations also are related to the degree of privacy sensitivity and privacy concerns for data protection of their citizens or corporations and are known to strongly correlate to with the regulatory approach of a nation (Rudraswamy & Vance, 2001). Also, global companies need to recognize that local cultures and beliefs influence organizational effectiveness and, therefore, need to adopt appropriate leadership styles and organizational structure suitable for each culture (Okunoye & Karsten, 2002).

There is, however, a need to be cautious in regard to the understanding of what constitutes a "national culture" (Myers & Tan, 2002; Straub, Loch, Evaristo, Karahanna, & Srite, 2002). Emerging research in this area suggests "that IS researchers interested in conducting research on culture and globalization should adopt a more dynamic view of culture—one that sees culture as contested, temporal and emergent" (Myers & Tan, 2002, p. 24).

STRATEGIES FOR THE GROWTH OF GLOBAL ELECTRONIC COMMERCE

A commitment to entry into the information economy guided by policies that emphasize both market forces and government structures is likely to prove to be a better strategy than relying on just one or the other alone. ICTs are enablers of globalization, boosting the movement of tangible and intangible assets over connected networks across national boundaries. Poor telecommunication infrastructure in a country may result in that country lagging behind even further in the emerging global digital economy, widening the digital divide. Countries that have restructured state services to provide better access to ICTs for their citizens and have formulated government programs to develop necessary skills through increase in education levels, job training, adult education, and in particular, programs that invest in achieving higher level skills in science and technology for long-term advantages have been able to achieve higher Internet growth rates and growth in related markets (Campbell, 2001). The growth of research and development and entrepreneurial activities is tied into the availability of high-level skills. Therefore, investment in higher education is an integral part of policies; for example, the growth of the ICT industry in Bangalore in India, Xian in China, and Silicon Valley in the United States is tied to their close proximity to well-known universities (Campbell, 2001).

In just the past decade, many countries have adopted strategies that seek to facilitate market access through privatization of the telecommunication industry, the extent of which varies among the many world regions. Privatization drives down prices, making telecommunication infrastructure more affordable, more accessible, and of improved quality. Governments are also providing tax incentive subsidies to attract foreign direct investments and to procure capital to compensate for lack of venture capital within the country. Similar strategies being adopted worldwide by various governments consist of providing tax incentives for companies to invest in research and development initiatives and continuing education in information-technology-related areas, providing tax subsidies for software and hardware development and for sharing scant telecommunication resources to optimize their collective usage, building a local pool of knowledge workers to spur local entrepreneurial activities and facilitating access to local and international financial capital and markets to build domestic and international demand. Another way for governments to spur corporate investment in Internet technologies is to adopt these technologies for governmental business, and thus provide a model of effectiveness of these technologies as well as providing accessibility and transparency.

Implementing some of these strategies has helped several developing countries to become increasingly integrated into the global economy through growth in the export of software and skill-intensive software services, such as call-centers. India is one such example. According to the case study in the final report of the Digital Opportunity Initiative produced in July 2001, the Indian software industry has grown at a staggering annual rate of over 50%, from a mere $150 million in 1991–1992 to $5.7 billion in 1999–2000 (Appendix 3 National ICT Approaches, 2002).

CONCLUSIONS AND FUTURE DIRECTIONS

Global electronic commerce, with its promise of globalization on an unprecedented scale, raises some pertinent issues about the equitable distribution of its benefits among developed and developing nations and closing the digital divide among them. Strategies formulated within the context of a holistic understanding of the specific environment within a country will be the key to the overall success of Internet diffusion there. Strategies that provide for integration and cooperation between the public and the private sectors are essential for the successful transfer of knowledge and technological infrastructure needed for e-commerce growth. Global e-commerce can only become a reality if the underlying technology is diffused throughout a society. Countries that have been most successful in harnessing the Internet have done this as a result of the combined efforts of government and the private sector through high investments in higher education and human capital, building infrastructures through deregulation, and providing a climate that promote fair competition and economic freedom. Governments that have yet to formulate national policies with clear and stable frameworks to reduce complexities and costs of participating in the global digital economy may be placing limitations on the growth of entrepreneurial ventures in the digital marketplace and denying their citizens access to the benefits of the Internet. Given the imperative need to combat the digital divide, various international coalitions such as Asia–Pacific Economic Cooperation (APEC) have formed to combat it and to improve the quality of life for disadvantaged people by 2010.

Longitudinal studies on the impact of the Internet and electronic commerce on the economic development at a societal and individual level are needed. This research can help developing countries formulate and adopt policies and strategies that can help narrow the digital divide and focus on providing basic and higher education, producing local and relevant content, and effectively using technology throughout society to provide opportunities to realize the benefits of electronic commerce for their citizens.

ACKNOWLEDGMENT

The author would like to thank the anonymous referees for their helpful suggestions and comments to improve this paper.

GLOSSARY

Authentication The process of ensuring the identity of another entity on the Internet.

Business-to-business (B-to-B) A term that describes transactions among businesses.

Business-to-consumer (B-to-C) A term that describes transactions between an individual buyer and a business.

Digital certificates Software provided by a trusted third-party certification authority to authenticate the holder of a public encryption key.

Domain name system (DNS) Maps Internet addresses. An Internet host needs a domain name that has an associated Internet protocol (IP) address record. The DNS is a database system that looks up host IP addresses based upon domain names.

E-readiness Refers to a country's ability to take advantage of the Internet as an engine of economic growth and human development by utilizing its telecommunication infrastructure, human resources, financial resources, and legal and regulatory policies.

Electronic data interchange (EDI) A format for exchanging business data (e.g., purchase orders) among organizations developed by the Data Interchange Standards Association to facilitate commerce.

Encryption A technology using complex mathematical algorithms to encode and decode information for security purposes during transmission. Only those who have an authorized decryption key can decipher an encrypted message.

Firewall Hardware and software systems that isolate a private network from the public network by providing a location for monitoring security-related events and allowing only authorized traffic to pass.

Globalization Refers to the process of diffusion of people and their knowledge across nations, cultures, and economies.

Global digital divide Refers to asymmetry or inequity in the distribution, access, and use of information and communication technologies among world regions.

Global electronic commerce The buying and selling of business services over telecommunication networks where the two parties involved in this exchange are in different countries.

Information and communication technologies (ICTs) Refer to numbers of devices such as telephone lines, Internet hosts, or mobile phones per inhabitant that provide a measure of Internet diffusion in a country.

IP security An IP level security protocol that secures all TCP/IP traffic between two networks by encrypting each IP packet that leaves the system and authenticating packets that enter.

One-to-one marketing A form of marketing where a firm can customize its products and services to fit each customer's unique needs and characteristics.

Public key encryption Asymmetrical key encryption that uses a pair of public and private keys to ensure security and authenticity of a transmitted message online. A message is first encrypted by the sender using his or her private key and then using the widely available public key of the message recipient. This encrypted message can then be decrypted only by the recipient using a LIFO order of decryption—first using his or her private key, and then decrypting using the sender's widely available public key.

Secure electronic transmission (SET) A security protocol embedded in a users' Web browser and in Web servers to encrypt all online communications. It is designed to be transparent to the end user.

Secure socket layer (SSL) A security protocol initiated by Visa, MasterCard, Microsoft, Netscape, and others and designed to protect credit card transactions on the Internet. SSL allows client/server applications to communicate using encryption, where servers are always authenticated and clients are optionally authenticated. It sets up a secure end-to-end link over which http or any other application protocols can operate.

Supply chain Refers to a set of activities related to ordering, manufacturing, fulfillment, and payment flow that involve an organization, its suppliers, contract manufacturers, and distributors, and the end customers.

United Nations Commission on International Trade Law (UNCITRAL) Established by the United Nations General Assembly in 1966 to reduce or remove obstacles to international trade created by disparities in national laws. Its mandate is to work toward a progressive harmonization and unification of the law of international trade.

CROSS REFERENCES

See *Digital Divide; Feasibility of Global E-business Projects; Global Diffusion of the Internet; International Cyberlaw; Politics.*

REFERENCES

About Intellectual Property (2002). Retrieved April 11, 2002, from http://www.wipo.org/about-ip/en/

Ahuja, M., Gupta, B. & Raman, P. (forthcoming). An empirical investigation of online consumer purchasing behavior. *Communications of the ACM,* forthcoming.

Appendix 3 National ICT Approaches: Selected Case Studies—India. Retrieved April 11, 2002, from http://www.opt-init.org/framework/pages/appendix3Case4.html

Boudreau, M., Loch, K. D., Robey, D., & Straud, D. (1998). Going global: Using information technology to advance the competitiveness of the virtual transnational organization. *The Academy of Management Executive, 12*(4), 120–128.

Campbell, D. (2001). Can the digital divide be contained? *International Labour Review, 140*(2), 119–141.

ccTLD Resource Materials (2002). Retrieved May 14, 2002, from http://www.icann.org/cctlds/

Chen, Q., Tu, Q., & Lin,. B. (2002). Global it/is outsourcing: Expectations, considerations and implications. *Advances in Competitiveness Research, 10*(1), 100–111.

Comparison of E-Readiness Assessment Models, Final Draft, v. 2.13, (2001, March 14). Retrieved April 11, 2002, from http://www.bridges.org/ereadiness/report.html

The Convention on Cybercrime, a Unique Instrument for International Co-Operation. Retrieved May 9, 2002, from http://www.coe.int/T/E/Communication_and_Research/Press/Themes_Files/Cybercrime/e_CP893.asp#TopOfPage

Daly, H. E. (1998). Population and economic globalization. *Organization and Environment, 11*(4), 438–441.

de la Torre, J., & Moxon, R. W. (2001). Introduction to the Symposium on E-Commerce and Global Business. *Journal of International Business Studies, 32,* 617–700.

Farhoomand, A. F., Tuunainen, V. K., & Yee, L. W. (2000). Barriers to global electronic commerce: A cross-country study of Hong Kong and Finland. *Journal of Organizational Computing and Electronic Commerce, 10*(1), 23–48.

Forrester Projects $6.8 Trillion for 2004 (2002). Retrieved May 5, 2002, from http://www.glreach.com/eng/ed/art/2004.ecommerce.php3

Head, M., & Yuan, Y. (2001). Privacy protection in electronic commerce—A theoretical framework. *Human Systems Management, 20,* 149–160.

Hofstede, G. (1997). *Cultures and organizations: Software of the mind.* New York: McGraw-Hill.

Howells, J., & Mitchie, J. (1998). Technological competitiveness in an international arena. *International Journal of the Economics of Business, 5*(3), 279–193.

Iammarino, S., & Mitchie, J. (1998). The scope of technological globalisation. *International Journal of the Economics of Business, 5*(3), 335–353.

Iyer, L. S., Taube, L., & Raquet, J. (2002a). Global e-commerce: Rationale, digital divide, and strategies to bridge the divide. *Journal of Global Information Technology Management, 5*(1), 43–69.

Iyer, L. S., Gupta, B., & Taube, L. (2002b). Global e-commerce: Analysis of growth limitations. Paper presented at the 3rd Annual Global Information Technology Management (GITM) World Conference, June 23–25, 2002, New York.

Jensen, M. (2002). The African Internet—A status report. Retrieved May 13, 2002 from http://demiurge.wn.apc.org/africa/afstat.htm

Kaarst-Brown, M. L., & Evaristo, J. R. (2001). The role of culture in global electronic commerce. In P. Palvia et al. (Eds.). *Global Information Technology and Electronic Commerce: Issues for the new Millennium* (pp. 255–274). Marietta, GA: Ivy League Publishing, Ltd.

King, J. L., Gurbaxani, V., Kraemer, K. L., McFarlan, F. W., Raman, K. S., & Yap, C. S. (1994). Institutional factors in information technology innovation. *Information Systems Research, 5*(2), 139–169.

Kshetri, Nir (2001). Determinants of the locus of global e-commerce. *Electronic Markets, 11*(4), 250–257.

Kshetri, N., & Dholakia, N. (2002). Determinants of the global diffusion of B2B e-commerce. *Electronic Markets, 12*(2), 120–129.

Lu, M. (2001). Digital divide in developing countries. *Journal of Global Information Technology Management, 4*(3), 1–4.

Markus, M. L., & Soh, C. (2002). Structural influences on global e-commerce activity. *Journal of Global Information Management, 10,* 5–13.

McCusker, R. (2001). E-commerce security, the birth of technology, the death of common sense? *Journal of Financial Crime, 9*(1), 79–90.

McKenna, A. (2001). Playing fair with consumer privacy in the global online environment. *Information and Communications Technology Law, 10*(3), 339–355.

Montealegre, R. (1999). A temporal model of institutional interventions for information technology adoption in less-developed countries. *Journal of Management Information Systems, 16*(1), 207–232.

Myers, M. D., & Tan, F. B. (2002). Beyond models of national culture in information systems research. *Journal of Global Information Management, 10*(1), 24–32.

Okunoye, A., & Karsten, H. (2002). Where the global needs the local: Variation in enablers in the knowledge management process. *Journal of Global Information Technology and Management, 5*(3), 12–31.

Palvia, P. C., Palvia, S. C., & Whitworth, J. E. (2002). Global information technology: A meta analysis of key issues. *Information and Management, 39,* 403–414.

Porter, M. E. (2001). Innovation: Location matters. *MIT Sloan Management Review, 42*(4), 28–36.

Rudraswamy, V., & Vance, D. A. (2001). Transborder data flows: Adoption and diffusion of protective legislation in the global electronic commerce environment. *Logistics Information Systems, 14*(1–2), 127–136.

Shimanek, A. E. (2001). Do you want milk with those cookies? Complying with the safe harbor privacy principles. *Journal of Corporation Law, 26*(2), 455–477.

Straub, D., Loch, K., & Hill, C. (2001). Transfer of information technology to developing countries: A test of cultural influences modeling in the Arab world. *Journal of Global Information Technology Management, 9*(4), 6–28.

Straub, D., Loch, K., Evaristo, R., Karahanna, E., & Srite, M. (2002). Toward a theory-based measurement of culture. *Journal of Global Information Management, 10*(1), 13–23.

TRIPS (2002). Retrieved May 6, 2002 from http://www.wto.org/english/thewto_e/whatis_e/tif_e/agrm6_e.htm

United Nations Conference on Trade and Development Report (2002). E-commerce and development, Chapter 1, Table 7. Retrieved November 19, 2002, from http://r0.unctad.org/ecommerce/ecommerce_en/edr02_en.htm

USCIB Letter on Imposition of a Value Added Tax on Electronic Commerce Transactions (2002, February 7). Retrieved May 7, 2002, from http://www.uscib.org/index.asp?documentID = 1950

Venkatesh, A. (1999). Postmodernism perspectives for macromarketing: An inquiry into the global information and sign economy. *Journal of Macromarketing, 19*(2), 153–169.

Wolcott, P., Press, L., McHenry, W., Goodman, S., & Foster, W. (2001). A framework for assessing the global diffusion of Internet. *Journal of the Association for Information System, 2,* Article 6. Retrieved December 9, 2001, from http://jais.aisnet.org/contents.asp

Groupware

Pierre A. Balthazard, *Arizona State University West*
Richard E. Potter, *University of Illinois at Chicago*

INTRODUCTION

Groupware refers to computer- and networked-based technologies that help people accomplish a group's objective. Some forms provide greater support for real-time collaboration than others (such as electronic meeting support systems, chat rooms, shared whiteboards, and real-time videoconferencing), some can support the individual's work, as well as work that can only be accomplished collaboratively (e.g., collective writing). Helping people who are in the same place at the same time (e.g., meetings) is the forte of some types of groupware, and other forms excel at connecting group members who might be continents and time zones apart (such as e-mail and shared database access). Some are geared to support specific group subtasks (e.g., shared calendaring and scheduling, e-mail), whereas others are designed for a wider array of integrated support functions (such as dispersed project management). To make this more complicated, the Internet has driven both the amazing popularity and adoption of these tools, but it has also driven the creation of new classes of tools that are proving important (e.g., shared databases). In this chapter, we first remark on the for purpose of groupware and then present some defining concepts and classifications for these myriad tools. We then explore some generic and common functionalities of groupware tools and turn to detailed descriptions of some popular modern examples and show how they map to our classifying scheme. We then adopt a management perspective and present a theory-based framework for mapping these functionalities to the varying circumstances of individual and group work. The final part of the chapter addresses issues of human productivity and groupware as well as issues surrounding the changes in organizational and human processes that often accompany the proliferation of groupware-supported work.

WHY GROUPWARE?

There has been more than five centuries of research devoted to group and team processes in an effort to understand why the desired synergy of group work is often elusive and why instead group work is often difficult and inefficient. Figure 1 lists the most common positive and negative influences on collaborative work.

Many things can go wrong with group work. Participants may fail to understand their goals, lack focus, or have hidden agendas. Some people may be afraid to speak up, and others may dominate the discussion. Misunderstandings occur through different interpretations of language, gesture, or expression. Besides being difficult, teamwork is expensive. A meeting of several managers or executives may cost several thousands of dollars per hour in salary costs alone. In Fortune 500 companies, there are more than 12 million formal meetings per day in the United States—more than 3 billion meetings per year. American managers typically spend about 20% of their time in formal meetings and up to 85% of their time communicating. For all its difficulties, group work is still essential. People must collaborate to solve tough problems. As business becomes more global in scope and computers become more ubiquitous in the workplace, the need for collaboration—and the means to achieve it—has surely continued to increase. Enter groupware, with functionalities that—when appropriately selected, applied, and managed—can bring users tremendous support for their group's tasks (Collaborative Strategies, 2002).

DEFINING AND CLASSIFYING GROUPWARE IN THE INTERNET AGE

Groupware is technology—typically software and networks—designed to support people working in groups. Under this broad definition, any technology that is used to communicate, cooperate, coordinate, solve problems, compete, or negotiate is classified as groupware. This definition can apply to some of the newest forms, such as informational Web pages and multiuser database applications, as well as to more established and familiar forms such as e-mail, newsgroups, videophones, or chat. The simplicity and breadth of this definition is intentional because definitions of groupware tend to evolve as rapidly as the technologies themselves, often becoming

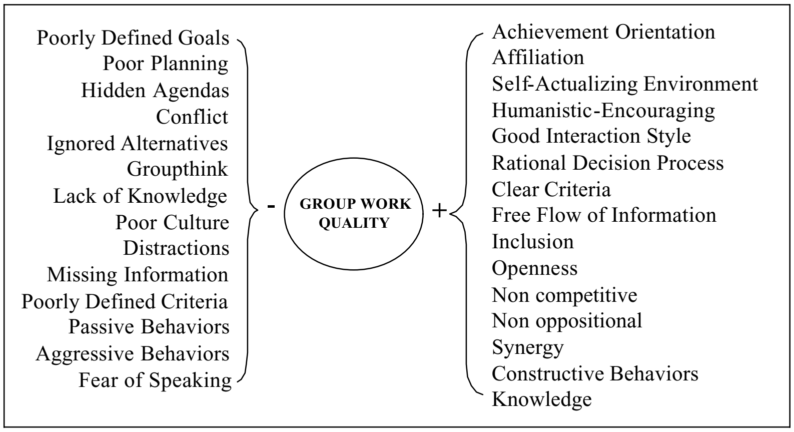


Figure 1: Influences on the quality of group work.

outdated because new technologies support group work in ways that were not represented. For example, an early definition for groupware was intentional group processes plus software to support them. At the time, this definition placed more emphasis on the activities of the group and much less on the technologies, which were limited. As we describe, although much of that early definition is valid, the reality that it does not address is that much group work contains a great deal of individual work. Many modern forms of groupware support not only the actual group work, but also much of the individual group members' work that may precede, be contemporaneous with, or follow that actual interactive group processes.

A later definition of groupware by Johansen (1988) is "specialized computer aids that are designed for the use of collaborative work groups." Again, this definition rightly focused on the actual group collaborative processes but does not mirror the current reality of how the tools extend to support the individual work processes that underpin group work. A major contribution by Johansen and others, however, was to provide a categorization scheme for groupware technologies that is still valuable today. As shown in Table 1, groupware technologies can largely be categorized along two primary dimensions:

- whether users of the groupware are working together at the same time ("real-time" or "synchronous" groupware) or different times ("asynchronous" groupware) and
- whether users are working together in the same place ("colocated" or "face-to-face") or in different places ("distributed" or "dispersed").

Another way to understand groupware is by the generic group processes that it supports: *communication* (pushing or pulling information), *collaboration* (shared information and building shared understanding), and *coordination* (delegation of task, group technique, sequential sign-offs, etc.). Some definitions of groupware call erroneously for the added dimension of a common goal. Although all systems require some agreement among participants (at minimum that they should agree to participate in the process), groupware can readily support negotiation-type interactions. Management of conflict is often a crucial feature of a groupware system (a voting tool is a good example of this incorporated in groupware designed to support meetings). This classification scheme also does not pay much attention to individual work processes that can be supported by groupware and are a component of group work, however, and it does not emphasize geographic location or temporality.

To build on these earlier perspectives and integrate them with the realities of individual work processes, as well as to reflect the ubiquity of access afforded by the Internet (and commonly, proprietary organizational networks and intranets), we offer a new model of groupware. Table 2 illustrates an emerging model for groupware as

Table 1 Traditional Dimensions of Groupware

	SAME TIME "SYNCHRONOUS"	DIFFERENT TIME "ASYNCHRONOUS"
Same Place: "Colocated"	Group support system Group decision support systems Presentation support	Shared work spaces
Different Place: "Dispersed"	Videophones Whiteboards Chat Virtual rooms	E-mail Threaded discussion Workflow

Table 2 A Categorization of Groupware Systems by Time and Targeted User Audience

	ASYNCHRONOUS GROUPWARE	ANYTIME/ANYPLACE GROUPWARE	SYNCHRONOUS GROUPWARE
Large	Database-oriented portals Collaborative portals E-community Newsgroup/usenet list Servers	E-learning/virtual classrooms	Web real-time communication Audio/video conferencing
Limited	Telework virtual workplaces Workflow systems Personal portals Group calendars	Distributed project management Collaborative writing Threaded discussion	Group decision support systems Meeting support Chat Shared whiteboard
Individual/Dyad	E-mail	Mobile/wireless messaging	Collaborative customer relationship management

represented by a 3 by 3 matrix of the nine discrete yet interrelated software groups. From a user's perspective, the horizontal axis represents a spectrum between synchronous and asynchronous functionality (in typical usage) but now includes "anytime/anyplace" to connote accessibility to information or functionality that the tool can provide. As we will see later, many modern forms of groupware, though primarily synchronous or asynchronous, have features that are accessible anytime and anyplace if one has an Internet or Intranet connection and a browser (Coleman, 1997). The vertical axis represents the group size targeted by the technology. In this representation, the "place" dimension of Table 1 above has disappeared—networks have so increased in speed and computers have become so ubiquitous and powerful, that geographical dispersion has lost its categorization power. For instance, any computer-supported meeting can be attended in face-to-face mode or by cyberspace connection.

GROUPWARE FUNCTIONALITIES IN THE INTERNET AGE

Common Groupware Functionalities

Although stand-alone groupware components such as e-mail can still be purchased separately, many modern products offer a number of common functionalities (e.g., e-mail) integrated into one suite. These common functionalities (or characteristics) might include the following.

Graphical User Interfaces

Most groupware programs today can take advantage of the color graphics and sound capabilities of multimedia computers. They are often made available via the Internet as either a client and server application or an application that uses a common browser interface.

WYSIWIS

This stands for What You See Is What I See. It is analogous to two people each at their own home watching the same television show at the same time. Computer technology extends this concept in groupware products and allows users to interact and communicate in a WYSIWIS environment.

Integration

Most groupware products integrate components of different applications into a standard interface so the product feels like a multifaceted tool.

Online Communication

Discussions can be held in real time, with active interaction between participants. Several products allow for textual and video exchanges in real time.

Replication and Synchronization

Replication is the process of duplicating and updating data in multiple computers on a network, some of which are permanently connected to the networks. Others, such as laptops, may only be connected at intermittent times. Synchronization is the goal of having all computers in a distributed network providing the same information at the same time. The replication and synchronization concept has two components: (a) everyone should have access to the same information in the database and be able to replicate the data locally and (b) several users can make changes to the same document, and somehow the system synchronizes all those changes into one master copy of the database. If the replication manager has multiple update requests to the same record, it sends out messages asking for clarification as to what the correct entry should be. An example of this function is the use of a shared database of product inventory by a distributed sales force. When each person sells some units, the database would update itself so that there is an accurate count of inventory available to all users.

Remote Access

Using database replication and synchronization, remote users can access and update information held in a groupware database (see example in previous section).

Workflow

Workflow is a typical groupware subsystem that manages a task through the steps required for its completion. It results in the automation of standard procedures (such as records management in personnel operations) by imposing a set of agreed on business rules on the procedure. Each task (or set of tasks), when finished, automatically initiates the next step in the process, until the entire procedure is completed. Although some users implement self-contained workflow systems, the technology often appears in more recent groupware and image management systems. Lotus Notes for Domino, for example, provides workflow applications that can be customized to an organization's particular needs.

Time Management

These groupware products offer group calendar diaries and group project schedules to help managers as well as subordinates better coordinate their time. Microsoft's Outlook, among other products, offers these and related functionalities.

Secretarial Functions

These are background processes that perform "clean up" or "overhead" services for users. For example, the secretary would manage the real-time messaging system to manage the receipt of messages by accepting high-priority interactions but blocking unnecessary interruptions and redirecting (and prioritizing) those less important messages to the users' electronic mail platform. These filtering functions are common in many email groupware products.

Project Management

Groupware's increasing adoption is due to its ability to support project management endeavors easily. As personal computers and networks have become more powerful, they have provided the necessary performance for today's groupware, imaging, and document management applications. These technologies, originally distinct, have become more integrated into many organizations with the aim of supporting project management. This type of system improves distributed project management through controlled management of documents, schedules, and personnel; allowing project personnel to be dispersed geographically without regard to location or time; providing communication among project personnel through a variety of means, including e-mail and discussion forums; and, because project personnel and project information are linked, project information remains current and accessible. The right information gets to the right people at the right time.

Cutting-Edge Groupware Functionalities

As we noted earlier, it is still useful to categorize groupware tools on the basis of whether they support synchronous or asynchronous work, although some modern groupware tools support processes that do not fit neatly into these two categories. For simplicity sake, we divide our discussion of modern Internet-age groupware into these two categories and comment on how the technologies fit into this classification scheme (see Table 2).

Asynchronous Groupware

Asynchronous groupware can take a variety of forms from the early and simple (but useful and enduring) such as e-mail to the most modern and complex such as Lotus Notes for Domino, described in the following paragraph. Although some forms such as e-mail were often freestanding organizational communication support systems, run on a firm's local area network, all of the forms we describe can and typically do run on the Internet (as well as on private portions of the Internet such as corporate intranets and extranets). Functional descriptions of some of these systems follow.

System profile: Lotus Notes for Domino—a database-oriented portal. Notes is a sophisticated asynchronous groupware system from the Lotus Corporation (http://www.lotus.com), a subsidiary of IBM. Notes provides a traditional group support functions that are tightly integrated to provide a virtual "workspace," library, and secretary for its users. Group members are able to communicate via e-mail, store all digitized information related to a project in a dedicated and secure database, and obtain scheduling and calendaring support aid in project management. These functions allow an organization's workers to interact so that users at different geographic locations can share information with each other, exchange comment publicly or privately, and keep track of project schedules, guidelines and procedures, plans, white papers, and so on. Notes keeps track of changes and makes updates to replications of all databases in use at any site.

Notes runs on a special server called the Lotus Domino Server. The servers and workstations use a client and server processing model, and replicated databases are updated using special programs called remote procedure call requests. Notes can be coordinated with Web servers and applications on a company's intranets.

In addition, Notes has strong workflow structuring capabilities and provides the basis for a document management system, well integrated with its database functions. Although it ships with these and a number of other customizable application "shells," Notes also provides simple facilities so that custom end-user applications can be developed. Given the product's inherent complexity, new users, as well as application developers, will face a steep learning curve.

Some aspects of Notes (e.g., the database functions) can be used by individuals as well as small and large groups. The database, for example, also would fit into our "anytime–anyplace" dimension by virtue of its Internet or intranet accessability. Other aspects such as e-mail fit more neatly into our "asynchronous" and our individual and dyad dimensions, although the latter could extend to large groups when used to support list serve functions.

Collaborative portals. Found on the Internet or intranets, collaborative portals (e.g., Blackboard 5, described in the next paragraph) are a single point of useful, comprehensive, ubiquitous, and integrated access. There are corporate, community (e-community), and personal portals that cater to specific communities of interest by providing tailored content. Groupware has recently been developed to automate the technical aspects of deploying a Web-browser-based portal such that end users are freed from the rigid limitations of traditional portal technology (Web page development) and enables them to conduct business by the terms of their own communities, interests, tasks, and job responsibilities. Portals increasingly are becoming collaborative tools: They can manage personal data, create spontaneous collaborative work spaces, allow real-time chat capabilities, and increase communication effectiveness by keeping track of who is online within the community (with pertinent information about them such as topics of interest and job function). These tools are also “anytime–anyplace” and can support small groups. They may be asynchronous, serving more as passive postings of information, or when enabled with the appropriate technology, support near-synchronous interaction (e.g., a chat room).

System profile: Blackboard 5. This community portal system features a highly customizable community portal environment that unifies academics, commerce, communities, and administrative services online through one integrated interface. This advanced functionality is backed by a sophisticated product architecture that runs on relational databases and can be scaled to support tens of thousands of users through a multiserver configuration.

Newsgroups/Usenet. Usenet is a collection of user-submitted notes or messages on various subjects that are posted to servers on a worldwide network. Each subject collection of posted notes is known as a newsgroup. There are thousands of newsgroups, and it is possible for a user to form a new one. Newsgroups are similar to e-mail systems in spirit but are intended for messages among large groups of people instead of one-on-one communication. In practice, the main difference between newsgroups and mailing lists is that newsgroups only show messages to a user when explicitly requested, whereas mailing lists deliver messages as they become available. Most browsers, such as those from Netscape and Microsoft, provide Usenet support and access to any newsgroups that users select. On the Web, Google and other sites provide a subject-oriented directory as well as a search approach to newsgroups and helps users register to participate in them. In addition, there are other newsgroup readers that run as separate programs. This technology is typically asynchronous with anytime–anyplace access via the network and supports large and small groups.

Listserv. A listserv is a small program that automatically redistributes e-mail to names on a mailing list. Users can subscribe to a mailing list by sending an e-mail note to a mailing list they learn about; listserv will automatically add the name and distribute future e-mail postings to every subscriber. (Requests to subscribe and unsubscribe are sent to a special address “the list servers” so that all subscribers do not see these requests.) This technology is typically asynchronous with anytime-anyplace access via the network and supports large and small groups.

Workflow systems. These systems allow digital versions of documents to be routed through organizations through a relatively fixed process. A simple example of a workflow application is an expense report in an organization: An employee enters an expense report and submits it; a copy is archived and then routed to the employee’s manager for approval. The manager receives the document, electronically approves it, and sends it on. The expense is registered to the group’s account and forwarded to the accounting department for payment. Workflow systems may provide features such as routing, development of forms, and support for differing roles and privileges.

Workflow systems can be described according to the type of process they are designed to deal with, and there are essentially three types of workflow systems:

1. **Image-based workflow systems** are designed to automate the flow of paper through an organization by transferring the paper to digital “images.” These were the first workflow systems that gained wide acceptance. These systems are closely associated with “imaging” technology and emphasize the routing and processing of digitized images. Lotus Notes, for example, supports both document management (via scanning or “imaging”) and workflow by directing the digitized material via user-defined routes.
2. **Form-based workflow systems** are designed to route forms intelligently throughout an organization. These forms, unlike images, are text-based and consist of editable fields. Forms are automatically routed according to the information entered on the form. In addition, these form-based systems can notify or remind people when action is due, providing a higher level of capability than image-based workflow systems. Notes can also provide customizable forms (e.g., invoices).
3. **Coordination-based workflow systems** are designed to facilitate the completion of work by providing a framework for coordination of action. The framework is aimed to address the domain of human concerns (business processes), rather than the optimization of information or material processes. Such systems have the potential to improve organizational productivity by addressing the issues necessary for customer satisfaction, rather than automating procedures that are not closely related to customer satisfaction. Customer relationship management systems, for example, can be set up to route customer problems to specialists.

Workflow systems can support individual and dyadic functions but are typically used to support larger groups. They are typically asynchronous and may permit anytime–anyplace access.

Group calendars. These systems allow scheduling, project management, and coordination among many people and may provide support for scheduling equipment as well. Typical features detect when schedules conflict or find meeting times that will work for everyone. Group calendars also help to locate people. Typical concerns are privacy (users may feel that certain activities are not public matters), completeness (users may neglect to enter their schedules into the system), and accuracy (users may feel that the time it takes to enter schedule information is not justified by the benefits of the calendar). Calendars are also typically asynchronous with anytime–anyplace access.

E-mail. By far the most common groupware application, basic e-mail technology is designed to pass simple messages between two people. Today, however, even relatively basic e-mail systems typically include interesting features for forwarding messages, filing messages, creating mailing groups, and attaching files with a message. Other features that have been explored include automatic sorting and processing of messages, automatic routing, and structured communication (messages requiring certain information). E-mail, although typically asynchronous, can appear to become "near synchronous" with a fast enough network.

Synchronous Groupware

Synchronous groupware systems (e.g., GroupSystems, described in the following paragraph) are interactive computer-based environments (physical and virtual) that support individuals (collocated or dispersed) working together at the same time. This emerging generation of groupware is also incorporating new technologies, such as Internet telephony, videoconferencing, speech recognition, and simplified yet secure connections with corporate databases. However, the key point of collaborative software remains connecting the "islands of knowledge" represented by each employee for the organization's greater good.

System profile: GroupSystems—a meeting support system. GroupSystems (http://www.groupsystems.com) offers a comprehensive suite of that can shorten the cycle time for business processes such as strategic planning, product development, and decision making. Three features that apply to all tools in the GroupSystems suite:

1. Simultaneous contribution, in which everyone is "speaking" at once, saving time and increasing productivity.
2. Anonymity, in which the identity of each contributor is unknown so participants tend to feel freer to express their opinions, and ideas are evaluated more objectively.
3. Complete records, in which, at the end of a meeting, users can easily produce a complete and accurate report of all ideas, comments, and vote results, in Word or rich text format (.rtf).

There are seven tools in the GroupSystems suite. Each focuses on a specific aspect of a group effort, including idea generation, evaluation, organization, and exploration. For example, the electronic brainstorming module, a tool that supports simultaneous and anonymous idea sharing, is commonly used for team building, broad or focused brainstorming, and visioning or strategic planning sessions. The group voting module supports eight voting methods, including a customizable point scale, and makes the decision process efficient, flexible, and powerful. Organizations use the voting module to evaluate, make decisions, and build consensus.

Early forms of the GroupSystems tool set were designed to run on local area networks, often in dedicated meeting rooms with networked personal computers, electronic whiteboards, and audiovisual support. More recent forms are now Web-enabled and provide similar functionality to distributed meetings. In addition, these systems are among the oldest and best researched types of groupware (e.g., Nunamaker, Briggs, Mittleman, Vogel, & Balthazard, 1997).

Typical synchronous groupware include the following.

Chat systems. These permit many people to write messages in real time in a public space. As each person submits a message, it appears at the bottom of a scrolling screen. Chat groups are usually formed by listing in a directory chat rooms by name, location, number of people, topic of discussion, and so on. Chat systems, like e-mail, may be accessed anytime–anyplace when hosted on an Internet server or may be available only at certain times or through certain connections if run on a proprietary intranet.

Shared whiteboards. An electronic whiteboard is one of several writeable presentation display systems that can be used in a classroom or videoconference. These generally fall into one of three categories: stand-alone copy boards, which allow users to scan and print out the content of the whiteboard; peripheral boards, which transfer information in the form of digital files to an attached computer; and interactive boards, which are like large touch-screen monitors that can be synchronized to an attached computer, allowing users to interact with the display, visit Web sites, and access databases directly from the board. Some peripheral boards can accommodate a projector that can be calibrated to the display, making them interactive. There are a number of add-on whiteboard digitizer products available that can also be used to make traditional dry-erase whiteboards interactive.

In cyberspace, a whiteboard is a space on the display screen in which one or more participants write or draw, using a mouse, keyboard, or other input device. It allows users to view and draw on a shared drawing surface even from different locations. This can be used, for instance, during a phone call, allowing each person to jot down notes (e.g., a name, phone number, or map) or to work collaboratively on a visual problem. Most shared whiteboards are designed for informal conversation, but they may also serve structured communications or more sophisticated drawing tasks, such as collaborative graphic design, publishing, or engineering applications. Microsoft's NetMeeting allows collaborative real-time use of its PowerPoint application via the Web. Users can pass control of the cursor to each other while sharing a single view of a common screen, thus collaborating on the development of a slide or graphic. Although whiteboards might be real physical devices or virtual representations, the information they capture can be accessed anytime–anyplace when they are connected to a network.

Collaborative customer relationship management (CRM). Collaborative CRM systems support the communication and coordination across the customer life cycle between channels and customer touch points. They are aimed at providing an enhanced understanding and management of the customer relationship—increasing knowledge of customer behavior, building switching costs, and increasing customer satisfaction and retention, especially relationship-centric, process-centric, and productivity-centric collaboration and a more mature collaborative CRM market. Collaborative CRM addresses the problems of traditional CRM solutions that are inwardly focused on an enterprise, not on the customer. By understanding the heterogeneous environment of most enterprise application portfolios, collaborative CRM provides a framework that sales, service, marketing, and product development organizations can work together but still maintain their unique way of doing business. Some CRM functions are similar to workflow and are asynchronous, but others support the synchronous interaction that organizational teams use to provide customer support.

Videoconference. A videoconference is a live connection between people in separate locations for the purpose of communication, usually involving audio and often text as well as video. At its simplest, videoconferencing provides transmission of static images and text between two locations. At its most sophisticated, it provides transmission of full-motion video images and high-quality audio between multiple locations.

Videoconferencing software is quickly becoming standard computer equipment. For example, Microsoft's NetMeeting is included in Windows 2000 and is also available for free download from the NetMeeting homepage. For personal use, free or inexpensive videoconference software and a digital camera afford the user easy and inexpensive live connections to distant friends and family. Although the audio and video quality of such a minimal setup is not high, the combined benefits of a video link and long-distance savings may be persuasive.

The tangible benefits for businesses using videoconferencing include lower travel costs and profits gained from offering videoconferencing as an aspect of customer service. The intangible benefits include the facilitation of group work among geographically distant teammates and a stronger sense of community among business contacts, both within and between companies. In terms of group work, users can chat, transfer files, share programs, send and receive graphic data, and operate computers from remote locations. On a more personal level, the face-to-face connection adds nonverbal communication to the exchange and allows participants to develop a stronger sense of familiarity with individuals they may never actually meet in person.

A videoconference can be thought of as a phone call with pictures—Microsoft refers to this aspect of its NetMeeting package as a "Web phone"—and indications suggest that videoconferencing will some day become the primary mode of distance communication. Web-enabled, videoconferencing is accessible anytime–anyplace, although bandwidth considerations can limit the quality of transmission and reception.

Audio–video communications. Real-time communication (RTC) systems allow two-way or multiway calling with live video—essentially a telephone system with an additional visual component. Cost and compatibility issues limited early use of video systems to scheduled videoconference meeting rooms. Video is advantageous when visual information is discussed but may not provide a substantial benefit in most cases in which conventional audio telephones are adequate. In addition to supporting conversations, video may also be used in less direct collaborative situations, such as by providing a view of activities at a remote location.

Many RTC systems allow for rooms with controlled access or with moderators to lead the discussions, but most of the topics of interest to researchers involve issues related to unmoderated RTC including anonymity, following the stream of conversation, scalability with number of users, and abusive users. Although chatlike systems are possible using nontext media, the text version of chat has the rather interesting aspect of having a direct transcript of the conversation, which not only has long-term value but also allows for backward reference during conversation making it easier for people to drop into a conversation midway through it and still pick up on the ongoing discussion.

Threaded-discussion systems. Threaded-discussions systems are a place where messages can be posted for public display. The groupware tool allows participants to communicate with each other asynchronously. Communication takes place by having participants create a continuous chain of posted messages on a single topic. This chain is known as a "thread" or a "threaded discussion." Several parallel threads can make up a full conversation. The threads can then be used to build a hierarchy with the content. In some instances, a threaded discussion may also be called a bulletin board or a forum. A discussion can be used as a place where participants can discuss general issues, discuss the specifics of an issue, or organize related content. With a threaded-discussion tool, users can read posted messages, compose new messages, attach documents to posted messages, search and compile posted messages. Like e-mail and its variants, these systems are anytime–anyplace accessible with the right connection and hardware.

Decision support systems. These systems are designed to facilitate groups in decision-making situations. They provide tools for brainstorming, critiquing ideas, putting weights and probabilities on events and alternatives, and voting. Such systems enable presumably more rational and even-handed decisions. Primarily designed to facilitate meetings, they encourage equal participation by, for instance, providing anonymity or enforcing process structure (like a nominal group technique or a Delphi technique). Although in the past some forms were primarily designed for same-place–same-time interaction, many individual tools can be installed on a network for anytime–anyplace access. Some integrated suites such as GroupSystems (profiled earlier) are now available in Web-enabled formats.

Virtual workplace. The virtual workplace concept seeks to create an electronic facsimile of an office workplace in which coworkers meet and work. This is typically

Table 3 The Groupware Grid

	COMMUNICATIONS SUPPORT	DELIBERATION SUPPORT	INFORMATION ACCESS SUPPORT
Concerted Work Level			
Coordinated Work Level	░	░	░
Individual Work Level	▓	▓	▓

manifested as a Web site that has pictures of team members, links to instant messaging, a link to asynchronous discussion boards, as well as access to all the documents and database links needed. The goal is to enhance also the sense of being on a team by giving dispersed team members a "place" to go, where they can all "meet" (i.e., communicate textually and possibly telephonically or visually in real time). Lotus Notes for Domino features a number of functionalities that support this goal.

Collaborative writing systems. These system may provide both real-time and offline support. Word processors may provide asynchronous support by showing authorship and by allowing users to track changes and make annotations to documents. Authors collaborating on a document may also be given tools to help plan and coordinate the authoring process, such as methods for locking parts of the document or linking separately authored documents. Synchronous support allows authors to see each other's changes as they make them and usually needs to provide an additional communication channel to the authors as they work (via videophones or chat). Microsoft's NetMeeting provides this synchronous functionality with its Word application and similar functionality with its spreadsheet (Excel), database (Access), and graphics programs (PowerPoint).

Mobile and wireless messaging. Messaging is the ability to see whether a chosen friend or coworker is connected to a network and, if they are, to exchange relatively short messages with them. Messaging differs from e-mail in the immediacy of the message exchange and also makes a continued exchange simpler than sending e-mail back and forth. Most exchanges are text only, although some services now allow voice and attachments.

For messaging to work, the sender and receiver must both subscribe to a network service and be online at the same time, and the intended recipient must be willing to accept instant messages. Under most conditions, messaging is truly "instant." Even during peak Internet usage periods, the delay is rarely more than a second or two. It is possible for two people to have a real-time online "conversation" by messaging each other back and forth.

Mapping Groupware Functionality to Work Processes

The broad (and expanding) range of modern groupware functionality means that those who adopt the technology would benefit from some guidance. To this end, the Groupware Grid (Table 3) can serve as a theory-based heuristic model for mapping the functionalities of groupware technology to organizational work. The Groupware Grid is useful for managers to understand how to understand the dimensions of individual and group work that can be supported by the different general functionalities. It is consistent with our technology-oriented classification scheme presented earlier in the chapter but places its emphasis on the type of work that would be supported.

Team Theory and the Groupware Grid

The horizontal axis of the Groupware Grid derives from the team theory of group productivity (adapted from Briggs, 1994). *Webster's Dictionary* defines a team as a collection of people working together for some specific purpose. Team theory is a causal model for the productivity of a team. It asserts that team members divide their limited attention resources among three cognitive processes: communication, deliberation, and information access. Team theory posits that these processes interfere with one another, limiting group productivity.

Team theory's communication construct posits that people devote attention to choosing words, behaviors, images, and artifacts, and presenting them through a medium to other team members. Team theory's deliberation construct asserts that people devote cognitive effort to forming intentions toward accomplishing the goal-solving activities: make sense of the problem, develop and evaluate alternatives, select and plan a course of action, monitor results, and so on. The information-access construct addresses the attention demands of finding, storing, processing, and retrieving the information the group members need to support their deliberation. Team theory posits that a key function of information is to increase the likelihood that the outcome one expects will be obtained by choosing one course of action over another. Information has value to the extent that it is available when a choice must be made, to the extent that it is accurate and complete. The value of information is offset by the cost of acquiring, storing, processing, and retrieving it, however.

Team theory also posits that the cognitive effort required for communication, deliberation, and information access is motivated by goal congruence—the degree to which the vested interests of individual team members are compatible with the group goal. Team members whose interests are aligned with those of the group will exert more effort to achieve the goal than those whose interests are not served by the group goal. The Groupware Grid does not address goal congruence because goal

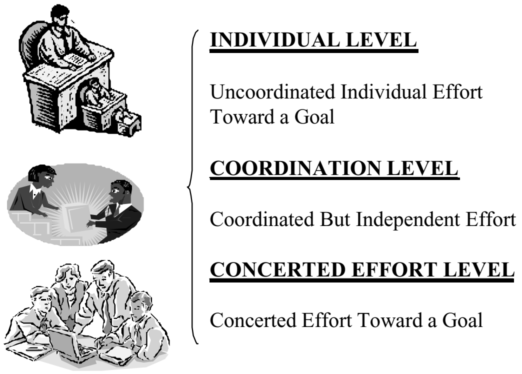

INDIVIDUAL LEVEL
Uncoordinated Individual Effort Toward a Goal
COORDINATION LEVEL
Coordinated But Independent Effort
CONCERTED EFFORT LEVEL
Concerted Effort Toward a Goal

Figure 2: Three levels of group work.

congruence may have more to do with the way a team wields the technology than with the technology's inherent nature. Therefore, the horizontal axis of the grid addresses the potential for technology to reduce the cognitive costs of joint effort. Groups may become less productive if the attention demands for communication, deliberation, or information access become too high. Groupware may improve productivity to the degree that it reduces the attention costs of these three processes.

Three Levels of Group Work and the Groupware Grid

The vertical axis of the Groupware Grid consists of three levels of group effort (Figure 2). Sometimes a team may operate at the individual work level, with members making individual efforts that require no coordination. Other times team members may interact at the coordinated work level. At this level, as with a technical sales team, the work requires careful coordination between otherwise independent individual efforts. Sometimes a team may operate at the concerted work level, that is, several parts of an organization must support the sales team to ensure future results. Teams working at this level must make a continuous concerted effort beyond short-lived coordination. For example, the sales department of an industrial manufacturer must routinely interact with its engineering department to help determine customer specifications on a new product. Today, there are even interorganizational cross-functional teams that provide concerted efforts to support increasingly complex value chains for products and services.

The demands placed on the team vary depending on the level of work in which they are engaged. There is groupware technology to support teams working at all three levels. The Groupware Grid can thus be used to map the contributions of a single groupware tool or an entire groupware environment (see Table 4). A given technology will probably provide support in more than one cell. The potential for productivity of different environments can be compared by examining their respective grids. For example, the shared database aspect of a product such as Lotus Notes is often used asynchronously and offers little support at the concerted work level but offers strong support for communication and information access at the coordination level. Furthermore, a team database offers only indirect support for deliberations at the coordination level, but a project management and workflow automation system offers strong support for deliberations at that level.

Synchronous groupware that supports meetings, for instance, offers a great deal of support for communication, deliberation, and information access at the concerted work level. For example, the parallel input and anonymous interventions can improve communication during a concerted effort. Other groupware tools affect deliberation in different ways. For instance, a brainstorming tool prevents a group from thinking deeply while encouraging them to diverge from familiar thinking patterns. An idea organizer, on the other hand, encourages a divergent group to focus quickly on a narrow set of key issues. With ubiquitous access to the Internet and client and server access to corporate databases, emerging synchronous groupware tools can support information access at the concerted work level by providing rapid access to the information in the minds of teammates by providing permanent transcripts of past electronic interactions, or by providing an information search engine.

Table 4 Potential Contributions of Groupware to Support Productivity Processes, as Viewed in the Groupware Grid

	COMMUNICATIONS SUPPORT	DELIBERATION SUPPORT	INFORMATION ACCESS SUPPORT
Concerted Work Level	Anonymity Parallel contributions Virtual presence Synchronous communication	Structured processes Distributed participation Alternative analysis Process focus	Database access Online search engines Organizational memory Concept classification
Coordinated Work Level	Asynchronous communication Sequential contributions	Distributed project management Scheduling Automated workflow	Shared data Coordinated filtering
Individual Work Level	Message preparation Communication structure	Modeling support Simulation support Hypothesis testing Logic analysis	Individual data Filtering Categorizing

GROUPWARE MANAGEMENT ISSUES: PRODUCTIVITY AND ORGANIZATIONAL IMPACT

Perhaps the most important management issues surrounding Web-based collaboration tools are how to use them to enhance productivity and how to manage their organizational impacts. Not all organizations that adopt groupware reap the potential benefits, and this is often due to the same factors that can plague the adoption of any information technology. Companies may try to implement a product that is not designed for its intended use, they may not think through exactly what benefits they wish to accrue from the technology, they may not invest sufficient efforts in training and support, and they may not understand that some of processes are amenable to team and group work (technologically supported or not) but others are not. Adoption of Web-based collaboration support systems is often tantamount to a large and often ill-defined management initiative such as total quality management (TQM) or business process reengineering—that is, success or failure of the adoption is often difficult to measure, benefits will emerge at different and often unpredictable rates, and human issues such as culture, politics, power, incentive and personality either facilitate success or bring the project to its knees. We cannot pursue systems planning and development here but instead devote some discussion first to groupware issues that derive from individual users and their interaction with groupware-supported team members and second to issues that derive from an organization's larger culture.

Groupware and Productivity

Organizations may not realize that some group or team members may not have the personality, motivation, or teamwork and communication skills to be effective in their roles. These issues are now beginning to receive attention from researchers examining the performance of groupware-supported workers and teams (e.g., Potter & Balthazard, 2002). Findings suggest that first, people collaborating via groupware do not leave their personalities at the virtual door simply because their new communication mode is textual and possibly asynchronous (e.g., Hiltz & Turoff, 1993). People's personalities drive how they interact with one another in both the face-to-face world and the virtual world; in both modes, people are similarly affected by each other's individual style of interacting. These effects can determine whether a group enjoys enhanced or diminished information sharing, enhanced, or diminished task performance (e.g., decision making) and positive or negative sentiments such as buy-in to the team's solution or product, satisfaction with the team's processes, and the team's cohesion. The important implication is that groupware means group work, and not all people are equally skilled in group communication and work processes or otherwise predisposed to work best in a team situation. Interventions such as training are needed to help some groupware-supported people adopt more appropriate interaction skills before they tackle a task. Alternatively, future groupware could incorporate interfaces, communication aids, protocols, and intelligent agents designed to promote, for example, participation, constructive interaction, and team building.

Organizational Impacts of Groupware

For several years, researchers have compiled evidence that information technology such as groupware or other forms of computer-mediated communication can alter a number of relationships and processes in the workplace (e.g., DeSanctis & Fulk, 2002). An organizational hierarchy that has traditionally exhibited mostly top-down dispersion of information may find that the communication support capacities of a collaborative information system create new lateral and bottom-up flows of information. In more radical manifestations, command-and-control structures in the organization can be circumvented as employees reinvent information-intensive processes based on expertise, interest, and resources, rather than restrictive operating procedures, protocols, or customs. Some researchers suggest that, once in place, collaborative systems become the vehicle for organizational redesign, and as such their more profound effects are largely emergent. In many cases, not all purposes or uses of the systems can be identified beforehand. Therefore, their management should reflect more of a "wait and see" attitude and an acceptance of the fact that they may give rise to a number of unanticipated benefits (e.g., Orlikowski, 1992).

Another common finding is the spontaneous creation of online communities of interest or professional practice. These can be the gateways into larger organizational strategic concerns such as knowledge management. Effectively managed collaborative technologies can and often do provide the venue for knowledge sharing—an important input into the knowledge management process, particularly when derived from communities of professional practice or interest (Constant, Sproull, & Kiesler, 1999; Pickering & King, 1999).

The national cultures of groupware users can also become a management issue for transnational organizations. Different traditions, social mores, and internal and national political realities can drive preferences for certain technologies and interaction modes (i.e., text, video, threaded discussion, e-mail; Potter & Balthazard, 1999). In addition, preference of various groupware technologies can also differ by gender (Potter, Balthazard, & Elder, 2000).

This more expansive "sociological" view of the management of these systems is informed by knowing the politics and culture of the organization on one level and the personalities of system users on another level. These—as much as the technology's functionality—ultimately will drive the use and potential benefits of the system. As noted earlier, some suspension of "standard operating procedure" or organizational status quo is a common result of these systems, and leaders often have to deal with relatively rapid changes in how things are done and, perhaps more important, how procedures and technologies are perceived in their organization. Many aspects of an organization's culture are "genies in a bottle" that are released when a new communication and collaboration system allows transcendence of predetermined, traditional communication paths. The good news is that installing the collaborative communication technology found in many groupware products into healthy organizations will likely spur new organizational knowledge, improved practices, and synergistic relationships to emerge.

CONCLUSION

Modern Internet-based groupware offers a plethora of useful functionality that is driving sizeable gains in organizational productivity by streamlining collaborative work that was traditionally encumbered by distance and time. It is also driving the reinvention of business processes, spontaneously altering the nature and connection of communication-driven relationships and enabling an unprecedented growth in organizational and professional knowledge. At the same time, such impact can represent important management challenges that need to receive increased attention from managers and researchers alike as the proliferation of Internet-based groupware continues. Thoughtful assessment of groupware functionalities matched with a true understanding of the human side as well as the work process side of the organizational context will ensure the creative and effective adoption of these revolutionary technologies in the foreseeable future.

GLOSSARY

Anytime–anywhere Groupware components and resources such as databases and threaded discussions that can be accessed via the Internet from any location at any time.

Asynchronous groupware Groupware that does not allow users to interact in real time.

Groupware Technology-typically software and networks—designed to support people working in groups.

Groupware grid A theory-based heuristic model for mapping the functionalities of groupware technology to organizational work.

Organizational impact Effects on work processes, social and authority hierarchies, and communication pathways and processes that can occur with groupware adoption.

Synchronous groupware Groupware designed to allow two or more users to interact in real time.

Team theory A causal model for the productivity of a team developed by Briggs (1994), asserting that team members divide their limited attention resources among three cognitive processes: communication, deliberation, and information access.

CROSS REFERENCES

See *Customer Relationship Management on the Web; E-mail and Instant Messaging; Intranets; Online Communities; Virtual Teams.*

REFERENCES

Briggs, R. O. (1994). *The focus theory of group productivity and its application to the design, development, and testing of electronic group support technology*. Unpublished doctoral dissertation, Management Information Systems Department, University of Arizona, Tucson.

Coleman, D. (1997). *Groupware: Collaborative strategies for corporate LANs and intranets*. Upper Saddle River, NJ: Prentice-Hall PTR.

Collaborative Strategies. (2002). Real-Time communication and collaboration 2002 industry report. Retrieved February 21, 2002, from http://gatedway.com/cs_shop/csi_shop_rtc_Rev01.php

Constant, D., Sproull, L., & Kiesler, S. (1999). The kindness of strangers: The usefulness of electronic weak ties for technical advice. In G. DeSanctis & J. Fulk (Eds.), *Shaping organization form: Communication, connection, and community* (pp. 415–444). Thousand Oaks, CA: Sage.

DeSanctis, G., & Fulk, J. (Eds.). (1999). *Shaping organization form: Communication, connection, and community*. Thousand Oaks, CA: Sage.

Hiltz, S. R., & Turoff, M. (1993). *The network nation: Human communication via computer*. Cambridge, MA: MIT Press.

Johansen, R. (1988). *Groupware*. New York: Free Press.

Nunamaker, J. F., Briggs, R., Mittleman, D. D., Vogel, D. R., & Balthazard, P. (1997). Lessons from a dozen years of group support systems research: A discussion of lab and field findings. *Journal of Management Information Systems, 13,* 163–207.

Orlikowski, W. (1992). Learning from NOTES: Organizational issues in groupware implementation. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work* (pp. 362–369). Toronto: ACM Press.

Pickering, J. M., & King, J. L. (1999). Hardwiring weak ties: Interorganizational computer-mediated communication, occupational communities, and organizational change. In G. DeSanctis & J. Fulk (Eds.), *Shaping organization form: Communication, connection, and community* (pp. 399–413). Thousand Oaks, CA: Sage.

Potter, R. E., & Balthazard, P. A. (1999). Supporting integrative negotiation via computer-mediated communication technologies: An empirical example with geographically dispersed Chinese and American negotiators. *Journal of International Consumer Marketing, 12,* 7–32.

Potter, R. E., & Balthazard, P. A. (2002). Understanding human interaction and performance in the virtual team. *The Journal of Information Technology Theory & Application*. Retrieved February 21, 2002, from http://www.peffers.net/journal/volume4_1/toc4_1.pdf

Potter, R. E., Balthazard, P. A, & Elder, K. L. (2000). Toward inclusive dialogue in the classroom: Participation and interaction in face-to-face and computer-mediated discussions. *Journal of Information Systems Education, 11,* 73–82.

Guidelines for a Comprehensive Security System

Margarita Maria Lenk, *Colorado State University*

INTRODUCTION

Security for Internet-related distributed networks is not a product, service, or status produced or purchased at one point in time, nor is it a one-time decision that can be expected to be valid over a period of time. Modern distributed networks are continuously changing. Each day there are new technologies, new risks, and new threats that challenge the status quo and must be considered. Security, therefore, is a complex continuous process that actively involves stakeholders and participants who have differing priorities for and valuations of the assets involved. A careful, systematic approach to the task of network security can provide the reliable assurance over time needed by organizations who utilize the Internet. This chapter models such a process for an organization's security team to utilize in designing, implementing, and maintaining comprehensive Internet-related system security. A review of the assets involved, their risk sources and probable effects, and a spectrum of prescriptive measures is provided. The inherent bias in this model, supported by many real-world security successes and failures, is the belief that a lack of management support and involvement in a systematic approach toward comprehensive security has been associated with many Internet-related network security problems. Many of the primary causes of security failures (e.g., lack of enough qualified IT staff, insufficient/ineffective funding for IT needs, communication barriers between the IT staff and top management) may have been mitigated or eliminated with active top management support and involvement (Stein, 1999; SANS, 2002; Allen, 2001).

A perfect comprehensive security system would protect all assets from all possible damaging events, whether intentional or not in nature. However, restricted resources, technological realities and advances, and human characteristics more often result in less-than-perfect comprehensive security systems. For-profit, nonprofit, and government agencies all operate with a limit on the utilization of capital for electronic security. Also, top management's attitude toward funding security can be less than enthusiastic for a variety of reasons, which are discussed in this chapter. The discovery and application of new technologies or new uses of existing technologies also change security risk sources and assessments. Finally, any system that involves human beings is subject to possibilities of unintentional errors due to fatigue, illness, lack of training and supervision, and malicious actions due to attention, resentment, revenge, and control needs (Garfinkel & Stafford, 2001; IOMA, 2000).

Even when powerfully effective individual and joint security elements are present in a network, the entire network security will only be as strong as its weakest link. In other words, it is the unlocked door that invites entry. A burglar is at best only momentarily delayed if he first finds a locked door with state-of-the art security before finding the unlocked door. Furthermore, Internet-connected networks and their environments are continuously changing and security teams need to continuously screen for new assets, new risks, and new solutions. The task for security teams, then, involves not only designing an efficient and effective security system, but also continually assessing and testing the system in efforts to find and correct its weakest links. This chapter will provide a structured process for this seemingly formidable task. Anderson (2001) states that "building systems in the face of malice is one of the most important, interesting, and difficult tasks facing engineers in the twenty-first century."

BUILDING YOUR SECURITY TEAM

Comprehensive security system management involves the creation and delegation of authority to a knowledgeable, trusted security team. The membership of that team should depend upon the size of the organization, the dependence of the organization and its business partners on their Internet-related networks for critical business processes (suppliers, customers, creditors, banks, etc.), and the level of risks associated with the organization's industry and practices (regulatory agencies, legal liabilities, etc.). All policies and resolutions emerging from this team should be formalized in writing, signed by each of the team members, and be guidance for each of the systems' administrators for implementation and maintenance purposes.

Suggested membership for the comprehensive security team would include top management, a finance officer, an internal audit officer, system administrators, major user group representatives, and major business partner representatives. Business partners may be suppliers, contract laborers, outsourced process providers, bankers, creditors, and customers. While many organizations prevent outsourced "employees" from becoming too familiar with their systems (e.g., need-to-know controls), new types of business affiliations and Web services are creating fuzzier lines between the organization and its environment. The different stakeholders, whether in the organization or in the nexus of surrounding business partners, have varying priorities for and valuations of the assets involved in the distributed network systems. Their needs and viewpoints need to be included, either as members of the security team or accessible by the security team, for the security system to be effective.

The inclusion of top management representatives and a top financial officer on these security teams is important for several reasons. First, top management's attention typically first goes to issues of market, customer, and product or service quality and fulfillment. The provision of Internet-related network security does not directly bring in any new markets, new customers, or more revenue from existing customers. However, just as security failures can cause business failures or significant negative impacts on corporate image, the indirect effects of effective visible security measures and the perceptions of trust therefrom can have a sizable positive effect on customers' choices.

Another reason for top management representation on the comprehensive security team is due to the importance of the influence of top management's participation on the culture, formal and informal, of the organization. Management variables related to security risk for an organization include the philosophies, operating styles, integrity, and competences of top management, the delegation of authorities and responsibilities in the organization (the structure), the quality of the board of director's audit committee, and the policies and procedures for human resources (Greenstein and Vasarhelyi, 2002).

The inclusion of a top finance officer in the security committee membership is important for two primary reasons. First, the cash outflows required for security are oftentimes poorly estimated or understood in timing, justification, and/or amount by nonfinance business people. Cash outflows related to system security goals are not clearly matched to correlated or resulting cash inflows. Second, "benchmark" comparisons to competitor's security-related cash outflows or security incidents may not performed, and without an IT security-knowledgeable finance officer, finance committees often reach to the IT budget to cut overhead costs. Involvement of a top finance officer can assist in the sustained success of investment in corporate security in all of these situations. The inclusion of an internal auditor on the security committee is primarily for the consideration of making sure that all security measures taken can be verified as to their effectiveness in some manner (otherwise, it may be impossible to deduct the benefit of the investment), as well as enabling the auditor to know the level of internal control inherent to the electronic processes performed on the network. Collaborations with the internal auditors in this regard can result in cost efficiencies in the internal auditing department, as well as providing documentation that may result in cost efficiencies for the external independent audit process.

Finally, for comprehensive security policies, procedures, tools, and equipment to be taken seriously by the rest of the employees and other stakeholders, the support and involvement of top management is essential. Participation in the security team is an efficient communication that management is willing to "walk the talk" of the importance of security for the organization.

THE ASSET-SECURITY CONTINUUM

A centuries-old personal reflection model from India presents a continuum model that relates a person's beliefs system to their destiny. The continuum illustrates the relationships between beliefs and thoughts, thoughts and choice of actions, actions and habits, and habits and ultimate destiny, or end state. The "lesson" within this saying is not about trying to understand how your beliefs are related to your destiny, but rather comes from the research and reflection performed at any one or more of the individual dyads within the continuum: beliefs-to-thoughts, thought-to-actions, actions-to-habits, and habits-to-destiny. The greater the correlation built into these dyads, the greater the probability that you will reach your desired destiny. This methodology of breaking down a large task into smaller localized tasks is similar to the milestone execution focus in project management and system development guidelines, an approach also recommended here in this chapter for the development and maintenance of a comprehensive security system.

This continuum conceptualization can be applied to comprehensive security systems. As can be seen in Figure 1, the asset-security continuum contains the individual elements of a business' assets, risks, threats, controls, and security goals. This continuum begins with the identification and prioritization of the business assets, whether tangible or intangible, physical or logical. Then the abstract risks that can cause loss, costs, or damage to an organization's assets are considered. Next, the specific threats, or causes of those risks, are identified. This is followed by the consideration of general and specific controls that may reduce or eliminate those threats. Finally, the specific effectiveness, efficiency, or trust-creating security goals of the organization are presented. Comprehensive security is maximized when careful research and reflection occurs for each of the dyads in the asset-security continuum. The power of this model lies in the correlation within the individual linkages rather than in the endpoints.

Assets are tangible, digital, and intangible aspects of a business which provide future value. Assets can be servers, payroll systems, customer lists, or even the knowledge of the database administrator. Assets need to be identified and then dynamically categorized in terms of their priority with respect to the corporate mission and strategies. The prioritization of assets becomes complicated when the timing of the risks and threats are taken

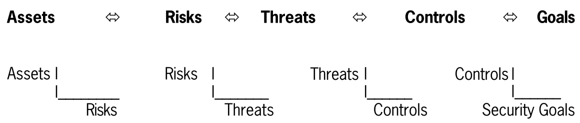


Figure 1: The asset-security continuum.

into consideration. For example, if the payroll application were to crash immediately after the electronic fund transfer instructions were sent to the bank, the corporation would have the time until the next pay-day to recover the payroll application. However, if the payroll application were to crash one hour before the normal fund transfer to employee accounts, the payroll system would have a much higher priority for the security system.

Risks are defined as events that cause an organization to lose value. Risks can create a very wide range of negative effects, including business interruption, legal or regulatory sanctions, excessive costs, lack of independence, losing revenues, losing customer loyalty, and losing intellectual property. Risks can be the reduced capacity to produce new intellectual property or reduced operating flexibility in the future.

Threats are the specific causes of risk. Threats can emerge from external or internal sources. Examples of threats from the external environment include a lack of or new regulatory or legal standards, new competitors, malicious or curious hackers, new technologies, misuse or abuse of your data by application service providers, earthquakes, floods, or war-related attacks. Examples of threats from the internal environment include malicious employees, software or hardware failures, unintentional human error, lack of buy-in toward the importance of security-related team-work, outdated systems, undocumented systems, and even unaudited systems.

Controls are policies, procedures, techniques, tools, habits, and leadership structures and attitudes designed to reduce threats to an organization. Alternatively, controls are designed to maximize the probabilities regarding the achievement of organizational goals and responsibilities of security, the efficiency and effectiveness of operations, and legal and regulatory compliance. Controls can be as pervasive as top management's attitude or style and the standard hiring practices utilized by an organization, or as specific as a reasonableness check on one field of transactions from one day in a database for a specific application. Controls can be classified as preventative if they restrict the event from occurring, detective if they sound an alarm when the event is discovered, and corrective if they help restore the situation to prior to the security-reducing event. Passwords are preventative controls because they are designed to keep unauthorized access from occurring. Logouts are detective controls that shut down log-in opportunities when a specific number of failed log-in attempts occur. Backups are corrective controls that restore the system to the state at the last time that the backups were made. Controls can be required or discretionary, depending upon the specifics of the situation or the judgment of the individual. Controls can be evaluated for effectiveness, which regards how the control helps to increase the probability that a particular security goal is achieved. Also controls can be categorized for efficiency, in terms of the number of threats mitigated by a particular control. Conversely, control redundancy (more than one control for a particular asset, process, or store) may be desired for highly vulnerable or valuable items for a business.

Security goals are the specific statements made by an organization to explain and communicate their security-related objectives. For example, organizations dependent upon their Web commerce may want their Web servers functional 24 hours per day, 7 days per week. Companies with proprietary Web services may want to keep their code or methods confidential. Companies negotiating online with their suppliers or their customers may want to keep the negotiations secret so that none of the other competitors learn of their decisions. The security objectives should be stated in a manner that encourages assessment with specific confidence intervals and system reliability goals clearly articulated.

Two examples of the utilization of the asset-security continuum are provided to illustrate its usefulness for comprehensive security teams. Consider purchasing negotiations with a primary supplier as the asset. Risks related to this asset include rising costs, for example, due to damages to the relationship with this supplier. Threats that increase this risk include errors in the online purchasing system, such as duplicate, erroneous, or nonauthorized orders. Control procedures designed to mitigate these threats include automated field verification procedures, encryption and digital signatures for all electronic communications with this supplier in order to achieve a security goal confidence that all communications have been confidential and authenticated. A second example involves a business's customer database as the asset. A risk to this asset includes reduced customer service. While there are several threats that could cause this risk, one of which could be a lack of integrity in the data regarding the customers (which could be due to hackers modifying the database, careless employees inputting errors, ineffective database design that does not allow important queries to be performed, etc.). Control plans that reduce these threats range from access controls (e.g., passwords and/or encryption) to application design and modification procedures and controls. The security goal for this example could be to have real-time accurate information on all customers. The next sections of this chapter review each of the individual components of the asset-security continuum: assets, risks, threats, controls, and security goals.

DYAD TABLES

One of the more powerful ways to utilize the asset-security continuum is to break down the whole model into four two-element dyads: assets-to-risks, risks-to-threats, threats-to-controls, and controls-to-security goals. Figure 1 illustrates these dyads below each of the linkages in the model. For example, the left-most dyad plots the organization's priority assets as the rows of the table and the abstract risks to the organization as the columns. The cells of the table can be discussed by the security team, and then filled if a particular risk column affects the asset listed in that row. Similarly, the next dyad plots the abstract risks with the specific threats, or causes of loss for the organization. In this dyad table, a particular risk may be caused by a number of threats; likewise, a particular threat may eventuate a business risk. Just as in the centuries-old Indian saying where the belief system assists the ultimate destiny through the personal work done on each of the individual dyads, the correlation between the assets and the comprehensive security system is maximized when the comprehensive security team studies each of the four dyads within the asset-security continuum.

IDENTIFYING AND CLASSIFYING YOUR ASSETS

Organizations derive their current and future value from their assets. Their assets may be physical or digital, such as property, equipment, data or software, or intangible such as intellectual capital or having a reputation of providing the best service. Organizational security teams must devise a system to first identify all value-providing assets of an organization and second, prioritize, or place a relative value, on each asset. This is not an easy task due to the different stakeholders present or represented on the team. Each party typically has a unique set of asset priorities that must be somehow reconciled into a comprehensive security system. Moreover, security teams must consider the dynamic nature of most environments and organizations; assets change for a variety of reasons (e.g., business restructurings, new technologies, new partnerships). Finally, the priorities of different assets are sometimes a function of the timing of the threat. For example, consider a two-week payroll application as the asset, and its inability to function as the risk. Threats related to this risk include hardware and software failure, malicious human actions, and even power outages. The priority of the payroll application asset will be much higher if the threat occurs immediately before direct deposits of employees' paychecks than if the threat occurs immediately after the employees get paid (which would provide you with two weeks in which to correct the problem).

Comprehensive security systems must therefore know an organization's assets to know how much to spend protecting which assets. Most organizations have current or near-current lists of the physical assets related to the distributed network: servers; motherboards; CPUs; memory modules; network interface, video, serial, and printer port cards; printers; scanners; multimedia devices; screens; power supplies; hard, floppy, and tape drives; CD ROM readers; modems; and network cables, hubs, switches, bridges, and routers. Proper documentation of the original copies of current software, service packs, and operating systems is recommended. Database administrators typically have lists and libraries containing descriptions of their system components.

Oftentimes many organizations do not have lists of their logical or intangible assets. Several alternative methods can help the security team get started. First, the team can begin by reviewing the organizational mission, strategies, and tactics. For each goal articulated in these items, the team can determine the assets required to make the desired outcomes occur. Alternatively, a framework provided by Lynn (2001) provides guidance by organizing security assets as related to communication, staff responsibilities, supplier issues, data and property protection, and operations.

Microsoft (2003) recommends organizing assets by asset type as the best practice. A multiattribute classification scheme for the organization's physical assets involves the following categories: asset type (e.g., hardware, software, library references, communication equipment or channel wiring, wireless and portable tools); scope of asset utilization (general or specific application); organizational owner of the asset; physical or logical residence for the asset (with the physical asset identification number if relevant); relevant contract ancillaries (service level, warranties, updates, key contacts, etc.); sensitivity of data (e.g., personnel files, customer information); backups; and whether audit records are kept. Nonphysical assets must also be categorized and evaluated, even though they may be more difficult to articulate and value. Nonphysical assets that create value for the organization include branding; public image; customer service and product quality reputation; information processing availability, confidentiality and integrity; and distributed network configurations.

Once identified, the assets need to be prioritized in terms of their relative importance for security-related investments. These priorities can differ between stakeholders (e.g., a salesperson may value a customer database more highly than a purchasing agent). Most security system methodologies utilize a four- to six-level method for priorities, with a wide range of labels for each level. For example, the highest level of asset may be termed "strategic" if impairment to the asset even for a small amount of time could cause serious interruption for the business. An example of a critical asset for an online bookstore would be its online shopping cart technology. The second highest level of priority could be labeled "tactical" meaning that while the company could probably continue to function for the short run in an impaired state, the company would suffer significant losses in the long run without these assets. The next level of priority may be termed "noncritical," indicating that the organization could function for a longer period without the asset; however, there may be cost or service effects on the long run that warrant consideration. The lowest level of asset priority could be labeled "insensitive," which means that if the asset is lost, the organization would easily find another way to create the value that the insensitive asset provided before the threat was realized.

Microsoft (2003) suggests that the best practices for data classifications are those that indicate the level of criticality or sensitivity toward loss or disclosure to unauthorized audiences. Sensitive data has a higher than average need for assurance of accuracy and completeness (e.g., financial data). Confidential data have a higher than average need for confidentiality (e.g., customer lists and ordering history). Private data are items that could lead to legal consequences for the organization if accessed inappropriately (e.g., personnel health information). Public data are data that will not create a loss of value for the organization if it were to be inadvertently accessed, even if there is a policy to keep that information proprietary. Note that these dimensions are not a continuum, and any one asset may register on more than one dimension.

IDENTIFYING RISKS, THREATS, AND PROBABLE LOSSES

There are many risks to our Internet-related wired and wireless networked environments. Exponential growth in the number of Internet transmissions is evidence that

Table 1 Sample Sources of Threats to Comprehensive Security

Advanced intrusion scanning techniques
Airwave jamming techniques
Burglary
Concurrency errors
Cross-site scripting
Digital payment alterations/fraud
Disabled automated audit functions
Distributed attack tools (that do not work)
Distributed denial of service attacks
Electronic and wave emissions
Encryption key mismanagement
External events
Fault tolerance errors or omissions
Filtering technologies bugs
Geographic separations from solutions
Hijacking sessions
Holes in operating system access controls
Human error
Identity theft
Intellectual property theft
Lack of employee training/supervision
Malicious human behavior
Misconfiguration of e-mail protocols
Naming convention errors
Out-of-date hardware protection systems
Out-of-date intrusion detection systems
Password mismanagement
Phone phreaking
Physical environment issues
Privacy violations
Race conditions (multistage attacks)
Remote access violations
Server and e-mail system bugs
Slack security culture
Sniffing
Spoofing
Stack overflow attacks
Stack smashing (memory overwrites)
Staged attacks
SYN flooding
Third-party services
Turnover in IT administration
Viruses, worms, and Trojan horses
Vulnerable CGI programs
Weak SNMP protocol
Web bugs (attached to HTML)

more and more organizations are utilizing the Internet for internal and external processes. This increased dependence upon the open Internet environment heightens the importance of being aware and prepared for the increased risks to security.

Table 1 provides a comprehensive list of the sources of threats to an organization's distributed network security. Threats can emerge from inside the organization as a result of leadership, technologies, humans, or physical environment problems. Threats can also result from external sources, either malicious independent parties, past or present business partners, or environmental changes (Rubin, 2001).

Significant risks can be caused by any of the stakeholders, employees, or unrelated hackers through viruses, manipulations of or errors/failures in software, hardware, telecommunications, or just plain old human error. Moreover, a comprehensive security team must discuss the importance of market or individual stakeholder perceptions of Internet-related risks, which tend to be higher than the actual risk of doing business online. These perceptions are fueled by the real fact that while a physical burglar can only enter one house at a time, a virtual burglar can enter many doors at one time, and it may be difficult to tell which doors have been opened.

Cisco (2001) identifies the primary perpetrators who reduce organizational security as unrelated hackers or snoops and unaware or disgruntled staff, either current or past employees. These perpetrators are interested in accessing, altering, damaging, copying, and stealing assets, whether digital or physical. Many times their most common vehicles include password guessing; introducing viruses, worms, and Trojan horses; reconnaissance attacks such as social engineering (whose purpose is information gathering); and the use of sniffers and scanners. Their primary targets are system access points; modification of authentication services and FTP functionality; interception of data; denial of service; and manipulation of e-mail accounts, databases, and confidential information. Other common threats include packet replaying, packet modification, and IP spoofing.

Microsoft (2003) recommends that the security team identify all of the possible threats that can access, alter, damage, destroy, or steal each asset. The threat assessment can be performed one asset at a time, each one considered independently or within groups of similar assets. For example, the asset of stored economic transaction information can face intentional and unintentional threats such as hardware or software errors; malicious codes such as viruses, Trojan horses, or worms; modifications made by authorized or unauthorized access, human error, negligence, or ignorance; changes in regulations or laws; natural disasters such as fires; or damage due to power surges, excess water or dust, or vandalism.

MAXIMUM COST OF CONTROLS

Once the assets have been identified and prioritized, the risks related to each of those assets articulated, and the threats that cause each of those risks researched and expressed, then the next step is to calculate the maximum cost of controls. The maximum cost calculation is an expected value algorithm founded on a simple economic principle that the security for an asset should never be more than the value of the asset, or the value of the loss an organization would experience should the asset become compromised.

The first step is to calculate the probability that the threats for each risk will affect the identified asset. This involves judgment, experience, and expertise. Microsoft

Table 2 Types of Comprehensive Security Controls

Ad blockers
Audit trails, logs, and alarms
Backups
Board of directors audit committee
Client-side digital certificates
Cluster server technologies
Code signing methods
Configuration structures and policies
Cookie crushers and other privacy protections
Cryptography
Digital identity certifications and authentications and public key infrastructures
Disaster recovery planning
DNS protection
Evidential issues
Firewalls and other filtering devices
Intrusion detection and response
Logging mechanisms
Mobile code and agents
Mobile code controls
Overall commitment to security through enforced policies
Overall management style and support
Password and other access technologies: user authorizations
Protection from threats such as viruses
Quality service providers
RAID techniques (disk mirroring)
Redundant capacity
Securing sessions: SSL and other options
SSL server certificates
Web application controls
Web page content access control

(2003) recommends that the best practices prepare the data for this decision either by external experts hired for this task or by employees working with the asset as the process increases security awareness and ownership. Likelihood of threat occurrence estimates can be developed from prior event records, past experiences, or published reports, or purchased from insurance or utility companies.

The second step is to calculate the conditional probability that a loss will occur given that the threat has occurred. This is a conditional probability and is based on the assumption that some of the time, the threats occur without causing loss to the organization. Examples of this are sniffing that occurs on insensitive data, or many types of human errors in daily activities.

The third step is to calculate the loss to the organization by estimating the loss in monetary terms. This estimate should include the replacement costs of the damaged assets but also the hours of time spent internally on the situation (time that could have been spent on productive tasks), as well as the intangible effects on relationships (e.g., the loss of trust, or timeliness and/or accuracy of customer orders). Again, past experience, current partners, and loss experts can assist with this estimation.

The fourth step is to calculate the expected value of the loss conditional on the threat occurring. This is the product of the monetary value estimate of the loss and the conditional probability that a loss will occur given that the threat has occurred. This multiplication product is the maximum cost that an organization should invest in security solutions for that asset for that threat/set of threats.

TYPES OF CONTROLS

Security controls, also called security tools, include an extensive choice of administrative structures and a variety of technologies ranging from antivirus software to audit-tracking solutions to dedicated hardware such as firewalls and intrusion detection systems. Table 2 lists many of the most common types of comprehensive security system controls.

Many security teams struggle to put some sort of organization on the pool of security controls. Some security teams organize the controls by type of asset, some by type of risk, and some by type of threat. Some teams categorize the control options into categories derived from the nature of how they work, as preventative, detective, or corrective of security violations. A similar classification scheme organizes controls as preventative, inspection, detection of internal failure, and correction of external failure. Other classification schemes separate the manual controls from the automated or computerized controls, or by whether the exercise of the control is mandatory or discretionary. Others organize controls with respect to where they apply to the distributed network asset components: hardware, software, mobile units, data, telecommunication channels, third-party issues, management, auditors, and employees. Yet others classify the controls according to their pervasiveness or specificity. Finally, many organizations classify their controls by the nature of the control (e.g., cryptography, firewall, intrusion detection system).

If the security team utilized databases to save their asset-security continuum information, they can write and retrieve efficient queries on assets, risks, threats, controls, or security goals (with either relational or indexed hierarchical databases). Regardless of how an organization organizes their information regarding security controls, security team members must (1) define the cost and purpose (expected benefit) of each control, (2) distribute knowledge and responsibility for the control, and (3) regularly audit the efficiency and effectiveness of each control.

Most comprehensive security teams write general organizational security policies that explain the culture toward security, top management's support for security, and top management's attitude toward intentional violations of security policy. These general policies typically discuss general e-mail, Internet access, password, remote access, and the security policies for communication of suspected violations and adverse events. Employees should be able to clearly understand the expectations for security-related behaviors, performance, and accountability. The two most common problems regarding general security policies are that (1) the policy is written in a manner that does not communicate clearly to the employees, and (2) the policy is not read or utilized by any employees. Oftentimes the general security policies are written to be in compliance with the recommendations of the corporate attorney or the independent auditor. Management may utilize a variety of tools, such as security newsletters, lobby videos, and guest speakers on security issues followed by question-and-answer periods, to increase the visibility of and commitment to general security policies.

The security controls that involve human interaction must be analyzed closely for effectiveness and hidden costs. First, there needs to be a balance between the benefit of the control procedure and the effect on the individual performing the procedure. Many times, too many controls on a task will make the task inefficient and the employee may skip the controls that slow down his or her performance. If controls are too loose, on the other hand, then there may be security threats due to lack of control. Examples of frustrating control plans and the human responses they elicit include too-long passwords that become post-its on monitors, too restrictive account lockouts that may result in denial of service, and too many door keypads that may motivate someone to prop the door open.

Cisco (2001) offers several "top" security controls to keep in mind as part of a comprehensive network security solution. The use of nonobvious passwords that are changed every three months, education of employees and business partners regarding the risks associated with e-mail attachments, and the use of the same current antivirus software throughout the organization are recommended. Immediate removal of network access when employees are terminated, the use of centrally administrated servers for all remote access traffic, and the removal of all nonutilized network services are also highlighted as important Internet-related controls for distributed networks. Cisco also recommends maintaining current versions of Web server software and the continual review of the organizations' assets and risks to determine the continued effectiveness of the comprehensive security solution.

Network configuration choices dramatically affect the security issues within the network. Network services should only involve the minimum necessary functionality, network administrator "back-doors" should all be closed off, and network policies and procedures should be clearly stated, kept current, and periodically audited. Networks are in a state of continuous change, with new topographies, resources, and users. Network-level controls include integrity-checking utilities and documentation of registry changes, system development and maintenance records, and network-testing results. Network scanning software is a preventative control that can be utilized to locate security weaknesses before there is an adverse security incident. Cisco (2001) recommends that organizations hire professionals with expertise for periodic security assessments on distributed networks.

Internet protocols (IP) determine the security rules with regards to Web servers and clients (browsers). Threats can attack Web servers directly or indirectly use the Web server to attack internal systems. Particularly vulnerable are Web servers' operating systems and Web server software. IP-based networks were originally engineered to be open systems, and efforts at maintaining openness while improving security can be inefficient. Even with sound security procedures, networks utilizing the IP protocols are still open to threats such as sniffing, man-in-the-middle attacks, and spoofing.

The most common forms of security control for Web servers are firewalls and intrusion detection systems, both of which filter users and packets coming in or going out to the Internet. These controls can be hardware, software, or hybrid solutions. Firewalls, for example, can stop unauthorized traffic both incoming and outgoing, can direct allowed incoming traffic, can protect network resources by hiding data and even internal systems from inquiring entry requests, and can have audit functionalities to log requests and traffic. Types of firewalls include packet filtering gateways, application gateways, and hybrid systems. Intrusion detection systems can work at the network, host, or application levels. Both firewalls and intrusion detection systems work to restrict access to authorized users and authorized types of packets coming in to or going out from internal authorized sources or users. The functionality is similar to the use of physical locks and peep-holes on doors: only those either recognized or with the appropriate key are allowed to enter.

Passwords are the most common access restriction control. Password policies should be developed by the network administrators and should clearly explain the responsibilities of the users for how and when their passwords are created, stored, utilized, changed, and deleted. Passwords need to be periodically modified by the users, and changes audited by the system administrators. Access management is simplified when user groups are established, with policies written specifically for each user group. Employees should be trained to delete all stored items, especially passwords, after a session on a computer at cyber cafes and airport lounges. Individuals should be allowed to try to log in with a password a few times before account lockout policies are activated. After lockout begins, the user must contact a system administrator for access.

Another essential control for networked environments is antivirus software. This software detects viruses entering or on a system, identifies the particular virus, and then removes the virus either before, during, or after it executes. Antivirus software utilizes three types of detection tools. Static detection is the strongest tool of the three, detecting the virus before execution. Further downstream is detection by interception, where the software catches a virus's attempts to infect either boot sectors, applications, or data files. Third is the method of detection by modification audits, which search for unexpected modification of executables. This control is the furthest downstream of the three methods, only detecting the virus well after it has been infecting the system. In this case, the antivirus software must systematically search through the system to remove all possible instances of the virus.

Controls related to validity and verification of identity involve authentication of the sender or receiver of transmitted messages. Controls related to authentication include digital certificates or public key certificates from certificate authorities, biometric devices, smart cards with private keys, and passwords. These authentication controls are recommended even in the situations where encrypted "tunnels" of secure traffic have been created on the Internet as virtual private networks. Other access measures include dual- and triple-item tests, challenge-response systems, and call-back procedures for remote access attempts.

Encryption tools can scramble characters within messages and/or messages within Web sessions (e.g., SSL) to make it more difficult for unauthorized access. Encryption can be utilized to ensure data integrity through the use of digital signatures, which are message digests encrypted with the sender's private key, to be decrypted usually by the receiver with the sender's public key. Public trust is encouraged because the sender first registers with a Certificate Authority who provides the sender with a private key for the encryption. The sender digests the message and then encrypts the message and its digest with the private, or secret, key. If the recipient of the message is doubtful about the sender, he/she can check with the Certificate Authority for validation of the identity of the sender.

Public-key, private-key dual encryption pairs are utilized to provide sender and receiver authentication and confidentiality. Pretty good privacy (PGP) and S/MIME (secure/multipurpose Internet mail extensions) are two methods that can authenticate the originator and provide message confidentiality. S/MIME digitally signs and encrypts each e-mail message: sender authentication (validating the origin), message integrity (detecting whether any changes have been made to the contents), message confidentiality (no unauthorized recipients), and nonrepudiation (validating origin, receipt, timing, and contents). After transmission, files stored on a network server may also need to be secured. Microsoft's encrypting files systems (EFS) is a Microsoft NT example of a symmetric key encryption for stored files.

Network auditing tools assist administrators to monitor activity in order to detect rather than prevent any problems. For example, audit logs contain records of logins, file accesses, password changes, and logoffs. Most audit software has extensive log files and audit trails that are useful to research the source of a bug or a break-in, and to determine the amount and scope of the damage. Audit logs are also useful documentation for legal testimony and for claims toward insurance recoveries. Risks associated with auditing include unauthorized modification to the audit log tables and system performance degradation.

Recovery controls include the creation of backups, and redundant processing capacity in the case of contingency or disaster recovery planning. Comprehensive security teams need to create backup policies, decide which files are backed up (e.g., databases, mail servers, user files, registries), how often they will be created, on which media, on which type of backup device (e.g., CD writers, tape drives, network shares), and when the backups will be periodically tested. Storage of backups can be onsite in a fireproof safe, or offsite, either in a hot site, cold site, or safety deposit box.

Disaster recovery plans typically identify an incident response team that is trained to know how to respond to an emergency. The location of reboot disks, the original software installation disks and license agreements, the names and addresses of vendor contacts, records of disk partition descriptions, and hardware configurations are all important to provide to the incident response team members.

Fault tolerance, disk mirroring, disk duplexing, and desk striping with parity are different types of redundant array of inline disks (RAID) techniques, which are methods of backing up distributed network servers. Each method has its benefits and drawbacks. For example, disk mirroring backs up more of what is on the original disk than some of the other methods, but it is the least efficient in terms of disk storage utilization. Desk striping with parity separates data and parity information so that if there is a disaster, the information can be rebuilt with greater probability. This control works best when large database read operations are performed more often than write operations, and is weak when high-speed data retrieval is needed, or when many write operations are required.

Cluster server technologies are groups of servers that have been programmed to act as a single, large unit. Resources and responsibilities are distributed between the servers within the cluster. One of the main advantages is the resulting low probability of down time because if one server fails, the other servers continue to function. The more servers and backup systems that are added to the cluster, the lower the risk of system failure and the greater the likelihood of system availability. Another benefit of cluster server technologies is load balancing, which maximizes performance of the network by improving the throughput, a primary goal of Web, e-commerce, and FTP application servers (Cisco, 2001).

Additional topics oftentimes included in comprehensive security controls and procedures involve event handling, software patches, operating system weaknesses, architectural segregation of internal/external zones when using firewalls, and application development policies related to secure access. The best practices for each of these topics depend on the platform and software involved. However, the creation, analysis, and maintenance of documentation regarding corporate error level philosophy, the number and type of errors, and clearly articulated committee responsibility are important across all of these topics.

Once the comprehensive security set of controls has been identified, prioritized, and organized, the security team must decide the specifics of the policies, methodologies, and technologies that will be utilized. Then the security team must decide how to communicate the new or existing policies, changes, and security alerts to all of the stakeholders utilizing the Internet-related distributed network. Third, the security team must design an implementation plan, execute the plan, and periodically audit and assess the continual effectiveness of the comprehensive security system.

SECURITY GOALS

The effectiveness of a comprehensive security system can only be measured against stated, measurable comprehensive security goals. Therefore, it is the responsibility of the security team to evaluate the mission and strategies of the organization along with the prioritized assets in order to develop a set of explicit security goals. These goals may be stated in terms of the organization's mission and strategies, the functions of the prioritized assets, or the most devastating threats that will be avoided or reduced by the security system.

Lynn (2001) suggests security goals to be organized in terms of corporate communications (involving employees, insurance companies, suppliers, customers, regulatory agencies, technical support services, etc.), staff responsibilities (in terms of the chain of command, individual responsibilities, and checks for task coverage), supplier issues (how long would it take to get you up and running?), data protection (system performance, system of backups, and schedule of testing), property protection (equipment, furnishings, library, supplies, etc.), and operations (software, operating systems, emergency lighting, sprinklers, smoke detectors, extinguishers, utility shutoffs, backup generators, evacuation procedures, exit routes).

CONCLUSION

The purpose of this chapter was to provide a state-of-the-art overview of the primary components, issues, and solutions for businesses interested in comprehensive Internet-related network security. This chapter, therefore, presents a comprehensive security strategy organizations can utilize to plan, analyze, implement, and maintain their Internet-related security systems. The first step creates a cross-functional security team with top management membership. The second step identifies both the priority assets (including property, data, and people) and risks affecting those assets. The third step evaluates the specific threats that create the risks, and calculates the maximum cost of controls. The fourth step researches and plans appropriate controls for the priority assets and risks. The nature of these controls ranges from general to specific controls, preventative to corrective, and manual versus computerized. Finally, the correlation between the chosen controls and the overall security goals for the organization is assessed.

This structure of the continuum emphasizes the important interrelationships between the elements of the asset-security model. The process of working through the dyads within the model will develop a deeper understanding of the security relationships. The importance of top management "buy-in" is emphasized, for validity, economic, and other methods of support to be taken seriously.

GLOSSARY

Access control A set of mechanisms and policies that restrict use of various computer resources.

Account lockout A security feature that closes log-in access to a user account if a number of failed logon attempts occur within a specified duration of time. Lockout is based on security policy lockout settings.

Antivirus software Used to prevent viruses from damaging a computer or to remove viruses from the computer.

Application gateway An application program that runs on a firewall system on a proxy-server between two networks to keep the computers behind the firewall protected. Two connections, one between the client and the proxy-server and the other between the proxy-server and the destination, are required. When all communication is conducted through the proxy-server, the computers behind the firewall are protected.

Audit-tracking solution An automated solution for activities like coordination, assignment, evaluation, and response development for internal and external audit findings and recommendations, which would otherwise be time consuming and prone to delays. A Web-based solution may make access to the audit information easy, secure, and very cost effective.

Audit trail A record that shows who has accessed a computer system, and where and what operations were performed during a given period of time. Audit trails are used for system reliability, accuracy, security, and recovery purposes.

Authentication A process of identifying an individual utilizing a computer resource, usually based on a username and password.

Authorization A process of giving access to a system based on the identity of the user.

Boot sector The first sector on every floppy and hard disk, containing an executable program that is executed every time a PC is switched on or booted.

Biometric devices Devices that use biometric techniques to identify and authenticate individuals or users. Biometrics are the authentication techniques that rely on measurable physical characteristics that can be automatically checked, including computer analysis of fingerprints, speech, and facial characteristics.

Callback procedure A procedure executed from a remote log-in attempt. Basically the network server or main computer sends a signal to the authorized computer node, which can be received only if the log-in attempt is being made at that authorized computer.

Certificate authority A trusted third-party organization or company that issues digital certificates that are used to create digital signatures and public–private key pairs. They are a critical component in data security and e-commerce because they guarantee the two parties exchanging information are really who they claim to be.

Cluster A collection of systems connected to share devices, whereupon all systems can read or write to all of the devices. The individual computers in a cluster are referred to as nodes.

Cluster server technologies Usually of two types—shared-disk and shared-nothing. The former implementation allows all cluster participants (nodes) to own and access cluster disk resources. In the latter, several nodes in the cluster may have access to a device or resource, but the resource is owned and managed by only one system at a time. Each node has its own memory, system disk, operating system, and subset of the cluster's resources.

Cold site A disaster recovery service that allows a business to continue computer and network operations even in the face of computer or equipment failure, by providing office space, but the customer provides and installs all the equipment needed to continue operations. It is less expensive, but takes longer to get an enterprise in full operation after the disaster.

Cross-site scripting (also called XSS) Occurs when a Web application gets malicious data from a user, usually in the form of a hyperlink. The request looks less suspicious to the user when clicked on, but may lead to threats like account hijacking, changing of user settings, cookie theft/poisoning, or false advertising.

Cryptography The art of protecting information by transforming it (by encrypting it) into an unreadable format, called cipher text. Only those who possess a secret key should be able to decipher (or decrypt) the message into plain text.

Cyber café A public or private location where one can connect via a computer to the Internet, typically for a flat fee plus a time-for-services fee.

Database A software application designed to help users organize information, so that a computer program can quickly find and select desired information. Databases can be thought of as an electronic filing system.

Digital certificate An attachment to an electronic message to verify that a user sending a message is who he or she claims to be, and to provide the receiver with the means to encode a reply. It is issued by a certification authority (CA), containing your name, a serial number, expiration dates, and a copy of the certificate holder's public key.

Digital signature Acts as the functional equivalent of a paper signature. A digital signature uniquely identifies a sender and can make a document binding. It can also be used to ensure that the original content of the message or document that has been conveyed is unchanged. They are becoming increasingly important in e-commerce as a component of authentication schemes.

Disk duplexing or mirroring A technique in which data are written to two duplicate disks simultaneously. If one of the disk drives fails, the system can instantly switch to the other disk without any loss of data or service. It is commonly used in online database systems where it is critical that the data be accessible at all times. In other words, it makes for reliable data redundancy.

Disk striping A technique for spreading data over multiple disk drives. The computer system breaks a body of data into units and spreads these units across available disks. Disk striping can speed up operations that retrieve data from disk storage.

Distributed network A network where the resources are divided up and shared among many computers to reduce the burden on any one system. A distributed network is the main characteristic of the Internet.

Encryption The translation or coding of data into a secret code for data security. Encrypted data is called "cipher text" while unencrypted data is called "plain text."

Fault tolerance The ability of a system to respond gracefully to an unexpected hardware or software failure. There are many levels of fault tolerance, from the ability to continue operation in the event of a power failure to performing every operation on two or more duplicate systems, so if one fails, the other can take over.

Firewall A system designed to prevent unauthorized access to or from a private network. Firewalls can be implemented in either hardware and/or software. All messages entering or leaving a network pass through the firewall, which examines each message and blocks those that do not meet specified security criteria.

FTP (file transfer protocol) Used on the Internet for sending or transferring files from one location to another.

Host A computer server in a network that runs applications used by or from other computers (e.g., Web servers, file servers, and application servers).

Hot site A commercial disaster recovery service that allows a business to continue computer and network operations even in the face of a failure. The affected enterprise can move all data processing operations to a hot site that has all the equipment needed for the enterprise to continue operation, including office space and furniture, telephone jacks, and computer equipment.

Integrity A measure to protect data or resources from unauthorized modification to ensure information security. Data whose integrity cannot be ascertained are said to be corrupted.

Integrity checking utilities Programs used for executing file system integrity checks on remote hosts from a server and sending reports via e-mail.

Internet The world-wide collection of interconnected networks and computers using and sharing information based on set rules defined by protocols like the TCP/IP.

Intrusion detection system An inspection system that checks all inbound and outbound network activities and identifies suspicious patterns that may indicate a network or system attack or break-in.

IP spoofing A technique to gain unauthorized access to computers. Here an intruder sends a message to a computer with an IP address indicating that the message is coming from a trusted host, thereby beguiling the user.

IP-based network A network running on the Internet protocol. Each computer on the Internet (known as a host) has at least one IP address that uniquely identifies it from all other computers.

Key A string of bits used in cryptography that allows people to encrypt and decrypt data. The key determines the mapping of the plaintext to the ciphertext.

Load balancing Distributing processing and communication activities evenly across a computer network so that no single device is overwhelmed. Busy Web sites typically employ two or more Web servers in a load balancing scheme. If one server starts to get overloaded, requests are forwarded to another server that has more capacity.

Log-in The opening procedure used to connect to a computer system by using credentials like "username" and "password." Also, the act of typing in your username and password.

Log-out The closing procedure used to disconnect from a computer system or formally end a session with the system. Once you log out, you must log back in to regain access.

Network Formed when any two or more computers are connected together to share resources. Networks are established to enable the sharing of information, files, and applications between users via multiple systems.

Network administrator "back-door" A hidden "entrance" for an administrator to access a network.

Online shopping cart A piece of software that acts as a catalog for an online store and also performs the ordering process. A shopping cart is an interface between a company's Web site and its infrastructure that allows consumers to select merchandise, review what they have selected, make necessary modifications or additions, and eventually, purchase the items in the cart.

Operating system A software that serves as an interface between the system hardware and the users. The operating system is the most important program on a computer. Operating systems perform basic tasks, like recognizing input from the keyboard, sending output to the display screen, keeping track of files and directories on the hard drive, and controlling peripheral devices such as disk drives and printers.

Packet A piece or block of message/information/data transmitted over a network (packet-switching network) that contains the destination address labeled to the data. In IP networks, packets are often called datagrams.

Packet filtering A method of controlling access to a network by analyzing incoming and outgoing packets and letting them pass or stopping them based on the IP addresses of the source and destination, thereby serving as one of the many techniques used for implementing security firewalls.

Packet filtering gateway A packet filter that analyzes each IP packet at the network layer and determines whether to pass or block it based on a set of rules. If it allows communication between two specific addresses, packets are allowed to travel through the firewall to the specified address. If no rule is available for a given address, the packet is rejected and not allowed to pass through the firewall.

Packet modification Modifying a packet that leads to sending incorrect or corrupted information across the network.

Packet replaying Replaying a packet to find out what was previously sent.

Password A code used to log in or gain access to a locked or secure system.

PGP (pretty good privacy) A message encryption technique based on the public-key method, which uses two keys—one public key that you share with anyone from whom you want to receive a message, the other a private key that you use to decrypt messages that you receive. PGP is one of the most common ways to protect messages on the Internet because it is effective, easy to use, and free.

Phone phreaking The act of cracking the phone network or entering it in an unauthorized manner to use it, for example, to make "free" phone calls.

Private key The undisclosed key in a matched key pair (that is, the private key and the public key) that each party must safeguard for public key cryptography. The private key is known only by the recipient of the message while the public key is known to everyone.

Proxy server An Internet server that controls the access of the client computers to the Internet. It can stop clients from accessing undesirable Web addresses, improve performance by storing Web pages locally, and hide the internal network's identity, so that it is difficult for external users to monitor.

Public key certificate A digitally signed document that serves to validate the sender's authorization and name. The document consists of a specially formatted block of data that contains the name of the certificate holder, the holder's public key, and the digital signature of a certification authority for authentication.

Public key encryption A cryptographic system using two keys—a public key known to everyone and a private or secret key known only to the recipient of the message. For instance, when John wants to send a secure message to Jane, he uses Jane's public key to encrypt the message. Jane then uses her private key to decrypt it so that she can make any sense out of it.

Public key A value provided by some designated authority as a key which, when combined with a private key, can be used to effectively encrypt and decrypt messages and digital signatures.

Query A request for information to a database.

RAID [redundant array of independent (or inexpensive) disks] A combination of two or more drives (frequently on servers) used together for better fault tolerance and increased performance.

Reboot Restarting a computer.

Reboot disks Disks used to reboot a system.

Remote access The ability to log on to a network from a different, distant, or remote location.

Service pack An update to a version of software that fixes an existing problem, such as a bug, or provides enhancements to the product. These enhancements and fixes will appear in the next version of the product.

S/MIME Multipurpose Internet mail extension (MIME) is originally a standard for defining the types of files attached to standard Internet mail messages. It is used in situations where one computer program needs to communicate with another program about what kind of file is being sent. Secure multipurpose internet mail extension (S/MIME) describes how encryption information and a digital certificate can be included as part of the message body.

Smart card A small electronic device about the size of a credit card that contains electronic memory, and possibly an embedded integrated circuit (IC).

Sniffer A program or device that monitors the flow of data in a network. Can be used for both legitimate network management functions and stealing information off a network. Sniffers are very dangerous because they are virtually impossible to detect and can be inserted almost anywhere.

SSL (secure sockets layer) A program layer that manages the security of message transmission in a network.

Software Computer instructions or data that are stored electronically. In contrast, storage devices and display devices that show or operate the instructions or data are hardware.

SYN flooding Bombarding a system with dozens of false connection requests a minute that leads to degradation of the system's ability to give service to legitimate connection requests. This attack is also called the denial-of-service attack.

System access point A hardware or software that acts as a communication hub for users of a wireless device to connect to a wired LAN. They are important for providing heightened wireless security.

System administrator The individual responsible for maintaining a multiuser computer system. Small organizations may have only one system administrator, while large organizations and enterprises may have a team of system administrators.

Trojan horses Impostor files that claim to be something desirable but, in fact, are malicious. Unlike viruses, Trojan horse programs do not replicate themselves. Trojans contain malicious code that when triggered cause loss, or even theft, of data.

Virtual burglar A type of hacker that steals information on the Web.

Virus A program or piece of code that is loaded onto your computer without your knowledge and runs against your wishes. Viruses can also replicate themselves. A simple virus can make a copy of itself over and over again and is very dangerous because it will quickly use all available memory and bring the system to a halt. An even more dangerous type of virus is capable of transmitting itself across networks and bypassing security systems.

Web commerce (e-commerce or electronic commerce) Business over the Internet, specifically, the electronic transfer of value across the Internet in exchange for the delivery of a service or product.

Web server A computer that stores and delivers Web pages to other computers.

Web server software The software designed to be used on a Web server, for accepting, managing, and responding to Internet requests for sessions.

Web server's operating system The operating system that controls the functionality of a Web server.

Wired network A network wired together through cables or physical connections as compared to a wireless network where radio frequencies are generally used to make connections.

Wireless network A network interconnected with radio waves or other signals instead of through a cabling or physical connection.

Worm A program or algorithm that replicates itself over a computer network, usually performing malicious actions, such as shutting the system down or using up the computer's resources. Worms replicate themselves from system to system without the use of a host file as opposed to viruses, which require the spreading of an infected host file.

CROSS REFERENCES

See *Authentication; Biometric Authentication; Computer Security Incident Response Teams (CSIRTs); Computer Viruses and Worms; Denials of Access Attacks; Digital Signatures and Electronic Signatures; Disaster Recovery Planning; Encryption; Firewalls; Forensic Science for Computer and Network Security; Internet Security Standards; Intrusion Detection Systems; Passwords; Physical Security; Public Key Infrastructure (PKI); Secure Electronic Transmissions (SET); Secure Sockets Layer (SSL).*

REFERENCES

Allen, J. (2001). *The CERT guide to system and network security practices.* New York: Addison-Wesley.

Anderson, R. (2001). *Security engineering: A guide to building dependable distributed systems.* New York: Wiley.

Cisco. (2001). *A beginner's guide to network security.* San Jose, CA: Cisco Systems.

Garfinkel, S., and Spafford, G. (2001). *Web security, privacy & e commerce.* Sebastopel, CA: O'Reilly.

Greenstein, M., and Vasarhelyi, M. (2002). *Electronic commerce: security, risk management, and control.* New York: McGraw–Hill/Irwin.

IOMA. (2000). *Outsourcing: How to know when to let someone else do the work* (Security Directors Rep. No. 00–8). New York: Institute of Management & Administration.

Lynn, J. (2001). Disaster planning: Are you prepared for the worst? *Commercial Law Bulletin, 16*(4), 30–31.

Microsoft (2000). *Best practices for enterprise security.* Retrieved March 10, 2002 from www.microsoft.com/technet/security/bestpracbpent/bpent/sec1/secplan.asp

Rubin, A. (2001). *White-hat security arsenal: Tackling the threats.* New York: Addison-Wesley.

SANS. (2002). *The twenty most critical Internet Security vulnerabilities (updated): The experts' consensus.* Retrieved March 20, 2002 from http://www.sans.org/top20.htm

Scalet, S. (2002). How to plan for the inevitable. *CIO, 15*(11), 74–82.

Stein, T. (1999). Security starts with sound practices. *Electronic Commerce World, 9*(10), 46–47.

FURTHER READING

CERT Coordination Center: http://www.cert.org

Internet Storm Center: http://isc.incidents.org

Internetworking technologies handbook, Chapter 51: Security technologies (pp. 1–12): http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/security.htm

Internet security policy: A technical guide: http://www.packetstormsecurity.org/docs/infosec/isptg/

H

Health Insurance and Managed Care

Etienne E. Pracht, *University of South Florida*

INTRODUCTION

Health insurance in the United States traces its origins to 1847 when coverage became available against rail and steamboat accidents. The next major development occurred almost 100 years later during World War II, when restrictions on wage increases impelled employers to offer fringe benefits such as health insurance as a means to attract additional laborers. Since then, the industry has grown increasingly sophisticated and, in the absence of national health insurance coverage, is now controlled largely by the private sector. The number of direct health and medical insurance carriers grew from 47 to 3,238 during the 20th century (Raffel & Barsukiewicz, 2002; U.S. Census Bureau, 2000). In 2001, 70.9% of the population purchased insurance coverage from private firms, primarily through their place of employment. Government health plans, primarily Medicare and Medicaid, covered almost 24.7% of the population, and a fraction of the population received health insurance coverage from a combination of private health plans, Medicare, and Medicaid. (The insurance status figures contain some duplication. According to the U.S. Census Bureau, *Current Population Survey,* 1.99% of the population received coverage from Medicaid and private insurance, 7.61% received coverage from Medicare and private insurance, and 1.89% received coverage from Medicare and Medicaid. The privately insured population obtains coverage primarily through employment [64% of the total population].) In that same year, 14.6% of the population did not have any health insurance coverage (U.S. Census Bureau, 2002).

Total U.S. national health care expenditures in 2000 were $1.3 trillion, accounting for more than 13% of the gross domestic product. According to projections from the U.S. Centers for Medicare & Medicaid Services, health care expenditures will grow to $2.8 trillion by the year 2011 (Centers for Medicare & Medicaid Services, 2001). Health insurance coverage, the fee for service structure in particular, along with advances in medical technology, has been identified as a major contributor to the enormous growth in U.S. health care expenditures (Dorenfest, 2000; Fuchs, 1996; Weisbrod, 1991). Understandably, emphasis has shifted to cost containment, and through pressure from employers and the government, insurers are increasingly forced to compete on the basis of cost. Competition among insurers has traditionally focused on reducing premiums by manipulating deductibles and copayments and avoiding high-risk individuals. More recently, managed care has introduced provider incentives to reduce utilization while promoting outcomes-based strategies to implement most effective courses of treatment or "best practices." The Internet revolution has created the potential for cost reductions through improved productivity in the administration of health insurance and, more broadly, the health care industry.

This paper discusses the application of the Internet and e-commerce technologies in the health insurance industry. Current examples augment an industry system-wide perspective for the discussion. The structure of the industry, and in particular the lack of standardization, is highlighted as one of the defining parameters that determine the implementation and effective use of information technologies. The next section discusses the relationships between existing technologies and the various stakeholders of the industry. The section Technical and Practical Applications examines the technical and practical considerations of implementing information technologies. The section E-application analyzes five categories of applications of the Internet and e-commerce in the health insurance industry. The final section presents a summary and conclusions.

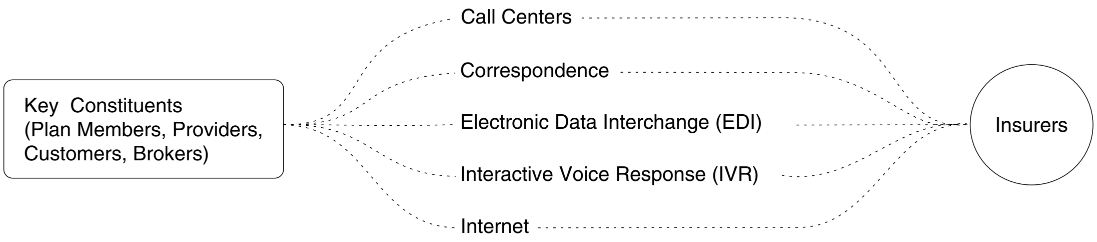

Key Constituents
(Plan Members, Providers,
Customers, Brokers)
Call Centers
Correspondence
Electronic Data Interchange (EDI)
Interactive Voice Response (IVR)
Internet
Insurers

Figure 1: Methods of interaction between insurers and their key constituents.

HEALTH INSURANCE, THE INTERNET, AND E-COMMERCE

Interactions between insurers, as third-party payers, and their main constituents, enrollees, employers, and providers, take place to process basic insurance functions. Internet applications concerning these functions are discussed in more detail under E-application. Enrollees and employers may contact insurers to purchase, verify, change, or drop coverage. Providers may contact (or be contacted by) third-party payers regarding patient eligibility confirmation, claims submission, or patient referral information. Insurers, in turn, may use this medium to facilitate some functions relating to contractual agreements with providers such as certification or recertification. The Internet, as noted below, is a viable alternative tool to interaction between third-party payers and their key constituents (Figure 1).

Internet functions as both a replacement and complement to existing information technology (IT). For example, in 1999, PreferredOne, a health benefits management company serving the upper Midwest, reported that a clinic reduced its telephone calls by more than 50% after implementing an online system for claim status inquiries, eligibility verification, and payer's contact information (Hofflander, 1999). The advantages are obvious: online services per contact episode are substantially less costly compared to telephone calls and, once implemented, require minimal labor resources, yielding additional savings. According to a California PPO the cost of a typical call to their center was approximately $7, compared to $0.20 for an online system (Runy, 2000). At the same time, enrollee services may be greatly improved.

In addition to speed, the Internet offers a quantum leap forward in functionality, touting a new level of sophistication in information display using a graphical user interface, interactive capability, and vastly improved organization of complex information. It represents a platform for tremendous enhancement and integration of functions. Voice, video, and data traffic can all be combined under one platform to increase efficiency. Electronic data interchange (EDI) and interactive voice response (IVR) functions will likely be handled increasingly by Internet-based applications. Certainly, Web-based applications can make use of readily available e-mail technology to replace traditional postal service correspondence. The Internet also offers a greater potential for networking and connecting geographically dispersed organizations. An excellent example of this is the Colorado-Fayette Medical Center, which is spread over seven campuses in three towns and two counties, consisting of a 38-bed hospital, two rural health clinics, a nursing home, an assisted-living facility, an outpatient center, and a unit for Alzheimer's patient (Bacus & Zunke, 2001). A new wide area network with Internet access enabled the organization to "improve its Web presence, allow radiologic image viewing at all sites, negotiate more favorable prices from vendors, implement electronic communication for staff members, and take advantage of on-line education opportunities" (Bacus & Zunke, 2001).

In the case of EDI, the Internet also acts as an alternative platform for electronic transactions (Figure 2). Smaller provider groups that previously used paper forms because they found the resource requirements of more traditional EDI systems prohibitively expensive may find that the Internet offers a cost-effective method for implementing the approach (see National Research Council, 2000, p. 90, for a specific example). In this regard, the Internet is predicted to have a prominent role as the Health Insurance Portability and Accountability Act (HIPAA) is implemented. HIPAA regulations and practical and technical considerations pertaining to the use of the Internet for EDI are discussed in more detail under Technical and Practical Considerations.

Stakeholders

Groups affected by applications of the Internet and e-commerce technologies in health insurance and managed care are enrollees, employers, providers, brokers or intermediaries, the government, Internet applications

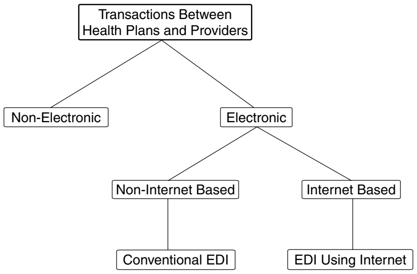


Figure 2: Electronic transactions between health plans and providers.

providers, and, of course, the insurers themselves. The information distributed or made available to the respective stakeholders must be properly structured and dedicated to specific needs. Enrollees require information concerning their coverage, personal health, and care. They may be interested in tracking administrative matters as well, involving claims, copayment schedules, and scheduling of secondary visits. Employers are primarily interested in cost and quality aspects, billing, and other administrative functions pertaining to the structure of insurance plans and employee utilization of health resources. Providers may be interested in the more technical aspects of medicine and efficient communication with various insurers. Finally, insurers are interested in remaining competitive, reducing the cost of care without compromising quality, and meeting state regulations. For these purposes, streamlining administrative procedures is crucial, a process in which the Internet can play an important role. Concerning patient care, efficient delivery involves not only tremendous amounts of information in real time, but also the cultivation of stable, consistent, factual, and good relationships between the various stakeholders, particularly providers and plan members. The behavioral relationship between plan members and providers has been identified as an important component in the delivery of efficient and cost-effective care. While further discussion of this topic is beyond the scope of this chapter, it has been widely discussed in the literature, for example, in the context of Medicaid managed care. For more in-depth discussions of this and related issues, see, for example, Hurley, Freund, and Paul (1993); Luft (1978); Manning, Leibowitz, Goldberg, Rogers, and Newhouse (1984); Newhouse, Schwartz, Williams, and Witsberger (1985); and Pauly, Hillman, and Kerstein (1990).

The IT and labor resource demands of a system which satisfies these needs can be challenging. Insurers have two advantages and incentives over other private sector stakeholders for transforming the e-health marketplace from one dominated by independent, noninteractive, general information Web sites to one coordinated, engaging, and personalized. First, they have the strongest fiscal incentives, since successful implementation of a streamlined and effective e-health mechanism would be extremely helpful concerning their primary goal of containing costs and maximizing profit. Second, insurers already possess the required administrative and financial resources and the information required, most of which is already in electronic format (Korpman, 2001).

The Internet and Intermediaries

A profound impact of the Internet is the transparency of information. The ease of online comparisons of products, based on defining characteristics such as price and quality, has made sellers susceptible to greater competition and given consumers enhanced decision-making power. In addition to simplifying information and reducing the cost of price and quality comparisons, the Internet acts as an effective tool for coordinating activities traditionally involved with selecting and purchasing goods and services. While consumers as a group are unlikely to make purchasing decisions (e.g., health insurance plan selection, copayment levels, and deductibles) based solely on Internet searches, insurers will be forced to modify traditional marketing practices more and more as the population becomes increasingly savvy about the Web.

Cost reductions generated by the Internet through advances in operational efficiency are likely to be transferred to customers due to increased price transparency. (While the focus here is on cost, it should be noted that enhanced transparency might result in increased competition based on other aspects instead, for example quality, while holding the price constant. In any case, the consumer should benefit from the change.) As the segment of the population that embraces the Internet grows, competition will increase, further narrowing operating margins. Consequently, increasing pressure on insurers to reduce costs will force them to market directly to consumers and decrease reliance on brokers or intermediaries, reducing or eliminating commissions, which can reach as high as 30% of administrative expenses of managed care plans (Goldsmith, 2000). This process is known as disintermediation. Present examples include investment firms eliminating stockbrokers by offering services online and large automotive corporations placing their parts specifications online to solicit competitive bids from manufacturers. Amazon.com, Priceline.com, and Monster.com (which have taken on the status of household names) all act to eliminate or replace the traditional middleman. The e-health marketplace is already utilizing this direct marketing function of the Internet (see E-application) on a limited scale.

TECHNICAL AND PRACTICAL CONSIDERATIONS

Considerations pertaining to the implementation of Internet, or electronic information technology, may be grossly classified as nontechnical and technical. Nontechnical considerations range from general ignorance and reluctance to important logistical and managerial challenges involved with adoption of new or standardized electronic information technologies, whether they are Internet-based or otherwise. The magnitude of these challenges and, perhaps more importantly, the associated operational risks will vary according to factors such as the scale of the proposed change, the organization's size, current information technology, and management structure. The fragmentation of existing systems concerning types of applications, platforms, operating systems, and network protocols (see discussion below) is only one in a series of obstacles the industry faces. Transforming an existing system, whether it is paper-based, consists of a single self-contained computer system, or is comprised of disparate systems, may be costly and time-consuming as one moves through the stages of implementation, including installing, testing, training, and maintaining. Prior to implementation, additional resources may be required to gather the information necessary to make sensible decisions involving the new system. In large organizations, particularly those that are geographically dispersed and maintain separate management structures, reaching an agreement over, for example, the software vendor, the type of system, and the functionality of that system may be difficult. This may be an especially sensitive concern if one considers long-term maintenance contracts that may accompany new systems. Individual providers may be particularly reluctant to enter such agreements.

The National Research Council (2000) identifies and discusses the relative importance of five technical considerations:

Bandwidth—The rate, measured in bits per second, at which information is transmitted through a network.

Latency—The time required to transmit data across the network. Real-time applications, for example, require low latency, while other applications, e.g., e-mail, are less demanding in this regard.

Ubiquity—The relative accessibility of a network. This function depends on such factors as geographic dispersion and regulations regarding participation.

Availability—The continuous accessibility of the network, the applications of which it is composed, and the services it offers.

Security—Availability, confidentiality, and integrity of information.

With regard to bandwidth and latency, advances in technology, coupled with the relatively modest requirements of administrative and customer service tasks, make these technological concerns largely nonissues. The significance of ubiquity, or relative accessibility of a network, requirements is contingent on the scale and frequency of use and the dispersion of users. The Internet is expected to play an important role in reducing the degree of ubiquity required by providing low-cost access to providers and administrators of geographically dispersed organizations (National Research Council, 2000). The importance of availability depends largely on the urgency of the transaction being executed. For example, referral certification and prior authorization requests concerning an enrollee's immediate care needs require relatively high availability. Other, less formidable and sometimes trivial, obstacles include lack of processing capacity, poor quality of applications, and inadequate understanding of the Internet and its potential benefits (Schaich, 1998).

Security was identified as the most important technical consideration in the adoption of Web-based and other e-commerce applications by the health insurance industry (National Research Council, 2000). The discussion here presents only an outline of the security issues. A more in-depth analysis is beyond the scope of this chapter. Interested readers may refer to Department of Health and Human Services (2000) and National Research Council (2000). Current information may be found at http://www.hhs.gov/ocr/hipaa/finalreg.html (accessed October 9, 2002).

The most sensitive aspect of security in electronic commerce concerns privacy, or, in other words, the confidentiality and authorized use of patient data. Many administrative transactions, for example medical claims, include highly sensitive and personal information pertaining to patients' health status, ailments, diagnoses, and treatments. Potential misuses of individual health information include unfavorable insurance decisions, employment decisions, and other adverse social outcomes (Lumpkin, 2000).

In addition to the technical aspects, privacy of personal health data involves important political and ethical considerations. Indeed, patient privacy has been debated in Congress since HIPAA was enacted in 1996. In 1998, Donna E. Shalala, the Secretary of the U.S. Department of Health and Human Services (HHS), submitted recommendations to Congress concerning security requirements for patient records. The recommendations called for specific legislation to protect patient privacy or individually identifiable health information. In the absence of such legislation, a 1996 HIPAA mandate stipulated that HHS would impose its own privacy, confidentiality, and security regulations. The final regulations resulting from the HIPAA mandate apply to insurers, health care clearinghouses, and certain health care providers. The objective of the resulting rules is to give consumers increased control over health information. Essentially, providers and insurers are required to seek consent from patients before information can be shared for treatment, payment, and other health care operations. The regulations also specify disclosure requirements, appropriate uses of related information, and available recourse if privacy protections are violated.

The second, far less politically charged, aspect of security relates to the process of maintaining and transmitting information. Protection of information is of vital importance both when it exists at rest (e.g., in a physician's office) or in transit (on paper or electronically) during transmission. Protection of printed information involves measures such as controlling and monitoring access and implementing organizational procedures that focus on the security of health information. Protecting electronic information, particularly during transmission, involves a different set of challenges. It is noteworthy that these security concerns are gradually diminishing with advances in encryption, authentication, and certification technologies (Department of Health and Human Services, 2000; Gerberick, Huber, & Rudloff, 1999).

Secure access to information, both in reality and as perceived by health plan constituents, is required if the Internet is to be used successfully in managing basic health insurance functions. The rapid growth in the number of health-specific Internet users and Web sites indicates that consumers as a group have reached a reasonably high comfort level with this technology. According to recent surveys (Fox & Raine, 2000; Taylor, 1999) 60 to 70 million U.S. residents went online in search of health information at the turn of the millennium, and 17,000 specialized Web sites had been created in response to this demand (Taylor, 1999).

The legislation is not without its critics. The health insurance industry argued that it is already held to rigorous security standards and that implementation of the proposed rules would impose substantial new costs, which would, at least in part, offset the benefits of standardization. Critics also point out that the 1996 HIPAA legislation was formulated in a pre-Internet environment, and therefore was ignorant of important new entities such as Internet and application service providers (Cunningham, 2000). Nonetheless, the use of security and privacy standards, both in the public and private sectors, is expected to boost adoption of electronic communications and improve the efficiency and effectiveness of the health insurance industry.

Lack of Standardization

Although most information demanded by interested parties is already flowing through EDI systems in electronic formats, the health insurance industry (and, indeed, the health industry as a whole) has been relatively slow to adopt Web-based applications on a larger scale. It should be noted that, while a large proportion of demanded information is already in electronic format, many systems housing the data are not interoperable with the Internet. Numerous obstacles, including the nontechnical and five technical considerations discussed above, have been cited for this apparent lag. However, the most serious impediment to the proliferation of Web-based applications is the extreme lack of standardization that characterizes the industry. In addition to the more obvious source of variation across organizations, resulting from the adoption of different applications, mergers and acquisition also attributed to considerable variation within institutions.

Problems associated with a lack of standardization are heightened in large integrated delivery systems (IDS) where communication between member organizations is a key factor in performance. A large IDS may be composed of a number of geographically dispersed member organizations, perhaps acting as independent decision-making units, with a wide array of applications running on different platforms. These may, in turn, be controlled by different operating systems and network protocols. The magnitude of the problem and the potential benefits from a seamless information-sharing arrangement are positively correlated. The industry has recognized this fact and many organizations have accepted the Internet as the most promising platform for achieving the vital objective of seamless integration of diverse information sources.

The lack of standardization in health information systems (e.g., transfer protocols, information content, and response protocols) affects the willingness of individual providers to adopt new platforms for managing administrative tasks. The provider's benefit of a new interface depends on the percentage of her business that will be affected. The larger the number of insurers a given provider has relationships with, the smaller will be the share of total business associated with any individual plan, and, therefore, the smaller the benefits (and therefore the incentive) of adopting a new e-health system sponsored by a given health plan. Standardization of processes into a single interface, allowing providers to communicate with insurers in a consistent manner, will significantly reduce their reluctance to undergo what may constitute a major overhaul of their current operations. In part, legislation specified in the HIPAA in 1996 was aimed at resolving these problems.

HIPAA Administrative Simplification Legislation

The HIPAA was passed by Congress in 1996 to stabilize health coverage for individuals during disruptive events such as changing or losing jobs, pregnancy, moving, or divorce. A less noted provision of HIPAA, but of profound importance within the context of health insurance and e-commerce, is administrative simplification. This provision stipulates the establishment of standards for health information and measures relating to the security, confidentiality, and privacy of that information. Information-sensitive activities that will be affected are health claims processing, enrollment and disenrollment of members, eligibility verification, referral certification and authorization, health plan premium payments, health claims status confirmation, health care payment and remittance advice, and coordination of benefits (Department of Health and Human Services, 2000). Furthermore, HIPAA will, subsequently, require standardization of injury reports and claims attachments. The standards will apply to all private sector and government health plans, health care clearinghouses, and health care providers that elect to make use of affected electronic transactions.

According to the Centers for Medicare and Medicaid Services (CMS), approximately 400 different formats are currently in use for health care claims. The deadline for compliance with the standardization of administrative transactions is currently set for October 2003 (http://www.wedi.org), three years after they were adopted (Department of Health and Human Services, 2000). Standardization of electronic claims and other transactions will allow providers to use the same format for submitting forms to any of the insurers with whom they are affiliated. Providers will also receive electronic transactions, for example referral authorizations, in standardized formats from insurers. As a consequence, providers should find electronic data interchange to be a relatively attractive alternative to paper claims. Transactions conducted over the Internet must also conform to the standards in the area of data content. Some flexibility is allowed to ensure compatibility with online transmission modes.

Primarily insurers, insurance companies, and other payers who will be required to adapt or replace their current systems will feel the initial impact of the HIPAA standardization rules. Much of the administrative simplification provision is aimed at increasing the use of EDI, whether it is Internet based or otherwise. Providers are affected only to the extent that they submit claims electronically. Considering the substantial benefits and financial incentives, the majority of providers are expected to join the effort. For providers who previously were required to maintain multiple formats for doing business with different plans, standardization will make it easier to conduct or adopt EDI. Savings are expected from several sources: reduced labor requirements for managing multiple formats; reduced need for software development and maintenance required to handle multiple formats; elimination or minimization of costly errors resulting from needless duplication and manual data entry; and reduced

Table 1 Ten-Year Projection of Costs and Savings from HIPAA

	10 year total[a]
Costs	
HC provider	3.3
Health plan	3.3
Total	6.8
Savings from claims processing	
HC provider	7.7
Health plan	6.5
Total	14.2
Savings from other transactions	
HC provider	6.2
Health plan	4.9
Total	11.1
Savings from manual transactions	
HC provider	0.2
Health plan	0.1
Total	0.3
Total savings	
HC provider	14.1
Health plan	11.6
Total	25.6
Net	
HC provider	10.8
Health plan	8.3
Total	19.07

Note. From *Health insurance reform: Standards for electronic transactions; announcement of designated standard maintenance organizations; final rule and notice* (Table 4), by Department of Health and Human Resources, August 17, 2000. Reprinted with permission.
[a] Discounted using a 7% rate.

reliance on time-consuming, relatively costly, and uncertain postal services.

The Department of Health and Human Services (2000) estimated the growth in EDI use for physicians, hospitals, and other providers. EDI use by physicians before HIPAA is relatively low compared to the remaining provider groups, implying that the largest gains are expected to be realized in that segment of the industry. The percent growth of EDI use among physicians is assumed to average 22% two years after implementation of the standards. The growth in EDI use by hospitals and nonphysician providers is estimated to average 6 to 7% between 2004 and 2011. The growth in EDI use of this group during the two years immediately following the deadline for compliance is estimated to be between 3 and 6%.

Table 1 shows the Department of Health and Human Services' estimates of cost savings resulting from HIPAA legislation for the next decade. The cost estimates are based on the size and complexity of existing systems, the ability to implement using existing low-cost translator software, and reliance on health care clearinghouses to create standard transactions (Department of Health and Human Services, 2000). Total discounted net savings resulting from the HIPAA-legislated standardization is projected to be $19.7 billion between 2002 and 2011.

The effect of the administrative simplification provision is not limited to the mere implementation of EDI by insurers and providers. Standards for financial and administrative transactions are expected to have a profound impact on health care practice and delivery. A key element of the evidence-based approach to medical practice is effective information management and retrieval, which relies heavily on consistency and uniformity of identifiers, values, and formats of units of data which compose a patient record. HIPAA represents an important step in this direction by establishing uniform rules for the identification and content of more than 400 data fields commonly used in administrative transactions. System-wide standardization initiated by this legislation has the potential to promote more effective medical management by enabling transparency of "best practices" both within and across health care organizations. It is noteworthy that significant practical considerations, particularly those involving patient privacy and confidentiality (see discussion above concerning practical and technical considerations), must be resolved before these benefits can be realized.

Internet Communications Security and HCFA Privacy Act-Protected Information

The potential role of the Internet concerning EDI cannot be ignored. However, the advantages of the Internet are associated with a substantially increased risk to the confidentiality and integrity of information. Recognizing that its design makes it difficult, if not impossible, to eliminate security risks HCFA temporarily prohibited the use of the Internet for the transmission of "HCFA Privacy Act-protected and other sensitive HCFA information by its components and Medicare/Medicaid partners, as well as other entities authorized to use this data." (*WEDI/AFEHCT Internet Security Interoperability Pilot Certification Authority Task Group,* 1999; Workgroup for Electronic Data Interchange [WEDI] and the Association for Electronic Health Care Transactions [AFEHCT]. The HCFA Privacy Act is directly linked to the Privacy Act of 1974. According to Section 5 U.S.C. 552a (e) (10) of that Act, federal systems must "establish appropriate administrative, technical, and physical safeguards to insure the security and confidentiality of records and to protect against any anticipated threats or hazards to their security or integrity which could result in substantial harm, embarrassment, inconvenience, or unfairness to any individual on whom information is maintained.") The WEDI/AFEHCT was charged in 1998 with examining the privacy concerns underlying the Internet and transmission of HCFA Privacy Act-protected information. The resulting policy specifies acceptable methods of encryption to secure confidentiality and integrity of the data and the use of authentication or identification procedures that assure that both the sender and recipient are known to each other and are authorized to receive and decrypt the information. The policy also specifies that organizations that wish to use the Internet to transmit HCFA Privacy Act-protected data have a security plan and comply with HCFA audits designed to verify that all requirements are satisfied.

Non-Internet data transmissions, including local data-at-rest or local host or network systems, are not affected by these additional security requirements. Although organizations are expected to act responsibly with data-at-rest or information stored on local site networks, the extra security requirements associated with transmissions over the Internet are likely to make the latter a relatively less attractive alternative. Indeed, estimates of Internet use involving sensitive information, for example submission of insurance claims, are low (National Research Council, 2000). Large-scale adoption of the Internet for transmission of sensitive information by HCFA and Medicare and Medicaid may act as a signal to private sector entities that the technology is safe and trustworthy. Until then, the private sector may be expected to take a "wait and observe" approach.

E-APPLICATION

The Internet, and, more broadly, information technologies, can influence the health insurance industry in two ways. First, traditional health insurance administrative and customer service tasks are highly adaptable to the Internet. These functions include network development and management, enrollment and eligibility verification, and claims submission and payment. The benefits of a Web-based health benefits management system include reductions in transactions costs, increased velocity of transactions, and increased transparency of the customer service process. The potential for cost reductions from a streamlined Web-based operations process is substantial. According to one estimate, successful implementation of such a system could reduce traditional health insurance-related administrative functions by as much as 70 to 90% (Goldsmith, 2000). The distribution, consulting, and administrative expense of health benefits were estimated at $18 billion annually, with sales and marketing, benefits consultants, and health plan overhead claiming, respectively, $5 billion, $3 billion, and $10 billion (cited in Goldsmith, 2000, from J. P. Lathrop, G. Ahlquist, & D. Knott [2000, Second Quarter], Healthcare's new electronic marketplace, *Strategy+Business, 36*). Net savings from the adoption of e-commerce solutions could be at least $3.6 billion according to a study by Ernst and Young (also cited in Goldsmith, 2000, from P. Kongstvedt [2000, February], 1999 *Managed care benchmarking study,* Washington: Ernst and Young).

A second area where electronic information technologies may play an important role relates to the area of medical management. The contribution of information technologies in this area is difficult to measure, but is, nonetheless, of tremendous interest, particularly to the managed care segment of the industry. Minimization of medical uncertainty and the pursuit of best practices remains an important goal of managed care. Information technology can help realize this goal by enhancing consistency and comprehensiveness of patient records and streamlining the coordination of care.

This section discusses application of the Internet and e-commerce technologies concerning plan selection, benefits management, self-funded insurance, network development, and claims management. Next, it examines the use of the Internet as a tool to educate and guide enrollees. This is followed by a discussion of the role of information technology in medical management.

Plan Selection, Benefits Management, and Self-Funded Insurance

To the extent that choice of health insurance plans is possible, the Internet can act as a powerful tool to facilitate selection. In addition to premium rates, plan selection depends on a variety of quality measurements such as customer service ratings, access to specialists, and the degree of emphasis on preventive care. Employers' interest in these quality measures is clear considering the substantial opportunity costs associated with employee sick days.

Quality information regarding health insurance plans on the Internet ranges from advertising on corporate Web sites to online report cards issued by independent proprietary or not-for-profit organizations (Bates & Gawande, 2000). For example, the National Committee for Quality Assurance (NCQA) offers free online health plan report cards "based on an evaluation of clinical quality, member satisfaction and a comprehensive assessment of key systems and processes" (http://www.ncqa.org). An interactive Web site allows users to compare health insurance plans based on type and location. Another independent organization, HealthScope, created by a coalition of large employers, the Pacific Business Group on Health (PBGH), which used its leverage to require health plans to publish certain quality measurement data, provides customized comparisons for California health insurance plans (Bates & Gawande, 2000). HealthScope provides online comparisons based on health insurance plans' care for staying healthy, care for getting better, care for living with illness, doctor communication and services, and plan service. Plans are given a score in each area, categorizing them as poor, fair, good, or excellent. Each of the five quality measure is derived from a questionnaire. For example, a health plan's score in "care for staying healthy" depends on its activity concerning immunizations, screening for certain types of cancers, prenatal care, and screening for sexually transmitted diseases (for more detail, visit http://www.healthScope.org).

Traditionally, once a corporation selects a group plan, little, if any, potential exists for customization based on personal needs of individuals. The Internet provides an easy and effective method for allowing employees to personalize their health insurance coverage from a large menu of options. Employers' role as the financiers of health insurance coverage, perhaps in the form of a lump sum payment per employee, remains largely unchanged. Whether employers or employees bear the eventual financial burden of health insurance coverage is debatable (Pauly, 1997). For the purposes of this chapter, we focus on the organization that sponsors the health insurance policy. A move away from "one size fits all" to personalized, but financially viable, health insurance plans has important implications for efficiency and cost-effective delivery of medical care by helping patients effectively manage their own health.

Direct competition by insurers for corporate business is another way in which the Internet is influencing the industry. In 1999, Hewitt Associates piloted a program termed a Health Maintenance Organization (HMO) Internet Auction in which more than 50 insurers competed for the business of three large employers (IBM, Morgan Stanley, and Ikonwidth) on a secure Web site. In a series of iterations, plans were allowed to make and modify bids. With the help of unique identifiers, insurers, along with participating employers, were able to view a comparative listing of their bids alongside other measures such as a quality ranking and the level and proposed change of their rates from the previous year.

Hewitt Associates reports that the transparency of the bidding process produced the expected results (Beauregard, 2000). To the satisfaction of the participating employers, the average rate of increase was 2% lower compared to the experience of similar employers in comparable markets. However, two concerns were raised by this pilot test. First, despite security measures, insurers were concerned about placing proprietary data, such as premiums, on the Internet. Second, insurers were worried that employers would make selections based solely on price while ignoring quality and consumer satisfaction.

A potential manifestation of the increased transparency made possible by the Internet is a growth in self-funded health insurance. Many employers, usually those representing large and stable risk groups, take over the risk-bearing function of insurance, and the role of traditional health insurance firms is reduced to providing administrative and consulting support. Four administrative areas in which the Internet can promote this strategy and make a substantial impact are human resource management, medical claims management, medical management, and network development (Goldsmith, 2000). Human resource management functions are often labor intensive and are particularly adaptable to the Internet. Benefits arise from the large potential for cost savings through reductions in labor resources and the possible customization of health insurance plans based on individual needs. Once a plan has been selected, directly by an individual or indirectly through an employer, the Internet can serve as a navigational or educational tool to help members better understand their coverage and assume a more active role in managing their health (see the section below on Navigations and Education).

Network Development

A recent use of the Internet is customization of health coverage for individuals. Internet startups such as Definitely Health (http://www.healthcare.com) and Vivius (http://www.vivius.com) offer individuals a way to tailor a health network to their own unique needs. Participants select a primary care physician, a team of specialists, and a hospital that forms their own personal health network. Among other things, customers are promised greater control over their out-of-pocket costs, employers are promised greater predictability over health care benefits expenditures, and providers are promised control over their own pre-paid fee levels. Vivius advertises elimination of medical claims and referral forms and competitive bidding as a way to contain costs while allowing customers the freedom to choose their own providers. At the same time, providers gain more control over the prices they charge.

This model may serve the self-employed and those who do not have the option of less expensive group coverage. There are sizable obstacles that will have to be overcome before employers and employees will abandon more traditional health insurance models. A move to a do-it-yourself network may imply losing the benefits from large group coverage and special taxation arrangements associated with traditional employment-based health insurance coverage. In addition to employees and the labor unions that represent them, there are several other interest groups that can be expected to oppose a large scale move to the new system, for example, providers who would face capitated risks (Goldsmith, 2000).

Claims Management

Management of insurance claims-related functions is one of the most important administrative tasks performed by the typical provider. Claims management tasks include filing, verification of insurance coverage, determination of co-payment levels, getting approval for referral to a specialist, checking on the status of previously submitted claims, adjudication in case of disputes, and payment of claims. Provider organizations generally submit claims to multiple payers, each using different formats. Substantial labor resources may be required to handle these complicated transactions. Similarly, insurers employ large staffs in call centers to process calls from physicians' offices, plan members, and employers. For example, Regional Physicians Group of Columbus, GA employs more than 20 full-time staff devoted to insurance transactions. By moving to a Web-based system, the group expects to save approximately $0.5 million annually in personnel costs.

Compared to the other claims management functions, submitting is perhaps most amenable to an electronic medium, Internet-based or otherwise. According to estimates from the Department of Health and Human Services (see Table 1), at least 53% of physicians claims will be filed electronically in 2002. In 1999, 64.5% of claims processed by private and public health insurance plans were transmitted electronically (Faulkner & Gray, 2000). Faulkner and Gray (2000) also report that 84% of hospital claims and 89% of pharmacy claims were transmitted electronically in 1999. Hospitals and nonphysician providers have embraced electronic claims filing at a higher rate, reaching, respectively, estimated minimum use rates of 87 and 83%. Medicare and Medicaid, the two largest government-sponsored health plans, and Blue Cross/Blue Shield organizations have been much more supportive of this technology than HMOs and other commercial insurers. In 1999, HMOs and commercial insurers received, respectively, 18 and 45% of their claims electronically, compared to a nation wide total of 64.5% (National Research Council, 2000).

The percentage of electronically filed claims processed through the Internet is not clear. According to an assessment of the National Research Council (2000), relatively few provider organizations and third-party payers have the ability to transmit or receive claims through the Internet. Furthermore, even if this ability is present, security considerations (see Technical and Practical Considerations) make this medium a relatively less attractive alternative. However, a number of developments have taken place, suggesting a dramatically increased use of Web-based applications. In the public sector, Centers for Medicare and Medicaid Services has implemented the Internet-accessible Medicare Data Communications Network (National Research Council, 2000). In the private sector, Internet upstarts and insurers have also started using, or are planning to use, Web-based applications for managing claims-related functions. In an effort to increase its physician subscriber base and shift provider—payer transactions to the Internet, Healtheon/WebMD acquired Envoy Corporation, Medical Manager Corporation, and CareInsite (Danzon & Furukawa, 2001). Envoy Corporation was an electronic health transaction firm that processed approximately one billion claims annually. To challenge these Internet service providers and avoid having to use intermediaries, health insurance companies are actively developing and implementing Web-based applications. The most significant of such challenges is a system called MedUnite, a Web-based transaction system connecting health care providers and insurers (Baldwin, 2002; Danzon & Furukawa, 2001). Seven major insurance companies formed MedUnite, Inc.,: Aetna, Anthem, Cigna, Health Net, Oxford, PacifiCare, and WellPoint Health Networks.

Submitting claims through the Internet is only the tip of the iceberg. The same applications that allow providers to submit claims through the Internet could open the door to the wide variety of related functions, subsequently ballooning its potential to promote efficiency. Verifying a patient's eligibility, in particular, can be integrated. Reduced administrative expense constitutes the most obvious benefit of Internet-based patient eligibility verification to insurers and providers. The accuracy of information retrieved from an Internet-based real-time eligibility verification system can minimize, or even eliminate, delays in payments or denials of claims.

The financial and time burden on providers from eligibility verification using traditional EDI and telephone systems can be imposing. EDI clearinghouse fees range from $0.40 to $0.60 per inquiry. When labor costs are included, telephone eligibility verification for payers ranges from $2.00 to $4.00, while providers face a cost between $10 and $20 per transaction (Bingham, 2001). The cost of reprocessing claims, which were rejected due to ineligibility, is even higher at an estimated $20 to $40 (Bingham, 2001). The potential benefits of accurate real-time eligibility verification are demonstrated in the outpatient sector. In 1997, the Health Care Financing Administration estimated that a staggering 129 million claims are rejected annually as a result of ineligibility (Bingham, 2001). Bingham also provides a detailed analysis of inpatient and outpatient eligibility verification costs per encounter using EDI and telephone systems for managed care plans, indemnity insurers, and providers, and estimates a benefits-to-cost ratio of 6.41 for payers from implementing an Internet-based eligibility verification application. (The author uses an estimate of $0.10 per member per month cost to a software vendor and assumes an insurer covering 195,000 lives, a 40% eligibility verification rate, and estimated hardware and installation costs of $130,000 to derive a total estimated cost of $364,000. The cost savings from eliminating the need for EDI or telephone eligibility verification for 78,000 claims (40% of 195,000) are estimated at $2,331,927. The ratio of these estimated benefits and costs is 6.406.) This ratio does not include secondary benefits, which would accrue to providers, such as time saved, the ability to identify and take care of copayment arrangements at the point of contact, and elimination of bad debt.

In addition to allowing providers to submit claims and determine a patient's eligibility, Web-based real-time communications can significantly expedite referral authorizations. In most cases referral authorizations could be obtained while the patient is still present. In the remaining cases, the request may be transmitted electronically to a medical director for review who then returns it to the requesting office by the same means. In either case, a Web-based system represents a fast improvement over conventional methods, which generally involve lengthy and time-consuming paper forms. This type of real-time and two-way communication between physicians and payers constitutes one of the most important benefits of an Internet-based system. It is noteworthy that many managed care organizations are now reducing or eliminating their prior and referral authorization requirements, therefore, greatly diminishing the value of this function of the Internet.

There are three categories of online communications approaches between payers and physicians (Figure 3). The first consists of separate and independent communication between a physician's office and a single payer. The second approach involves an "electronic intermediary" (e.g., Healtheon/WebMD), which allows physicians to communicate with multiple plans using a standardized format. The third method is similar to the second, but replaces the independent "electronic intermediary" with a "payer portal" (e.g., MedUnite) administered directly by multiple collaborating insurers (Baldwin, 2002; Danzon & Furukawa, 2001). From the physicians' point of view, the latter two methods may be more attractive since they eliminate the need for managing multiple formats and protocols dictated by different plans. Whether a single approach will end up dominating the marketplace is not certain. First, the HIPAA administrative simplification provision was designed specifically to eliminate the difficulties and inefficiencies of managing multiple formats and protocols, thereby lessening some of the distinctions between the approaches. Furthermore, in a competitive marketplace where insurers and intermediaries alike must cater

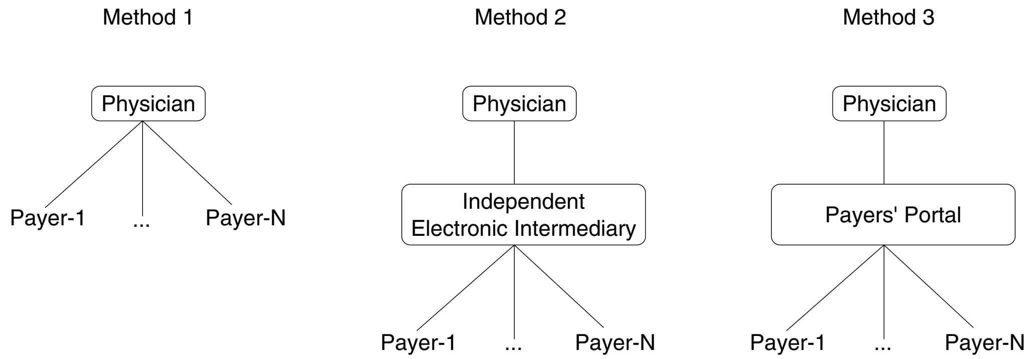


Figure 3: Online communications approaches between physicians and payers.

to a heterogeneous population of health care providers, employees, and employers, a variety of products is likely.

Navigation and Education

A logical start for insurer Web sites was to provide general information without any personal or confidential references. Accordingly, the first wave of third-party payer Web sites featured electronic brochures with general descriptions of the institutions, their products and services, news releases, employment ads, generic health and wellness information, and links to other health information sources. As consumers increasingly turned to the Internet for health information to better manage their own health, payer Web sites became gradually more sophisticated, with some organizations, for example UnitedHealthCare, a business segment of UnitedHealth Group, providing customized pages for members' personal interests. Other customized information found online includes utilization and financial reports, handbooks, directories, and guidelines to customer groups and providers. More interactive features include provider directories based on user-specified parameters such as location, specialty, and language.

Because of large information gaps between patients and providers, concerning appropriateness of diagnoses, treatments, and medical specialties, patients have traditionally assumed a passive role in their own care. (Asymmetries in information are identified as a principal reason for market failure in the health care industry. For an excellent analysis of this topic, see Arrow, 1963.) The complexities of medical care, combined with the laws that govern the delivery of that care, mean that patients will continue to rely on the expertise of professionals to a large extent. However, the Internet has opened the door to a tremendous amount of free information, both general and specific. Information on the Internet is not limited to treatment options, diseases, and diagnoses, but includes provider evaluations or report cards, and information about alternative medicine. By researching their medical conditions before contacting the health care system, consumers are playing an increasingly active role in decisions relevant to their care instead of passively relying on physicians, nurses, and other health care professionals.

Educational Web sites, containing both condition-specific and general educational material, are plentiful (see discussion of secure access under Technical and Practical Considerations). Unfortunately, much of this plentitude can be attributed to a lack of information quality control on the Internet (Winker et al., 2000). To combat medical misinformation on the Internet, the Health on the Net Foundation (HON) was founded in 1995. Although not widely utilized, HON offers their endorsement to those Web sites complying with the HON code of ethics, which emphasizes complementarity, confidentiality, attribution, justifiability, transparency of authorship, transparency of sponsorship, and honesty in advertising and editorial policy (see http://www.hon.ch). A survey of Web sites of 11 top HMOs revealed that none contained the HONCode icon to identify their subscription to the organization (Witherspoon, 2001). Despite the efforts of HON and similar organizations, "caveat emptor" is the recommended approach to information obtained from the Internet.

Web communication may be one-way or interactive. One-way communication serves primarily as a means to disseminate information. For example, this type of communication may be used to dispatch routine reminders for checkups and immunizations, information concerning specific illnesses and diagnoses, or educational information aimed at disease prevention. The Internet distinguishes itself from older technologies because of its ability to serve as a platform for efficient and effective interactive communication. Routine administrative functions, such as tracking medical claims, verification of enrollment, and analysis of benefits coverage, are greatly enhanced by the Internet's capabilities and are becoming increasingly available to plan members. Some insurers have begun using the Internet as a tool to assess members' needs, desires, and inquiries, or as a platform to facilitate dialog between patients and their physicians.

In the survey of 11 top-ranked HMOs mentioned above, Witherspoon (2001) assessed the extent to which these organizations make use of interactive technology on their

Web sites. The HMOs were selected based on their relative standings in rankings by *U.S. News & World Report* and *Newsweek*. Witherspoon rated interactivity of Web sites based on the availability and accessibility of nine functions. The functions include e-mail access to specific physicians or other providers, consumer feedback mechanisms, invitations to subscribe for newsletters and a listserv, ability to register for chat groups moderated by a physician or registered nurse, availability of a provider search engine, and the ability to register for classes via online technology. The following is a summary of some of the author's findings. Of the 11 sites examined, none offered e-mail links to specific health care providers, 6 included a consumer or visitor feedback mechanism, none provided a chat group of any kind, 2 offered subscription to an online newsletter, 5 provided an online provider search mechanism, and 2 allowed members to register online for health classes.

Following Witherspoon (2001), we conducted our own survey to determine how HMO Web sites have developed during the past year. Unfortunately, the sample of organizations could not be duplicated. Two organizations in the original sample either discontinued operating a particular plan or merged with another: the Kaiser Foundation, which ceased its operations in the Northeast, and the Matthew Thorton Health Plan in New Hampshire, which does not operate its own Web site but is incorporated into the regional Blue Cross/Blue Shield. The remaining sample was augmented using the top 40 of the 1999 *U.S. News & World Report* ranking of HMOs and point of service organizations. Of these organizations, several, for example Cigna and the Blue Cross/Blue Shield organizations, used the same Web site. After eliminating duplicates and Web sites that were offline, the sample contained 24 HMOs (see the Appendix). Unlike Witherspoon (2001) we do not rate a Web site's interactivity, but only assess the availability of specific functions. Our analysis suffers from the same shortcoming as the Witherspoon study. Although the sample size is substantially larger, it is, nonetheless, small from a statistical point of view, and only representative of the top HMOs. Furthermore, many of the Web sites analyzed included a section accessible only for members, requiring a log-in password. Some functions we searched for may have been available only through this section.

All 24 sites analyzed for this study, compared to fewer than 50% in the Witherspoon survey, offered a provider search engine. With a single exception, the provider search procedure was interactive, allowing users to specify such parameters as location, specialty, language, and gender of the provider. In addition to a provider search engine, 8 sites (33%) allowed members to change or select a primary care physician. Seven sites (29%) also provided brief physician profiles, offering members information on board certification status, hospital affiliation, medical school, and place of residency. However, none included e-mail links to specific providers. Eighteen of the sites (75%) included a newsletter or other plan-specific publication, and 9 (37.5%) allowed members to register for health classes online. Fourteen sites (58%) allowed members to submit feedback to the health plan through either e-mail or a special form available for download or printing. Similar to the Witherspoon survey, none of the sites invited members to register for a listserv or any kind of chat group.

Three sites (12.5%) offered, or will offer, members the ability to personalize their site experience. Personal Web pages, accessed through the organizations site, may include customized lists of health programs, lists of physicians and other providers, and financial information. Finally, 23 sites (96%) provided links to information supporting healthy lifestyles and preventive care.

Medical Management

The presence of health insurance tends to make consumers insensitive to the cost of medical care, leading to overuse of resources, or the use of unnecessary care with only minor or no added benefits. This phenomenon is sometimes referred to as moral hazard (Arrow, 1963; Pauly, 1968). Partly in response to this, the promotion of efficient delivery of health care remains a cornerstone of managed care. Efficiency suggests the elimination of inappropriate approaches to health care and the determination of the optimal quantity of care without compromising the quality of services provided to patients. Information plays a critical role in this regard. During their training and practice, physicians develop different beliefs concerning the safety and efficacy of treatments and the probability of specific diseases. The resulting variations in practice styles, both spatially and contemporaneously, are well documented in the literature (Wennberg, 1987; Wennberg et al., 1987). To the extent that these differences exist due to physicians' uncertainty and ignorance over best medical practices, the role of information is obvious. The more accurately the most appropriate therapeutic strategy can be defined, the better patients and physicians will be equipped to select an optimal treatment plan. The welfare gains from pursuing evidence-based "best" practices can be substantial (Phelps, 1992, 1995; Phelps & Parente, 1990).

In this regard, the importance of information, seamlessly integrated from all sources, cannot be overstated. Theoretically, information is a key ingredient of efficiency. Most importantly, information promotes efficient choices based on the value and marginal cost of the product. The outcomes-driven approach is still in its infancy and is stifled, at least in part, by the information challenge. The Internet, in its capacity to synthesize and present large amounts of information, is perfectly suited to promote this strategy. Web-based physician-oriented portals are a low-cost method for providing physicians with access to state-of-the-art decision support systems, targeted medical information, and continuing education (Danzon & Furukawa, 2001; Goldsmith, 2000). The tools used to promote outcomes-driven care include utilization review mechanisms. Utilization review can be after the fact, or retrospective, in advance, or prospective, or at the time of treatment, or concurrent. An example that highlights the advantages of the Internet over conventional methods, such as phone calls or mailings, is concurrent utilization review where expediency and immediacy of information are much more important than prospective or retrospective review. Approximately 83% of preferred provider organizations require concurrent utilization review for hospital stays.

Linking evidence-based health outcomes research with the Internet can reduce the fragmentation and lack of communication among caregivers, which is widely cited as a significant source of inefficiency and medical errors in the U.S. health care system. The Internet makes it possible to give all relevant providers, ranging from primary care physicians to specialists and inpatient departments, real-time shared access to patient-specific guidelines and records, in a fast and efficient manner. In the long run, the Internet can promote interaction between providers, actively involving them in patients' care from start to finish. Large integrated delivery systems, in particular, would benefit from such an approach, whether the information integration is accomplished through the Internet or some other platform, such as a private network. Internet technologies represent a cost-effective and logical option to achieve a high level of information integration in a relatively short period of time.

On the demand side of care, consumers are assuming an increasingly active role in the determination and delivery of their care (see the section Navigation and Education). In addition to disseminating information on diseases, diagnoses, treatments, and medical management criteria, insurers are involving members in the process of promoting cost-effective care with incentives, such as cost-sharing arrangements to avoid high-cost or high-risk providers (Goldsmith, 2000). This two-tiered system is used by preferred provider organizations and point of service organizations primarily as an incentive to encourage members to seek care from providers within the network. The Internet, in its capacity to guide consumers through the health care process, provides an excellent medium for insurers to publish their internal data on practice variations to help subscribers identify best practices.

CONCLUSION

Health insurance plays a central role in the health care industry. The health care revolution and the rapid increase in the number of lives covered by insurance created a tremendous demand for increased efficiency. Efforts of the health insurance industry to increase efficiency have focused on basic administrative functions, particularly those involving transactions between insurers and their constituents. The transition from paper-based transactions to electronic transfers was an important step forward, but did not reach stakeholders on a system-wide scale. The subsequent revolution of the Internet introduced exciting new possibilities but was, until recently, largely ignored by the industry. Fortunately, awareness of Internet technology is growing rapidly and will, undoubtedly, play an important and lasting role in achieving long sought after efficiencies. Concerning information technology, health insurance organizations have substantial advantages and incentives compared to other stakeholders, and are therefore expected to play a prominent role in its advancement in the industry.

A key characteristic of the Internet is its ability to generate efficiencies affecting directly, not only the largest stakeholders of the industry, but also individual health plan members and providers. However, there are significant obstacles the health insurance industry must cope with before it can explore the full potential of the Internet. Applications of e-commerce and Internet technologies in the health insurance industry concern communications between insurers, plan enrollees, and providers. The obstacles associated with implementation of these applications depend in large part on the type of information transferred between interested parties. Communication between insurers and plan members primarily concerns general and relatively less sensitive information. Therefore, implementation and proliferation of relevant Internet applications faced few obstacles. In contrast, information transfers between insurers and providers (insurer to provider, provider to insurer, insurer to insurer, and provider to provider) often involve sensitive, or HCFA Privacy Act-protected information, creating a special set of challenges particularly for Internet-based applications. Consequently, the industry has been relatively slow in adopting the Internet as a platform for these types of applications. Although advances in encryption, authentication, and certification make us more comfortable, security and confidentiality transcend technology, and will remain important considerations. Indeed, despite the HIPAA administrative simplification provision, privacy considerations will likely continue to suppress better information integration and standardization, particularly where it concerns the sharing of clinical information among providers and insurers.

In addition, important nontechnical obstacles to more widespread implementation of e-commerce and Internet technologies exist. At the institutional level, the operational risk involved with transforming information exchange technologies is identified as a major obstacle. Furthermore, while Internet use has increased dramatically across the population, substantial demographic inconsistencies remain (U.S. Department of Commerce, 2002), particularly based on age and income status. Apathy toward, and ignorance of, new technologies, both by providers and patients will continue to stifle more widespread adoption of such technologies. It will, therefore, be several years before the Internet's influence on the industry reaches its potential.

On the other hand, the Internet has demonstrated its staying power and its growth is certain. Web-based applications are becoming increasingly interactive, integrating information from a growing number of sources. This will eventually include medical records, making them accessible to providers across the spectrum of health care. What will be the impact on efficiency when all stakeholders are able to communicate effectively?

APPENDIX

Table A-1 HMOs and POSs included in survey.

	Web site function[a]											
Company name	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Fallon Community Health Plan, HMO (MA)	1	0	1	1	0	0	0	0	0	0	0	1
Finger Lakes-Blue Choice, HMO	1	0	1	0	0	0	0	1	0	0	1	0
Harvard Pilgrim Health Care, HMO	1	0	.	1	.	1	1	1	.	.	1	.
Welborn Health Plans	1	0	0	0	0	0	0	0	0	0	0	0
Healthsource Massachusetts	1	0	1	0	0	1	0	1	0	0	1	0
BC/BS HMO Choice Maine	1	0	1	0	0	0	0	1	0	0	1	1
ConnectiCare,	1	0	1	0	0	0	0	1	0	0	1	0
Capital District Physicians' Health Plan	1	0	0	0	0	0	1	1	0	0	1	1
Blue Cross and Blue Shield of Massachusetts	1	0	0	0	0	0	1	1	0	0	1	0
Kaiser Foundation-Southern California	1	0	0	0	0	0	0	1	0	0	1	1
Preferred Care	1	0	1	0	0	0	0	1	0	0	1	1
Group Health Cooperative of South Central Wisconsin	1	0	0	0	0	0	0	0	0	0	1	1
Health New England	1	0	0	0	0	0	0	0	0	0	1	0
HealthAmerica of Pennsylvania	1	0	0	0	0	0	1	1	0	0	1	0
Health Alliance Medical Plans	1	0	0	0	0	0	0	0	0	0	1	0
Independent Health-Western New York	1	0	0	0	0	0	1	1	0	0	1	1
Group Health Cooperative of Eau Claire	1	0	0	0	0	0	0	0	0	0	1	1
Physicians Plus Insurance, HMO (WI) 83	1	0	1	0	0	0	0	1	0	0	1	1
HealthPlus of Michigan, HMO	1	0	0	0	0	0	0	1	0	0	1	0
Priority Health—Michigan	1	0	0	1	0	1	1	1	0	0	1	1
Kaiser Foundation—Ohio	1	0	0	0	0	0	0	1	0	0	1	1
Kaiser Permanente—Colorado	1	0	1	1	0	0	1	1	0	0	1	1
Kaiser Permanente—Hawaii	1	0	0	1	0	0	1	1	0	0	1	1
Kaiser Permanent—Mid-Atlantic	1	0	0	1	0	0	1	1	0	0	1	1
	24	**0**	**8**	**7**	**0**	**3**	**9**	**18**	**0**	**0**	**23**	**14**
Percent	1	0	33.3	29.2	0	12.5	37.5	75	0	0	96	58

[a] See Table A-2 for an explanation of the Web site function. A value of "1" indicates that the function was present on the Web site. A value of "0" implies that the function could not be located on the Web site. A period means that the availability of the option was unclear or links were indicated on the main Web site but could not be accessed at the time of the survey.

Table A-2 Web Site Functions Examined in Survey.

Column	Description of function
A	Asearch engine for providers
B	E-mail links to specific physicians
C	An option for selecting or changing a primary care provider
D	Physician profile
E	E-mail links to nonphysician providers
F	Ability to create and maintain a personalized Web page
G	Option to register for classes via the Internet
H	An online newsletter or other plan-specific publications
I	Option to register for an online listserv
J	Option to register for an online chat group moderated by a physician or registered nurse
K	Links to information supporting a healthy lifestyle and preventative care
L	A mechanism, other than a postal address, for consumer feedback. This includes either an e-mail address or an online form.

ACKNOWLEDGMENTS

The author thanks Jodi B. Pracht for research support and valuable comments that were instrumental in shaping this paper.

GLOSSARY

Availability The continuous accessibility of a network, the applications of which it is composed, and the services it offers.

Bandwidth The rate, measured in bits per second, at which information is transmitted through a network.

EDI Electronic data interchange.

CMS Centers for Medicare and Medicaid Services (formerly, the Health Care Financing Administration).

HCFA Health Care Financing Administration (see CMS).

HHS U.S. Department of Health and Human Services.

HMO Health Maintenance Organization.

HIPAA Health Insurance Portability and Accountability Act.

HON Health on the Net Foundation.

IDS Integrated delivery system.

POS Point of service.

IVR Interactive voice response.

IT Information technology.

Latency The time required for transmitting data across a network.

NCQA National Committee for Quality Assurance

PPO Preferred Provider Organization

Security The availability, confidentiality, and integrity of information.

Ubiquity The relative accessibility of a network.

WEDI/AFEHCT Workgroup for Electronic Data Interchange/Association for Electronic Health Care Transactions.

CROSS REFERENCES

See *Electronic Data Interchange (EDI); Health Issues; Internet Literacy; Medical Care Delivery.*

REFERENCES

Arrow, K. J. (1963, December). Uncertainty and the welfare economics of medical care. *American Economic Review, 53*(5), 941–969.

Bacus, R., & Zunke, R. (2001). Achieving Internet-based efficiencies in a rural IDS: A case study. *Healthcare Financing Management 55*(9), 68–71.

Baldwin, G. (2002, April). Improving communications with payers. *Technology in Practice.* Retrieved February 11, 2002, from http://www.technologyinpractice.com

Bates, D. W., & Gawande, A. A. (2000, November/December). The impact of the Internet on quality measurement. *Health Affairs, 19*(6), 104–114.

Beauregard, T. R. (2000). Linking Internet strategy to high-quality, cost-effective managed care. *Benefits Quarterly, 16*(2), 15–19.

Bingham, A. (2001, February). Internet-based eligibility verification lowers costs, improves payment timeliness. *Healthcare Financial Management, 33*(2), 47–49.

Centers for Medicare & Medicaid Services. (2001). *Table 1: National health expenditures and selected economic indicators, levels and average annual percent change: Selected calendar years 1980–2011.* Retrieved October 29, 2002, from http://cms.hhs.gov/statistics/nhe/projections-2001/t1.asp

Cunningham, B. (2000, November/December). Old before its rime: HIPAA and e-health policy. *Health Affairs, 19*(6), 231–238.

Danzon, P. M., & Furukawa, M. F. (2001). Health care: Competition and productivity. In R. E. Litan, et al. (Eds.), *The economic payoff from the Internet revolution* (pp. 189–234). Washington, DC: Internet Policy Institute, Brookings Institution Press.

Department of Health and Human Services, Office of the Secretary Health Care Financing Administration (2000, August 17). *Health insurance reform: Standards for electronic transactions; announcement of designated standard maintenance organizations; final rule and notice* (45 CFR Parts 160 and 162).

Dorenfest, S. (2000, August). The decade of the '90s: Poor use of IT investment contributes to the growing healthcare crisis. *Health Informatics, 17,* 64–67.

Faulkner & Gray. (2000). *Health data directory 2000.* New York: Faulkner & Gray.

Fox, S., & Raine, L. (2000). *The online health care revolution: How the Web helps Americans take better care of themselves.* Washington, DC: Pew Charitable Trusts.

Fuchs, V. R. (1996). Economics, values, and health care reform. *American Economic Review, 86*(1), 1–24.

Gerberick, D., Huber, D., & Rudloff, R. (1999, Summer). Defining a technical architecture for health care security. *Information Security Systems, 8*(2), 37–49.

Goldsmith, J. (2000, November/December). The Internet and managed care: A new wave of innovation. *Health Affairs, 19*(6), 42–56.

Hofflander, J. (1999, July). Managing healthcare administration in the information age. *Health Management Technology, 20*(6), 34–37. Retrieved February 11, 2002, from http://www.healthmgttech.com/cgi-bin/arttop.asp?Page=mancare0799.html

Hurley, R. E., Freund, D. A., & Paul, J. E. (1993). *Managed care in Medicaid. Lessons for policy and program design.* Ann Arbor, MI: Health Administration Press.

Korpman, R. A. (2001, February). Managed care and e-health. *Health Management Technology, 22*(2), 12–17. Retrieved March 5, 2002, from http://www.healthmgttech.com/cgi-bin/arttop.asp?Page=h0201 managed.htm

Luft, H. S. (1978). How do health-maintenance organizations achieve their "savings." *New England Journal of Medicine, 298*(24), 1336–1343.

Lumpkin, J. R. (2000). E-Health, HIPAA, and beyond. *Health Affairs, 19*(6), 149–151.

Manning, W. G., Leibowitz, A., Goldberg, G. A., Rogers, W. H., & Newhouse. J. P. (1984, June 7). A controlled trial of the effect of a prepaid group practice on use of services. *New England Journal of Medicine,* 1505–1510.

National Research Council, Computer Science and Telecommunications Board (2000). *Networking health:*

Prescriptions for the Internet. Washington, DC: National Academy Press.

Newhouse, J., Schwartz, W. B., Williams, A. P., & Witsberger, C. (1985). Are fee-for-service costs increasing faster than HMO costs? *Medical Care 23*(8), 961.

Pauly, M. V. (1968, June). The economics of moral hazard: Comment. *American Economic Review, 58,* 531–537.

Pauly, M. V. (1997). *Health benefits at work.* Ann Arbor, MI: University of Michigan Press.

Pauly, M. V., Hillman, A. L., & Kerstein, J. (1990). Managing physician incentives in managed care. *Medical Care 28*(11), 1013–1023.

Phelps, C. E. (1992, Summer). Diffusion of information in medical care. *Journal of Economic Perspectives, 6,* 23–42.

Phelps, C. E. (1995, June). Welfare loss from variations. *Journal of Health Economics, 14*(2), 253–260.

Phelps, C. E., & Parente, S. T. (1990, August). Priority setting in medical technology and medical practice assessment. *Medical Care 28*(8), 703–723.

Raffel, M. W., & Barsukiewicz, C. K. (2002). *The U.S. health system: Origins and functions.* Albany, NY: Delmar.

Runy, L. A. (2000, December). Consumers in control: How the Net will reshape health care. *Hospitals and Health Networks, 74*(12), 4–7.

Schaich, R. L. (1998, June). Internet commerce and managed care. *Health Management Technology, 19*(7), 43–46. Retrieved February 11, 2002, from http://www.healthmgttech.com

Taylor, H. (1999, August 5). Explosive growth of "cyberchondriacs" continues. *The Harris Poll #47.* Retrieved February 11, 2002, from http://www.harrisinteractive.com/harris_poll/index.asp?PID=117

U.S. Census Bureau (2000). *County business patterns.* Retrieved March 7, 2002, from http://www.census.gov/epcd/cbp/view/cbpview.html

U.S. Census Bureau (2002, September). *Health insurance coverage 2001: Current population reports.* Retrieved February 11, 2002, from http://www.census.gov/hhes/www/hlthins.html

U.S. Department of Commerce, National Telecommunications and Information Administration: Economics and Statistics Administration (2002). *A nation online: How Americans are expanding their use of the Internet.* Washington, DC: Author.

WEDI/AFEHCT Internet Security Interoperability Pilot Certification Authority Task Group (1999, December 10). Retrieved October 11, 2002, from http://www.wedi.org

Weisbrod, B. A. (1991, June). The health care quadrilemma: An essay on technological change, insurance, quality of care, and cost containment. *Journal of Economic Literature, 29,* 523–552.

Wennberg, J. E. (1987). Population illness rates do not explain population hospitalization rates. *Medical Care, 25,* 354–359.

Wennberg, J. E., Freeman, J. L., & Culp, W. J. (1987). *Are hospital services rationed in New Haven or overutilized in Boston? Lancet, 1,* 1185–1189.

Winker, M. A., Flanagin, A., Chi-Lum, B., et al. (2000). Guidelines for medical and health information sites on the Internet: Principles governing AMA Web sites. *Journal of the American Medical Association, 283*(12), 1600–1606.

Witherspoon, E. (2001). A pound of cure: A content analysis of health information on Web sites of top-ranked HMOs. In R. E. Rice & J. E. Katz (Eds.), *The Internet and health communication: Experiences and expectations* (pp. 189–212). London: Sage.

Health Issues

David Lukoff, *Saybrook Graduate School and Research Center*
Jayne Gackenbach, *Athabasca University, Canada*

INTRODUCTION

An oft-repeated phrase from the best-selling *Being Digital* (Negroponte, 1995) is that computing "is not about computers any more. It is about living" (p. 6). This claim captures the pervasiveness with which our lives have become intertwined with computer-mediated communication. The question of the Internet's positive and negative effects has increasingly become a focus of public discussion. Media scholars point out that such concerns have a long history, however. The earliest forms of mediated communication (meaning non–face-to-face) date from prehistoric human art and symbols. Plato wanted to ban written poetry because he feared it might corrupt even the best individuals. Pythagoras forbid reading in his school out of fear that reliance on written communication would diminish mental capacities. When the telephone and television were introduced, there were also fears of loss of privacy and the homogenization of culture as well as concerns regarding indecent communication.

Yet mediated communication has not only prevailed, each new form has been adopted more rapidly and pervasively. The amazing growth of the Internet has far exceeded all other mass communication media and has implications for health and for the work of health professionals.

PSYCHOLOGICAL HEALTH ISSUES

Definitions of Self

In *Life on the Screen: Identity in the Age of the Internet,* Turkle (1995) discusses how online experiences can dramatically help or hinder the self through experimentation with multiple identities: "The Internet has become a significant social laboratory for experimenting with the constructions and reconstructions of self that characterize postmodern life" (p. 180). By interacting anonymously online, some people have found unparalleled opportunities to explore self, obtaining experiences difficult to find in real life—from pretending to be a member of the opposite sex to slaying a dragon. The Internet offers a rich place to act out or work through psychological issues. When used by someone who has a fragmented self in face-to-face reality, however, this can lead to problems, such as college students who spend many hours playing online games, leading to academic failure and even psychotic episodes. (Such dysfunctions are discussed further in the section on Internet addiction).

People whose lives might be changed the most by the Internet are those who have some antisocial aspect to their identity, such as extreme views or practices considered deviant by others in their social network. Those who initially communicate a taboo identity are subsequently more likely to both incorporate that identity into their sense of self and to disclose it to face-to-face friends. Thus, as often claimed, the Internet can be a kind of "social laboratory" in which people test their identities on strangers before embracing them and sharing them with friends and family (Tyler, 2002). Yet in some contexts, Internet use promotes behavior that is more normative and socially influenced. Social influence, rather than being filtered out, can be even stronger in computer-mediated communication compared to face-to-face (Spears, Postmes, Lea, & Wolbert, 2002).

Humanistic psychologists have warned that as humans become more dependent on technology and inhabit an increasingly technological world, "we too are subtly made-over and reinvented by the very technological world we invent" (Dodson, 2002, p. 17). As people try out multiple new identities online and experiment with virtual reality simulations of reality, the self becomes more mutable and fragmentary than it has ever been in human history. "To value machines and virtual reality over our humanity is a choice that will clash head-on with 500,000 years of biological wisdom. . . . The question now is: how are the relationships we are forming with computers changing what it means to be human?" (Heckler, 2001, p. 278).

One example of this impact is how readily the computer metaphor for the brain has been adopted both in the health professions and in popular culture. The choice of linguistic metaphor has profound implications for how that phenomenon is viewed and interacted with (Lakoff & Johnson, 1983). In the case of health issues,

> Instead of confronting how we are creating and reacting to the stresses that cause human conflict and sickness, we look only to how we can maintain and repair the computer-brain. This translates into servicing the "human machine" with drugs; drugs for sleeping, drugs for waking up, drugs for shifting moods, drugs for recreation, for upset stomachs, for having sex . . . and so on. As if taking ourselves in for a lube job and oil change will fix the problem. Our acquiescence to technology has lowered us to its level, rather than pulling it up to our level. (Heckler, 2001)

Even technology's critics are also searching for ways that it can lead to enhanced connections with others and a deeper sense of self, however.

Consciousness

The Internet is potentially a vehicle for expanding consciousness as it expands human contact with "altered realities." The sharing and communication capacities of the Internet are creating what has been termed an *electronic neurosphere,* a global mind that will change health care, theology, and the way we think about ourselves. The interface of human brains and computers will extend both human and machine intelligence and lead to changes in consciousness. With increased practice in absorption (deep experiential immersion characteristic of online communication) in altered constructed realities online, humans average experience of "altered reality" will expand (Gackenbach, 1998).

Social Interactions

An early and well-publicized study of the psychological effects of the Internet reported that heavy Internet use was associated with lower social involvement, less family communication, and depressed mood (Kraut et al., 1998). A 3-year follow-up by the same researchers, however, found the subjects no longer showed these effects (Kraut et al., 2002). As users became more experienced with the Internet and their competence improves, they become more skilled at getting social support from others online. Internet use leads to positive personal and social effects such as an increase in communication, social involvement, and psychological well-being, factors associated with improved—not reduced—mental health.

Research into computer use by the elderly also reveals its positive social potential. A study of a computer instructional program on the quality of life of senior citizens found that at first participants showed reluctance but then enthusiasm about using the computer as they progressed in functional skills. Eventually, participants showed improvements in level of independence, more interest in other activities, a shift in mood toward a more positive temperament, and transfer of physical skills to other areas (Groves & Slack, 1994).

CHESS is an online computer-based patient support system that targeted elderly women with breast cancer. In a study by Gustafson et al. (2001), at 2-month follow-up, the CHESS group was significantly more competent at seeking information, more comfortable participating in care, and had greater confidence in doctor(s). At 5-month follow-up, the CHESS group had significantly better social support and also greater information competence. Obtaining social support for health challenges is one of the major and oldest uses of the Internet. Newsgroups, which are online asynchronous threaded discussions, predate the World Wide Web by more than 20 years, and even though Web-based support groups have made it even easier to communicate online, newsgroups are still actively used to obtain social support.

As people become more mobile, the Internet also becomes more important to maintain contact with family and friends. Internet usage by people who have relocated jobs has been found to ease their adjustment and also provide a means for enhancing business and personal relationships (Haupt, 1998). In general, the Internet has had positive social consequences in people's everyday lives because it increases the frequency and quality of interpersonal communications. Early fears that Internet use would be associated with social isolation and dependence seem unsupported by more recent findings.

Internet Relationships

Disinhibition refers to a lack of interpersonal feedback characteristic of face-to-face communication. This can lead to enhanced self-disclosure and also personal attacks as many people feel freer to express themselves during online communication. Three potential causes of disinhibited behavior online are as follows:

1. **Deindividuation—**reduction in accountability cues that leads to decreased self-regulation
2. **Reduced Social Cues—**the limited nonverbal cues to communicate meaning
3. **Social Presence—**a lowered sense of being there and accountable for one's actions

Online relationships, contrary to early beliefs, can become close, meaningful, and lasting (McKenna, Green, & Gleason, 2002). People often bring relationships formed on the Internet into the real world. Studies show that female Internet users had more online contact than their male counterparts, with more contacts for romantic relationships than for friendships. Further studies found that relationships formed initially online were as stable after 2 years as those formed face-to-face. It appears that many of the concerns expressed about the quality of Internet interactions are unfounded.

Family Life

The Internet can contribute to dysfunction when family life is disrupted by technologically captured moments, such as always-on instant messaging, telemarketing phone calls during dinner, and unattended television and radio creating background noise. Dysfunctional use of the Internet by children as well as adults can result in diminished participation in the family with members becoming isolated, each person wrapped in his or her own "techno-cocoon."

At the end of the day, Mom is preparing dinner while checking the answer machine, head glued to the portable phone while she returns calls. One child is playing games on the computer in his bedroom, another is talking on her own phone, and the youngest is playing Nintendo. Dad comes home later from work and goes immediately to the computer (Weil & Rosen, n.d.).

When the children become the family computer and Internet expert, parents, although proud, may also be disconcerted by the shift in authority in the family structure (Kiesler, Bozena, Lundmark, & Kraut, 2000).

Childhood and Adolescence

The gaming industry that began with video games and now includes online gaming grosses more per year than the movie industry. It is a force that all parents and health workers need to contend with in their interactions with children. Emes (1997), in a review of the research on the effects of video games on children's well-being, concluded as follows:

> Results show that playing video games is associated with a variety of physical effects including increased metabolic and heart rate, seizures, and tendonitis. Aggressive behavior may result from playing video games, especially among younger children. There is no direct relationship between psychopathology or academic performance and playing video games. Overall findings indicate that although video games have some adverse effects, they are also valuable learning tools. (p. 409)

Game playing improves dynamic spatial ability, which is linked to measures of reaction time and speed of mental processing, which in turn are thought to be linked with general measures of intellectual ability (Jackson, Vernon, & Jackson, 1993). Video game play increases choice reaction time, spatial skills, scientific problem solving skills, and intelligence. Interactive media use also increases parallel processing abilities or multitasking (Greenfield, Brannon, & Lohr, 1996). Furthermore, there is no research data supporting the popular notion of loss of attention span associated with video game and Internet activities. Multitasking may have the negative consequence of losing depth of information processing while gaining breadth (Greenfield, 2000).

Despite these positive effects, exposing infants and young children to video games and the Internet often does not engage their major learning style, which is through their sensorimotor system, and thus may have negative effects on their cognitive development.

Access to pornography by minors is a major concern. Adult-oriented sites make up approximately 2% of all content on the World Wide Web. Although sexually explicit material comprises only a small fraction of online content, that fraction is highly visible and accounts for a significant amount of Web traffic. Although such material exists in magazines, television, movies, and billboards, pornography on the Web is much more accessible to children. Graphic images can find their way onto a child's computer screen without being actively sought. In addition, the anonymity of the Internet makes it easier for strangers to interact with children. Exposure to sexually violent material has been found to be related to attitudinal changes, with both men and women more likely to display callous attitudes toward female victims, such as stating that a rape was the fault of the victim or that she brought it on herself. Exposure to nonviolent sexual material increased the acceptability of premarital sex for teenagers compared with teenagers who were not similarly exposed (Huston, Wartella, & Donnerstein, 1998; Lawrence & Giles, 1999).

Most of the concern has focused on the adult online entertainment industry, which generates about $1 billion a year. According to the Nielsen/Net Ratings, nearly 16% of visitors to adult-oriented sites in February 2002 were under the age of 18. No studies have been able to tease apart the effects of Internet versus television versus print media exposure to sexually explicit material; however, the concern remains that the Internet provides increased access to such material, even when it is not being sought, such as when searches return sites with pornographic content unrelated or peripherally related to the search terms (e.g., fairies, cartoons). Some schools and libraries use filtering software to exclude such sites, and home computer versions of filtering software are also available (Thornburgh & Lin, 2002).

Socialization Effects

The influence of exposure to violence in games has been an area of concern. Most teenagers own their own video or computer games and spend 5 to 13 hours a week playing them. Additionally, nearly half of eighth and ninth graders report that they want more violence in their games. At-risk boys spend more time playing games than their peers, and they play more violent games. Youth who report an increased appetite for violence in video games are more likely to have gotten into physical fights in the previous year. Thus, there is reason to be concerned about exposure to violence in online games.

Most games have no female roles, and the ones that do typically involve damsels in distress. Because of the social element of Internet communications (e.g., instant messaging and chat), girls are quickly catching up with boys in their computer skills. Most adolescents use the Internet like a telephone rather than as an opportunity to try different personalities and identities. Girls are heavier users of chat rooms, whereas boys are heavier users of online games, but for both instant messaging was the most used element of online activities. Thus, Internet use is similar to traditional means of youth social interaction (Gross, Juvonen, & Gable, 2002).

Internet Addiction

Dysfunctional Internet use patterns are often referred to as Internet addiction, although there is no psychiatric diagnosis called Internet addiction listed in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM; American Psychiatric Association, 1994). To become an official diagnosis, research must show that Internet addiction disorder can be reliably diagnosed and that the diagnosis correlates with outcome, treatment results, histories, and prognosis.

Although some argue that the Internet is merely a communications medium and cannot be compared with addictive stimuli such as drugs and alcohol, others point to the growing number of cases related to overuse of the Internet seen in clinics and private practices. It seems closer to pathological gambling, an impulse disorder in the 4th edition of the *DSM* rather than a type of substance-related disorder. Studies are in fact establishing a connection between problematic Internet use and impulse disorder. For example, a study of 20 people with problematic Internet use found that all of them met diagnostic criteria for an impulse control disorder (Shapira, Goldsmith, Keck, Khosla, & McElroy, 2000).

Dysfunctional Internet usage occurs when people spend too much time online to the detriment of their social and financial well-being. Several types of problematic online behaviors have been documented, such as online gambling, online shopping, day trading, cybersex affairs, online gaming, violent computer games played by children, and participation in chat rooms. One study found that 6% of the 17,251 persons surveyed met criteria for compulsive Internet use (Greenfield, 1999). One of the earliest articles illustrating dysfunctional use of the Internet was the case study of a 43-year-old woman whose preoccupation with chat rooms eventually destroyed her 17-year marriage (Young, 1996).

Problems can also originate in the workplace. People whose jobs require extensive computer use tend to have pervasive and characteristic cognitive styles characterized by "multitasking" with high-speed processing. They can develop problems including loss of mid- and long-term goal directedness, diminished length of attention span, disrupted patterns of living (e.g., sleeping, eating), and detached or disturbed social relationships, often using the computer as the focal point for all contact with the world. (Fenichel, n.d.).

Assessment of Internet Addiction

Psychologist Kimberly S. Young (1998) conducted a 3-year study of Internet addiction that led her to propose a definition that involves having experienced four or more of the following symptoms during the past year:

1. Do you feel preoccupied with the Internet or online services and think about it while offline?
2. Do you feel a need to spend more and more time on line to achieve satisfaction?
3. Are you unable to control your online use?
4. Do you feel restless or irritable when attempting to cut down or stop your online use?
5. Do you go online to escape problems or relieve feelings such as helplessness, guilt, anxiety, or depression?
6. Do you lie to family members or friends to conceal how often and how long you stay online?
7. Do you risk the loss of a significant relationship, job, or educational or career opportunity?
8. Do you keep returning even after spending too much money on online fees?
9. Do you go through withdrawal when offline, such as increased depression, moodiness, or irritability?
10. Do you stay online longer than you originally intended?

Dr. Young (n.d.) has created interactive online tests for Internet addiction based on these criteria, although reliability and validity data are not reported. Dr. Young has created tests for specific Internet-related problems, including Cybersexual Addiction, Obsessive Online Traders, Compulsive Online Gambling, Online Auction Addiction, and A Parent's Assessment of Child's Internet Addiction.

No accepted guidelines for distinguishing "normal" enthusiasm from "pathological" preoccupation with the Internet exist as of yet. (Why aren't people who spend a lot of time reading called "book addicts" instead of being affectionately called "bookworms"?) It is clear, however, that people have lost their jobs, flunked out of school, and been divorced because Internet use consumed all their time. Such extreme cases demonstrate the need to pay attention to Internet use patterns. More research is needed in the area of Internet addiction to determine reliable and valid criteria.

Treatment

Suler (1999) proposed a treatment approach based on the observation that people become "addicted" to the Internet when they have disconnected from their face-to-face life. Their online activity becomes a world unto itself, which they don't talk about with the people in their face-to-face life. It can then become a walled-off substitute or escape from their life.

Because Internet addiction, as with all others types of addiction, entails isolation from others and prioritizing of the compulsive activity over all other aspects of life, Suler proposes that treatment be targeted to integrate online and offline worlds by

- Telling online companions about one's offline life,
- Telling offline companions about one's online life,
- Meeting online companions in person,
- Meeting offline companions online,
- Bringing online behavior offline, and
- Bringing offline behavior online.

Young, who founded the Center for Online Addiction, leads an online self-help group for Internet addictions. Although some would compare this to holding an Alcoholics Anonymous meeting in a bar, she has been able to engage people in reevaluating and altering their patterns of Internet usage using this approach. She believes that, as with other addictions, it is first necessary for the person to break through denial. She reports that Internet addiction does not require abstinence for a healthy and life-enhancing recovery. Her treatment approach focuses on finding a balance between Internet use and other life activities. The treatment model is similar for eating disorders or controlled drinking programs. The focus is to identify the triggers of binge behavior and relearn how to use the behavior in moderation (Young, n.d.).

Technostress

Technostress refers to any negative impact on attitudes, thoughts, behaviors, or body physiology that is caused either directly or indirectly by technology. One well-documented form of technostress is the escalating problem of information overload. Just as fat has replaced starvation as this nation's number one dietary concern, information overload has replaced information scarcity as an important new emotional, social, and political problem. Until recently, the production, distribution, and processing of information remained evenly balanced. People could receive and think about information at roughly the same pace it was generated. Since the mid-20th century, computers, television, satellites, and the Internet have created a condition of hyper-production and hyper-distribution that has surpassed human processing abilities (Shenk, 1998).

The impact of information overload is particularly apparent in the workplace as more and more people spend their time at work sorting through e-mail, voice messages, and Web pages while their day is interrupted by ringing phones, dinging e-mails, and squealing fax machines. A study of employees at Fortune 1000 companies found that they averaged 178 messages a day and three interruptions an hour (Rosen & Weil, 1997). A typical day can look like this:

> You walk in the door at 8:02 a.m. to be greeted by your secretary. She hands you a stack of 18 "While You Were Out" slips that she has transcribed from your voicemail. With your overstuffed briefcase in one hand, your laptop in the other, and the stack of folders under your arm. You consider trying to squeeze the messages between two available fingers. "Oh, wait," she says, and points to three Federal Express packages. No way to carry it all now, so you head off to unload the work you brought back from home.
>
> Your desk looks like a war zone—and your side lost. A quick glance at your computer reminds you to check your e-mail. What? 26 messages! How can that be? You checked it before you went to bed last night and spent 45 minutes answering all the messages. As you read your e-mail, you flip through yesterday's mail. The phone rings, and you half-listen to one of your employees as you scroll through your e-mail to decide what to answer now and what can wait until later.
>
> Your day has just begun and already you are exhausted. You feel like an octopus, with your arms and brain moving in multiple directions at the same time. By the time you have finished your work day—technically eight hours later (ha!)—you will have started and stopped dozens of tasks. The phone, the fax, the beep of an incoming e-mail all wrench your mind from what you are doing and thinking. When you finally return to your work, often you have lost your train of thought. (Rosen & Weil, 1998)

Symptoms of Technostress

Technostress can affect memory, leading people to lose track of what they wanted to do or say. Getting a peaceful night's sleep becomes difficult when overstimulated minds buzz and chatter, and enjoying laid back recreational activities is disrupted by preoccupation with to-do lists, calls, errands, memos, and more. Headaches, irritability, gastrointestinal discomfort, heart problems, and hypertension can also be related to technostress. In addition, information technologies allow people to work 5 p.m. to 9 a.m. as well as normal daytime hours from 8 to 5. Rosen, who has conducted research on human–computer interaction in the workplace, maintains, "Our brains aren't wired to 'multitask' the way our computers are. We're testing the limits of our human abilities" (Rosen & Weil, 2000).

The scientific understanding of the body's response to this assault on the mind is still in its infancy, but it draws on Selye's (1978) classic work on stress. Technostressed individuals not only live on the edge of constant crisis, some seem to prefer life that way. It has been hypothesized that they are addicted to their own adrenaline.

Coping With Technostress

People need to learn new ways to cope with the constant demand to learn new skills, meet speedier turnaround times, and be accessible 24 hours a day. The pace of technological innovation and intrusion into lives is unprecedented and requires a radical rethinking of how people relate to technology. Yet managed appropriately, technology can enhance both the quality and efficiency of everyday life.

Weil and Rosen (1997) offer some tips to help people manage the information flow:

Sift and Trash. Try to focus on the information you really need instead of news blips that distract. Think critically and separate the gems from the dross.

Set Limits. Ration the time you spend watching television, listening to the radio, and cruising the Internet. Designate the best times for people to call or fax you.

Respond on Your Own Time. Disable the e-mail ding and turn off the ringer on the fax machine. You can respond after you've finished other tasks.

Relax When Technology Makes You Wait. Instead of getting irritated while your e-mail boots or a company's telephone system puts you on hold, use that time to rest or tend to small tasks.

Use the Technologies That Work for You. You don't have to acquire every new technology. If beepers and cell phones cause you stress, stick with voicemail.

Schedule Time Away From Information. Set aside slots for exercise, sports, dinner with friends, and family vacations.

Technophobia

Surveys show that 85% of the U.S. population feels uncomfortable with technology. Only about 10–15% of the industrialized population are "eager adopters." Hesitant "prove its" make up about 50–60% of the population, and resisters 30% or so (Ellen, Beardon, & Sharma, 1991). A survey by Dell Computer identified 55% of Americans as technophobic (Dell Computer Corporation, 1993). The percentage of the population who are becoming hesitant and resistant is increasing. Over a 3-year period, clerical workers became more hesitant and resistant toward technology while managers and executives became more resistant (Rosen & Weil, 2000).

Technophobia can have detrimental effects on careers because workers who use computers on their jobs earn 10–15% more than those who do not even after holding education, income, occupation, and other characteristics constant. Thus, avoiding computers can be costly. People who don't feel comfortable with technology often feel inferior and intimidated, as though their boundaries are being invaded. Studies have shown that the key psychological factor determining resistance to technological change is the person's perceived ability to use a product successfully (which can be improved with training).

PHYSICAL HEALTH ISSUES

Repetitive Strain Injuries

Repetitive strain injuries (RSIs) are a response to excessive and repetitive demands placed on the body. RSIs all have a similar cause: excessive wear and tear on the soft tissues of the body (tendons, nerves, circulatory system, etc.) related to performing the same task over and over again, from clicking a mouse to straining to see the computer monitor. Disorders related to work that requires repetitive motion are increasing. They account for nearly half of all reported work-related illness, and carpal tunnel syndrome is estimated to account for more than 41% of these repetitive motion disorders. Researchers have defined six key risk factors in the workplace for the development of these disorders, including carpal tunnel syndrome (CTS): repetition, high force, awkward joint posture, direct pressure, vibration, and prolonged constrained posture. Experts estimate that between 7% and 16% of the population experience CTS. The incidence appears to be increasing. How much is directly related to computer use has not been established, but typing is considered a major source of RSIs (Szabo, 1998).

Other Health Problems

Computer vision syndrome is defined by the American Optometric Association the "complex of eye and vision problems related to near work, which are experienced during or related to computer use including eyestrain, blurred vision, dry and irritated eyes, slow refocusing, sensitivity to light, double vision, and color distortion.

Prevention

Experts in the field believe that a sound ergonomics program can resolve the majority of health issues associated with computer use. Linden (1995), a specialist in body and movement awareness education, advises, "The first and most basic element of computer safety is body awareness. We must be able to feel and understand the functioning of each body component, from our legs to our eyes, so we can detect strain and nip it in the bud" (p. 3).

Second, Linden prescribes proper movement breaks. Bodies are designed for movement, not for static work. Computer users should take a brief, 5-second movement break at the keyboard every 10 minutes. The third element is to choose equipment and workstation setup to ensure comfortable ergonomic use. The Centers for Disease Control (n.d.) publishes ergonomic guidelines for computer use covering work area, desk/workstation, chair, monitor, keyboard, mouse, lighting, work habits, laptop computer.

IMPACT ON THE HEALTH PROFESSIONS

Health Information

Consumers

More than 60% of Internet users sought health information in the past 6 months (Ferguson, 2002). The Internet is a revolutionary information technology that rivals speech, writing, and books in its potential to transform the exchange of human knowledge. The Internet already contains vast archives of scientific and health literature that is accessed not only by health professionals but also by consumers. A new area has developed called consumer health informatics that deals with this recent ability of consumers to access the entire body of published health care and scientific literature online, in abstract form via PsycINFO, MEDLINE, and other important searchable bibliographic databases. More than 30,000 Web sites offer medical information (Ferguson, 2002).

For consumers, information gathered online can have both helpful and potentially harmful impact. Certainly it can help by providing useful information regarding conditions, symptoms, or drugs, but this information can also contribute to risks such as the following:

1. Propagating inaccurate science and fraudulent cures;
2. Wasting time or money;
3. Instilling unnecessary guilt, fear, anger;
4. Making prognoses worse by delaying treatment; and
5. Leading to alienation or withdrawal from health care.

To some extent, these risks are part of all forms of information seeking. They essentially address a larger issue of how people react to and manage the vast amount of information available today, which can create anxiety and information overload.

Professionals

In addition to the vast amount of medical information now posted on the Internet, peer-reviewed research evidence that was once available only to clinicians is now accessible to everyone. The National Institutes of Health and CenterWatch maintain public online listings of investigative clinical trails and newly Food and Drug

Administration–approved therapies. Medical references such as the Merck Manual are accessible online. Patients now come to appointments armed with research evidence and knowledge about the latest treatment options, information that had previously been available only to health care professionals. This is changing some of the ways that health professionals interact with their patients because patients often are more involved in negotiating their treatments.

For health professionals, the Internet provides access to research on clinical interventions and online seminars for exchanging professional knowledge, news, and meeting announcements. In addition, the Internet contains numerous pamphlets, brochures, and other self-help resources that can be given to clients. Health professionals use e-mail to conduct research, write articles, and disseminate professional news and events.

Health professionals must be constantly engaged in learning activities. The Internet is fast becoming the prevailing infotechnology for learning, especially activities involving information retrieval, processing, creation, and communication. Its usefulness in training skills such as direct care in nursing and psychotherapy is still limited, but the Internet is a powerful infotechnology tool for staying current with research findings, learning about developments in the health field, and engaging in peer collaboration and supervision. This shift in educational principles and practices related to networked digital technologies will affect the whole clinical educational enterprise. As one example, the venerable clinical training institutions of ground rounds is now being placed online—at the University of California at Los Angeles (UCLA), Stanford, and University of Chicago.

Since Hippocrates, health professionals have held interdisciplinary case seminars to disseminate the latest understandings in their fields. The Internet allows health professionals to participate in online case seminars that expand their personal knowledge base and refine their professional skills. Like a case seminar, the Internet is inherently cross-disciplinary. During a search, resources on health are found from the perspectives of all health care disciplines (e.g., medicine, psychology, nursing, social work, etc.). The information exchange and communication capacities of the Internet create a global virtual case seminar which provides access to cross-disciplinary resources and colleagues.

Telehealth and Psychological Services

Telehealth is generally used as an umbrella term to describe all the possible variations of healthcare services using telecommunications, such as health assessment, diagnosis, intervention, consultation, supervision, education, and information across distance. Telehealth includes the delivery of health services via the Internet but also includes telephone, television, video, and fax. There is a convergence of functions and hardware in the telecommunications field. Videoconferencing, for example, can be conducted over the Internet using cameras connected to computers, or the data can be sent over telephone lines, or over a closed proprietary Intranet. Telehealth is likely to become a significant part of the future of health care. The U.S. government already heavily invests in telehealth—more than $20 billion in 2000 (mostly in the military). Congress specifically included funding for $200 million per year for telemedicine reimbursement in Medicare. Telehealth has already become a major part of certain specialties.

In medicine, several specialties have already developed approaches that make use of the Internet. Teleradiology, the sending of X rays, computed tomographic scans, or magnetic resonance images (store-and-forward images) is the most common application of telehealth in use today. There are hundreds of medical centers, clinics, and individual physicians who use some form of teleradiology. In telepathology, images of pathology slides are sent from one location to another for diagnostic consultation. Dermatology also makes use of interactive technology for exams. Digital images may be taken of skin conditions, and sent to a dermatologist for diagnosis.

Two-way interactive television (IATV) is used particularly in rural and wilderness health care when a "face-to-face" consultation is necessary. It is usually between the patient and his or her provider in one location and a specialist in another location. Videoconferencing equipment at both locations allow a "real-time" consultation to take place. It can make specialty care more accessible to underserved populations. Video consultations from a rural clinic to a specialist can alleviate prohibitive travel and associated costs for patients. Videoconferencing also opens up new possibilities for continuing education or training for isolated or rural health practitioners, who may not be able to leave a rural practice to take part in professional meetings or educational opportunities.

The military and some university research centers are involved in developing robotics equipment for telesurgery applications. A surgeon in one location can remotely control a robotics arm for surgery in another location. The military has developed this technology particularly for battlefield use, and some U.S. academic medical centers and research organizations are also testing and using the technology.

The first use of telecommunications in mental health projects date back to 1959, when the University of Nebraska School of Medicine began experimenting with a closed-circuit television link to provide psychiatric and other health services between the Nebraska Psychiatric Institute and Norfolk State Hospital. Telehealth is just taking off as a means of delivering professional service. Although barriers and challenges to the adoption of telehealth have been discussed since its inception, only in the last few years has a solid consensus emerged as to what those barriers are and just how they should be addressed. That consensus, coupled with consumer and commercial demand for affordable telecommunications technology, has spurred health care communications and other manufacturers to develop affordable, user-friendly systems. Finally, legislators, seeing affordable technological options, hearing consensus around both barriers and solutions, and receiving requests from undeserved constituents that the government help implement such solutions, have begun to act to ensure that telehealth technology becomes a significant part of the continuing evolution of health care (Kirby, Hardesty, & Nickelson, 1998).

Online Therapy

The first attempt to use computers for delivery of therapy dates back to a 1972 demonstration of a psychotherapy session between computers at Stanford and UCLA. The earliest known organized service to provide mental health advice online is Ask Uncle Ezra, a free service offered to students of Cornell University in Ithaca, New York (named for Ezra Cornell, the University's founder), which has been in continuous operation since September 1986. Fee-based mental health services offered to the public began to appear on the Internet in mid-1995. Most were of the "mental health advice" type, offering to answer one question for a small fee. Ainsworth (2002) maintains a Web site database that describes what services online therapists offer, ranging from e-mailed answers to questions to two-way video, and she also runs credentials checks to make sure they have the licenses and degrees they claim. She reports that in the fall of 1995, when she first did a search, she found 12 e-therapists practicing on the Internet. By 2002, her database had grown to include more than 300 private-practice Web sites on which e-therapists offer services, and the new e-clinic represent collectively nearly 700 more e-therapists.

Despite the increasing numbers, use of the Internet to provide psychotherapy is controversial. Even most therapists who deliver services via the Internet do not claim that what they are doing is psychotherapy. Most clearly state that the exchanges are educational rather than therapeutic in nature. Traditional psychotherapy does not seem to be part of the online offer of mental health services. Yet there is little doubt that many people have been helped, some profoundly, by interacting with mental health professionals over the Internet. Ainsworth (2002) describes her personal e-mail-based online therapy, which lasted 2 years without ever seeing or talking to her therapist.

> I would compare it to keeping a journal in that every day when I wrote him an e-mail I explored my thoughts and feelings in great depth. But usually when you keep a journal it doesn't talk back to you. He challenged me. The fact that we communicated by e-mail, I think, made me feel like he was inside my head and present in my life.

The International Society for Mental Health Online (http://ISMHO.org) was formed in 1997 to promote the understanding, use, and development of online communication, information and technology for the international mental health community. After considerable discussion and debate, they published "Suggested Principles for the Online Provision of Mental Health Services." Their guidelines address many issues that apply to all forms of therapy but have a unique application to online modes of delivery, such as the following:

- Ensuring competence in online delivery of therapy,
- Conformity with state online practice guidelines,
- Structure of the online services,
- Turnaround time, and
- Privacy of the counselor (the therapist's right not to have e-mail or other communications disseminated).

A promising area in online delivery of health services is virtual reality exposure therapy, which involves exposing the patient to a virtual environment containing the feared stimulus in place of taking the patient into a real environment or having the patient imagine the stimulus. A team of therapists and computer scientists developed a treatment for fear of heights (acrophobia) that was shown to be effective in reducing acrophobic subjects' anxiety and avoidance of heights and in improving attitudes toward heights. The developers note that this virtual reality approach is particularly amenable to telehealth:

> Virtual reality exposure therapy is appropriate for networked delivery of clinical psychology and psychiatry services to remote locations. Since the patient is receiving therapy within a virtual environment, the clinician conducting the therapy session could be present physically or participate via computer networks from a remote location. (Kooper, n.d.)

State regulations vary from state to state, however, and by virtue of state-based licensing system for health professionals, guidelines developed in one state would only apply to therapists licensed in that state. The patchwork of licensing regulations has functioned as an impediment to dissemination of telehealth initiatives. For example, in California, the Telemedicine Act of 1996, while mandating some forms of payment for telemedicine, does restrict the provision of mental health services to patients within the state boundary. Yet few doubt that telehealth will play an increasing role in health delivery, as the Internet expands it influence on a wider and wider range of human behavior and experience.

CONCLUSION

The Internet offers both consumers and professionals new opportunities to expand their knowledge base and to obtain support for their health problems and work. Yet, as with other technologies, there are costs to the fabric of the psyche and society that require adjustments in how people relate to the medium and each other. Despite its recent development, the rate of adoption of online communication is unparalleled in human history. It will be many more years before the full implications of the Internet for health and well-being are understood and addressed. Meanwhile, the Internet will expand in its usefulness and create new problems for society to address.

GLOSSARY

Disinhibition A lack of interpersonal feedback characteristic of face-to-face communication that can lead to enhanced self-disclosure as well as personal attacks.

Internet addiction Dysfunctional Internet use patterns.

Mediated communication Non–face-to-face interactions conducted with a communication medium.

Neurosphere A "global mind" created by the information sharing and communication capacities of the Internet.

CROSS REFERENCES

See *Health Insurance and Managed Care; Internet Literacy; Medical Care Delivery.*

REFERENCES

Ainsworth, M. (2002). Metanoia web site. Retrieved September 30, 2002, from http://www.metanoia.org/imhs

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Centers for Disease Control. (n.d.). Computer workstation ergonomics. Retrieved September 30, 2002, from http://www.cdc.gov/od/ohs/Ergonomics/compergo.htm

Dell Computer Corporation. (1993, July 26). Fear of technology is phobia of the '90s; Computer habits, attitudes determine "Techno-Type" (press release). Austin, TX: Author.

Dodson, E. (2001, December/2002, January). Technology, humanity, and humanistic psychology. *AHP Persective,* 17–19.

Ellen, P., Beardon, W., & Sharma, S. (1991). Resistance to technological innovations: An examination of the role of self-efficacy and performance satisfaction. *Journal of the Academy of Marketing Sciences, 19,* 297–307.

Emes, C. E. (1997). Is Mr. PacMan eating our children? A review of the effect of video games on children. *Canadian Journal of Psychiatry, 42,* 409–414.

Fenichel, M. (n.d.). "Internet addiction": Addictive behavior, transference or more? Retrieved September 30, 2002, from http://www.fenichel.com/addiction.shtml

Ferguson, T. (2002) The Ferguson Report. Retrieved September 30, 2002, from http://www.fergusonreport.com/articles/fr00901.htm

Gackenbach, J. (Ed.). (1998). *Psychology and the Internet: Intrapersonal, interpersonal, and transpersonal implications.* New York: Academic Press.

Greenfield, D. (1999). Virtual addiction: Help for netheads, cyberfreaks, and those who love them. Oakland, CA: New Harbinger.

Greenfield, P. (2000). Digital childhood: A research agenda on human development & technology. Webcast retrieved July 14, 2001, from http://www. dacadeofbehavior.org/digitalchild

Greenfield, P. M., Brannon, C., & Lohr, D. (1996). Two-dimensional representation of movement through three-dimensional space: The role of video game expertise.; In P. M. Greenfield & R. R. Cocking (Eds.); *Interacting with video. Advances in applied developmental psychology, Vol 11,* (pp. 169–185).

Gross, E., Juvonen, J., & Gable, S. (2002). Internet use and well-being in adolescence. *Journal of Social Issues, 58,* 75–90.

Groves, D. L., & Slack, T. (1994). Computers and their application to senior citizen therapy within a nursing home. *Journal of Instructional Psychology, 21,* 221–226.

Gustafson, D. H., Hawkins, R., Pingree, S., McTavish, F., Arora, N. K., Mendenhall, J., Cella, D. F., Serlin, R. C., Apantaku, F. M., Stewart, J., & Salner A. (2001). Computer support for elderly women with breast cancer. *Journal of General Internal Medicine, 16,* 435–445.

Haupt, P. A. (1998). *How does corporate-sponsored computer-mediated communications (CMC) technology affect the experience of expatriates who use the technology during their offshore assignment?* Unpublished Ph.D. dissertation, the Fielding Graduate Institute, Santa Barbara, CA.

Heckler, R. (2001). Somatics in cyberspace. In P. Denning (Ed.), *The invisible future: The seamless integration of technology into everyday* (pp. 277–294). Columbus, OH: McGraw–Hill.

Huston, A., Wartella, E., & Donnerstein, E. (1998). *Measuring the effects of sexual content in the media: A report to the Kaiser Family Foundation.* Retrieved September 30, 2002, from http://www.kff.org/content/archive/1389/content.html

Jackson, D. N., III, Vernon, P. A, & Jackson, D. N. (1993). Dynamic spatial performance and general intelligence. *Intelligence, 17,* 451–460.

Kiesler, S., Bozena, Z., Lundmark, V., & Kraut, R. (2000). Troubles with the Internet: The dynamics of help at home. *Human–Computer Interaction, 15,* 323–351.

Kirby, K., Hardesty, H., & Nickelson, D. (1998). Telehealth and the evolving health care system: Strategic opportunities for professional psychology, professional psychology. *Research and Practice, 29,* 527–535.

Kooper, R. (n.d.). Virtual reality exposure therapy. Retrieved September 30, 2002, from http://www.cc.gatech.edu/gvu/virtual/Phobia/phobia.html

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues, 58,* 49–74.

Kraut, R, Patterson, M, & Lundmark, V. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being. *American Psychologist, 53,* 1017–1031.

Lakoff, G., & Johnson, M. (1983). *Metaphors we live by.* Chicago: University of Chicago Press.

Lawrence, S., & Giles, C. (1999). Accessibility of information on the web. *Nature, 400,* 107.

Linden, P. (1995). *Compute in comfort: Body awareness training: A day-to-day guide to pain-free computing.* New York: Prentice-Hall.

Negroponte, N. (1995). *Being digital.* New York: Knopf.

Rosen, L., & Weil, M. (1997). *Technostress.* New York: John Wiley.

Rosen, L., & Weil, M. (1998). Multitasking madness. *Context Magazine.* Retrieved September 30, 2002, from http://www.contextmag.com/setFrameRedirect.asp?src=/archives/199809/InnerGameOfWork.asp

Rosen, L., & Weil, M. (2000). Results of our 49-month study of business attitudes show clerical/support staff, managers and executives using more technology at work and at home and becoming more hesitant toward new technology. Retrieved September 30, 2002, from http://www.technostress.com/busstudy2000.htm

Selye, H. (1978). *The stress of life.* Columbus, OH: McGraw-Hill.

Shapira, N. A., Goldsmith, T. D., Keck, P. E., Khosla, U. M., & McElroy, S. L. (2000). Psychiatric features of individuals with problematic internet use. *Journal of Affective Disorders, 57,* 91–107.

Shenk, D. (1998). *Data smog: Surviving the information glut.* San Francisco: Harper.

Spears, R., Postmes, T., Lea, M., & Wolbert, A. (2002). When are net effects gross products: The power of influence and the influence of power in computer-mediated communication. *Journal of Social Issues, 58,* 91–107.

Suler, J. (1999). *Computer and cyberspace addiction.* Retrieved September 30, 2002, from http://www.rider.edu/users/suler/psycyber/cybaddict.html

Szabo, R. M. (1998). Carpal tunnel syndrome as a repetitive motion disorder. *Clinical Orthopedics, 351,* 78–89.

Thornburgh, D., & Lin, H. (Eds.). (2002). Youth, pornography, and the Internet. Washington, DC: National Academy Press.

Turkle, S. (1997). *Life on the screen: Identity in the age of the Internet.* New York: Simon & Schuster.

Tyler, T. R. (2002). Is the Internet changing social life? It seems the more things change, the more they stay the same. *Journal of Social Issues, 58,* 195–205.

Young, K. (1996). Psychology of computer use: Addictive use of the Internet. *Psychological Reports, 79,* 899–902.

Young, K. (1998). *Caught in the Net: How to recognize the signs of internet addiction and a winning strategy for recovery.* New York: John Wiley & Sons.

Young, K. (n.d.). Internet Addiction Test Index. Retrieved September 30, 2002, from http://www.netaddiction.com/resources/iaindex.htm

Weil, M., & Rossen, L. (n.d.). A conversation with *TechnoStress* authors. Retrieved September 30, 2002, from http://www.technostress.com/tsconversation.htm

FURTHER READING

Bargh, J. A., McKenna, K., & Fitzsimons, G. (2002). Can you see the real me? Activation and expression of the "true self" on the Internet. *Journal of Social Issues, 58,* 33–48.

Bidgoli, H. (2002). *Electronic commerce: Principles and practice.* San Diego, CA: Academic Press.

CDC National Institute for Occupational Health and Safety. (1997). Carpal tunnel syndrome. Retrieved September 30, 2002, from http://www.cdc.gov/niosh/ctsfs.html

Cocking, R. (Ed.). Interacting with video. *Advances in applied developmental psychology* (pp. 169–185). Norwood, NJ: Ablex.

Crowley, D. & Heyer, P. (1999). *Communication in history: Technology, culture, society.* New York: Longman.

Farris, M., Bates, R., Resnick, H., & Stabler, N. (1994). Evaluation of computer games' impact upon cognitively impaired frail elderly [Special issue: *Electronic tools for social work practice and education: I*]. *Computers in Human Services, 11,* 219–228.

Fisher, D. R. (2001). On utopias and dystopias: Toward an understanding of the discourse surrounding the Internet. *Journal of Computer-Mediated Communication, 6*. Retrieved July 10, 2001, from http://www.ascusc.org/jcmc/vol6/issue2/fisher.html

Gainey, T. W., Kelley, D. E., & Hill, J. A. (1999). Telecommuting's impact on corporate culture and individual workers: Examining the effect of employee isolation. *SAM Advanced Management Journal, 64,* 4.

Jarvenpaa, S. (1998). Communication and trust in global virtual teams. *Journal of Computer-Mediated Communication, 3,* 4.

McKenna, K., Gren, A., & Gleason, M. (2002). Relationship formation on the Internet: What's the big attraction. *Journal of Social Issues, 58,* 9–31.

O'Keefe, B. (2000). *Summary and synthesis: Digital kids*. Retrieved July 14, 2001, from http://www.dacadeofbehavior.org/digitalchild/workgroupsumms.html

Podlas, K. (2000). Mistresses of their domain: How female entrepreneurs in cyberporn are initiating a gender power shift. *CyberPsychology and Behavior, 3,* 847–854.

Schleerlis, W. (1998). Internet paradox: A social technology that reduceds social involvement and psychological well-being. *American Psychologist, 53,* 1017–1031.

Sullivan, C., & Lewis, S. (2001). Home-based telework, gender, and the synchronization of work and family: Perspectives of teleworkers and their co-residents. *Gender, Work and Organization, 8,* 123–145.

Walsh, D. (2001). Fifth annual video and computer game report card. Retrieved July 15, 2001, from http://www.mediaandthefamily.org/research/vgrc/2000–2.shtml

History of the Internet

John Sherry, *Purdue University*
Colleen Brown, *Purdue University*

INTRODUCTION

In his prophetic and now famous 1945 *Atlantic Monthly* article, Vannevar Bush noted that inefficiencies in the exchange of new theory and discovery slowed scientific progress. He envisioned a system for rapid dissemination and organization of scientific information, instantly available to all who seek it, which he named the "memex." The memex was to be a "transparent platen" on which books, pictures, periodicals, newspapers, longhand notes, photographs, and other information would be provided to anyone who had access to the machine and knowledge of its indexing system. He foresaw scientists accessing information from around the world at unprecedented rates based on the same associational principles by which the human mind operates. Further, he saw scientists and professional data miners recording their journey through the data such that anyone who was interested could follow. A half century after the article appeared, his vision was realized in the form of the World Wide Web and related Internet technologies. This is the story of the inventors, ideas, and innovations that made Bush's vision a reality.

The Internet was created by a collection of visionaries and executed by hundreds of individuals whose contributions helped develop the technology (for a time line, see Figure 1). The result is the technological and intellectual infrastructure of the Information Age, supporting the majority of the emerging economy for the 21st century. How did this marvel come to be? This chapter tells the story of the creation of the Internet and the World Wide Web. Guided by a unique vision of computer communication in a world of calculating machines and driven by the desire to solve a series of complex engineering problems, the Internet emerged as an open-architecture marvel of scientific and engineering cooperation. The principles that guided the creation of the Internet typify the finest realization of American cooperative scientific ideals. The Internet as we know it is the result of dedication to a set of core principles: that this communication system be the result of cooperation among interested parties, be open to new ideas, and be scalable. The vision was realized in a series of fits and starts, problems and solutions, ideas and the implementation of those ideas by engineers pushing the boundaries of current technology. It is the result of intellectual and political argumentation and compromise, funded by the military–industrial complex, but realized in the world of the academic. This chapter can hardly be called a history, though, because the open architecture of the Internet encourages further experimentation. Instead, this chapter can best be called a beginning.

A UNIQUE VISION

The Industrial Revolution of the 19th century brought about the development of technologies that allowed people, ideas, and products to travel across long distances in a short amount of time, facilitating exploration of scientific and technological frontiers (Moschovitis, Poole, Schuyler, & Senft, 1999). Nineteenth-century scientific exploration laid the groundwork for the technological antecedents of modern computing and networking that came into popular use in the first half of the 20th century, from the telephone to such early colossal computers as ENIAC. These early 20th-century technologies were similarly dedicated to two goals: the desire to increase the ease of communication across distance and the desire to provide resources that assist humans in the efficient processing of information. However, it was the charged atmosphere surrounding the Cold War that provided the initial momentum to bring these two goals together as a matter of national defense and pride.

The Marriage of Science and National Defense

By the late 1940s, the Soviet Union had long-range bombers and atomic capabilities. The Department of Defense, under President Harry Truman, enlisted the help of researchers at RAND and the Massachusetts Institute of Technology (MIT) Lincoln Laboratory to develop the

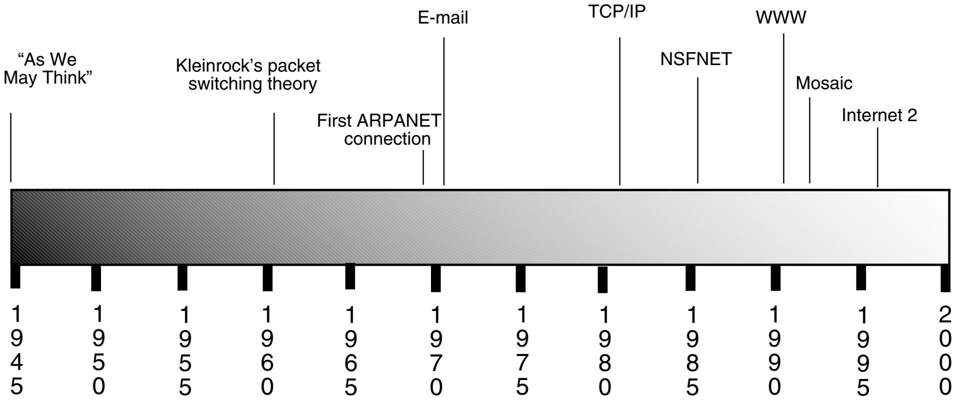


Figure 1: Timeline of major Internet events.

Semi-Automated Ground Environment (SAGE) to detect and counteract Soviet airborne attacks (Segaller, 1998). It was at this time, the 1940s and early 1950s, that many of the graduate students and scientists at MIT and Lincoln gained the preliminary experience necessary to create networked computers, including connecting computers to phone lines, using computers to handle real-time data coming from antennas and submarines, digitizing communications, and developing and refining faster, more reliable computers (Segaller, 1998).

On October 4, 1957, the Soviets launched Sputnik, the first human-made Earth satellite. The Eisenhower administration's fears of technological weakness were ignited; the Soviet Union was leading the United States in space technology. Science and technology were instantly wed to national defense and the relationship was thrust into the social and political limelight as President Eisenhower, a strident supporter of science prior to Sputnik, ensured the panicked public that he was committed to providing massive support for defense research. Eisenhower immediately convened a meeting of his Presidential Science Advisory Committee, and in November 1957, he appointed MIT president James Killian as the official presidential scientific advisor (Hafner & Lyon, 1996).

ARPA Is Created

Working closely with President Eisenhower and the Secretary of Defense, Neil McElroy, Killian recommended that Congress create the Advanced Research Projects Agency (ARPA) within the Department of Defense (Zakon, 2002). Congress quickly approved start-up funds of $520 million and an annual budget of $2 billion. In early 1958, ARPA was officially established as the U.S. government's sole research agency dedicated to the development of space-related military technology. The vigorous national commitment to beating the Soviets in defense and space technology had officially begun for the government and the American public. Although much of the computer-related funding at this time came from the Department of Defense, many of the scientists involved with ARPA used the surge in governmental support as an opportunity to realize their academic pursuits (Abbate, 1999). The intersection of military, governmental, academic, and public goals remains an integral part of the creation and maintenance of the modern Internet.

In 1958, the National Aeronautics and Space Administration (NASA) was split from the computer research section at ARPA in order to oversee all space exploration and missile research and to assure that this technology was developed in the civilian sector. ARPA was left with the relatively meager budget of about $150 million. However, the removal of space and missile research from ARPA's agenda allowed researchers to focus greater effort on the emerging fields of computer science and information processing. At this point, ARPA directly funded most of the top computer science researchers in the United States and at least partially funded projects at most of the research centers and universities engaged in high-tech endeavors (Hafner & Lyon, 1996).

Networking Visions Emerge

In 1962, the first director of ARPA, Jack Ruina, hired mathematician, behavioral psychologist, and burgeoning computer science connoisseur Joseph C. R. "Lick" Licklider to head ARPA's command and control division. In his groundbreaking essay "Man-Computer Symbiosis," Licklider (1960) proposed the idea of interactive computing, an idea taken for granted by every personal computer user today. Inspired by the visionary works by Bush (1945) and Alan Turing (1950) on the use of computers to augment human intelligence, Licklider also prophesized that computers would become more than mere calculating tools. He envisioned that the relationship between humans and computers would evolve and result in cooperative decision making and problem solving in real time, that is, without the long delays that characterized computing in the 1950s (Licklider, 1960).

In addition to envisioning the eventual state of interaction between human and computer, Licklider was the first to envision the linking of many computers at locations distant from each other. Nine years before it became a reality, he wrote:

> It seems reasonable to envision . . . a "thinking center" that will incorporate the functions of present-day libraries together with anticipated

> advances in information storage and retrieval... The picture readily enlarges itself into a network of such centers, connected to one another by wide-band communication lines and to individual users by leased-wire services. In such a system, the speed of the computers would be balanced, and the cost of the gigantic memories and the sophisticated programs would be divided by the number of users. (Licklider, 1960, p. 15)

Licklider expanded on the idea of a system of linked computers in a series of historically important memorandums written with Wesley Clark, entitled "On-Line Man Computer Communications" (Licklider & Clark, 1962). They wrote of a "galactic network" where computers and information would be linked and accessible to everyone. Larry Roberts, a successor of Licklider's at MIT, commented that, "[t]he vision [of linking many computers together] was really Lick's originally. Lick saw this vision in the early sixties. He didn't have a clue how to build it. But he knew it was important" (Segaller, 1998, p. 40).

Licklider's outline of an interactive network linking people and resources together was also seen as a possible solution to a practical problem surrounding computing in the early 1960s: stretched resources. In the 1950s and 1960s, ARPA had commissioned laboratories around the country to carry out a variety of projects in science and technology, providing each with large, expensive mainframe computers manufactured by different companies. Individual laboratories altered their computers, adding different applications, packages, and hardware, and each subsequent laboratory wanted the latest computing capabilities available. The enormous costs of building separate units for each project began to snowball.

In early 1966, Robert Taylor, director of the Information Processing Techniques Office (IPTO, the computing division of ARPA), proposed a remedy to the scarce resource problem that incorporated the latest research and developments: Build a system of electronic connections between research computers at different geographical locations across the country to allow convenient, quick, and inexpensive sharing of resources between many scientists (Hafner & Lyon, 1996). The idea of networking was already being implemented on a smaller scale, for example linking two computers together (Marill & Roberts, 1966), but nothing that approached the scale or complexity of Taylor's proposition had been attempted so far. The technology of the time did support such a notion. In 1958, communications engineers at Bell Laboratories had built the modulator–demodulator, or simply "modem" (Anderberg, 2002). The modem made it possible to convert data from the digital computer format into the analog telephone line format, send it over existing phone lines, and translate it back to digital computer format. By using the long-distance infrastructure already in place, the task of linking many computers for the purpose of sharing resources was indeed an attractive and financially sound proposal. In 1966, Taylor pitched the idea to his boss, ARPA Director Charlie Hertzfeld, and received approval and financial support to build the experimental network. In 1967, Taylor enticed Larry Roberts to leave Lincoln Lab and join him at ARPA to begin planning the proposed computer network, which would soon become known as the ARPANET (Segaller, 1998).

Networking: From Theory to Practice

Although Licklider imagined what a distributed, interactive computer network could accomplish and Taylor imagined what problems it could solve, Leonard Kleinrock had theorized in 1961 the new communication technology that could make such a network possible (Hafner & Lyon, 1996). Two major obstacles to creating a long-distance computer network were (a) being able to receive and process data over existing phone lines fast enough to allow efficient, real-time processing, and (b) creating a network that could support the unique needs involved in the exchange of computer-generated data. While an MIT graduate student in 1961, Kleinrock proposed what would later be called "packet-switching" technology, the most efficient way to send data over a network (Kleinrock, 1961, 1964). Two other researchers, Bob Davies at the British National Physical Laboratory (NPL) and Paul Baran at the RAND Corporation, were also investigating the idea of packet switching in unrelated research projects in the early 1960s (Baran, 1964; Davis, Bartlett, Cantlebury, & Wilkinson, 1967). Ultimately, Davies and his group of researchers at NPL would be the first to test a packet-switching network within a building, in Middlesex, England, in 1967 (Zakon, 2002).

Packets, Queuing, Demand Access—The Theory of Networking

Packet switching involved chopping messages into packets, assigning each packet a number, sending the packets of data independently through the network, and reassembling the packets at the other end in the proper order in readable form (Hafner & Lyon, 1996). The packets could queue up at nodes in the network until they were requested (demand access). In this fashion, network resources could be used as needed and not tied up with open circuits. Researchers hypothesized that packet switching would be a dramatic improvement over circuit-switched telephone networking because it avoided the inefficiency of a continuously open connection. In order to put packet-switching data transfer theory into action, however, ARPANET researchers had to address several key obstacles.

ARPANET planners were well aware that the computers that would be connected to the network were incompatible. Each ARPA-funded site had purchased different mainframe computers from different companies, using different operating systems and different languages (Segaller, 1998). Moreover, each site research team was only familiar with the operation of the computer in their respective laboratory. How could these mainframe computers be connected, share data, and communicate with one another with a minimum of errors? Taylor called a meeting of ARPA-funded researchers at the University of Michigan in early 1967 to discuss this and other unresolved network design issues (Anderberg, 2002).

Building on an idea proposed by Wesley Clark, a computer scientist at Washington University in St. Louis, the ARPA researchers envisioned a system of smaller,

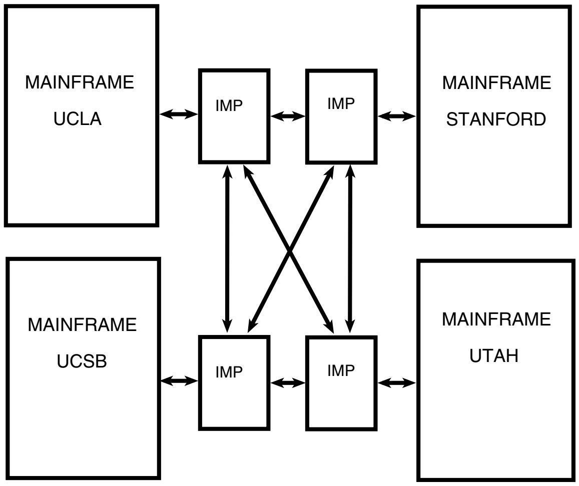


Figure 2: The interface message processing design.

intermediate computers called Interface Message Processors (IMPs) to solve the mainframe-to-mainframe incompatibility problem (Abbate, 1999). Each IMP would act as an interface between the network and its mainframe computer (see Figure 2). Every mainframe had to send the same format packets to the IMPS and there was a standard interface all mainframes had to implement. NCP dealt with incompatibilities on an end-to-end basis. Each site's mainframe would communicate with its own IMP, and all IMPs would be designed to communicate with one another using the same operating system and speaking the same language. Roberts distributed a Request for Quotation (RFQ) to design and build the IMPs in July 1968 and received more than 100 proposals. He awarded the contract to Frank Heart and his team of computer scientists at Bolt, Benerak, & Newman (BBN) in Cambridge, Massachusetts, in December 1968 (Anderberg, 2002). The final step was deciding which sites would initially be connected to the network. Four sites, each with its own particular research specialties, were chosen: UCLA; the University of California, Santa Barbara (UCSB); Stanford Research Institute; and the University of Utah. The deadline for establishing the network at UCLA was set, somewhat arbitrarily according to Heart, for September 1, 1969, only nine months after the initial proposal was presented. Heart and his group would have $1 million and less than a year to turn theory into a working system (Hafner & Lyon, 1996).

In the summer of 1968, IPTO organized a meeting for graduate students from the universities that would be the first sites connected to the network to discuss the technological aspects of the proposed network. The meeting gave rise to the Network Working Group (NWG), which solidified the collaborative nature of the ARPANET project (Abbate, 1999). Rather than one single scientist taking charge of the direction of the group, the attendees recorded their thoughts and deliberations in a set of notes that Steve Crocker soon named Request for Comment (RFC). Crocker published the first RFC in April 1969, detailing the interface between the hosts and IMPs and asking for input on problems (Crocker, 1969). The RFC became firmly established as a means of communicating information and seeking feedback about networking issues, and it was supported by all levels of researchers. This open manner of sharing ideas in the realm of scientific discovery remains in place to this day and, by many accounts, influenced the eventual spirit of community associated with the modern-day Internet (Internet Society, 1999). The entire series of RFCs have been collected and catalogued by Jon Postel, the RFC editor for more than 30 years, and are available for viewing (RFC Editor Homepage, n.d.).

It is important to note that at this point, the number of researchers working on the development of ARPANET was relatively small. In fact, as news of the proposed network spread to scientific circles across the country, the idea was met with little enthusiasm. To the many ARPA-funded research teams, the idea of sharing resources across a network meant that "outsiders" would be meddling in their private computing centers and that their hard-earned computing equipment and power would be depleted. To researchers who still hoped that ARPA would fund their projects, the idea of a network meant that ARPA might not purchase new mainframes for them. Roberts, Taylor, and Licklider tried to win the support of the largely uninterested and skeptical ARPA-funded researchers, eventually resorting to subtle "blackmail" (Segaller, 1998), reminding each site that as they were being supported by ARPA to the tune of hundreds of thousands, if not millions, of dollars a year, they should be just as enthusiastic about the project as the project managers were. Needless to say, ARPA-funded researchers started to outwardly come around, though inwardly most doubted the venture would work (Moschovitis, Poole, Schuyler, & Senft, 1999).

ARPANET Comes Alive

By early 1969, the researchers at BBN were working around the clock to develop the IMPs by the deadline, while the research teams at each university were busy developing the custom software and hardware that would allow communication between their mainframe and the IMP. Significant challenges arose every day, all of which had to be resolved under the pressure of the looming deadline. To everyone's amazement, the UCLA IMP was delivered in time for the September 1 deadline and the IMP-mainframe connection was ready to go a few days later (Zakon, 1993). The first test of the packet-switching technology, between the UCLA IMP and its mainframe, was a success. On September 2, 1969, the two machines began talking to each other. The second IMP arrived, at Stanford Research Institute, on schedule one month later and their first official test occurred on October 1, 1969 (Hafner & Lyon, 1996).

The plan was to log-on from the UCLA host to the Stanford host by typing "LOG IN." An L was typed at UCLA and Stanford received an L. An O was typed at UCLA and Stanford received an O. A G was attempted and the system crashed. A couple of hours later, however, the system was up and running again, and the researchers successfully logged onto the network, performed some minimal operations, and logged off. The connection between UCLA and Stanford was a success, and the remaining two nodes, one at UCSB, and one at the University of Utah, were working by the end of the year (Zakon, 1993). The technology of ARPANET proved that a long-distance packet-switching computer network was possible. Eventually, the ARPANET helped ARPA cut its computer budget by 30% and impressed the skeptics by providing them with more computing power and allowing researchers at different institutes to exchange papers, data, ideas, and software programs and upgrades.

GROWING THE NETWORK

The fifth ARPANET node was installed at BBN in March 1970, and over the next few years, Vint Cerf and Steve Crocker, under Kleinrock at UCLA, and Frank Heart and Bob Kahn, then at BBN, all worked to push the network to the limit, intentionally breaking the network in order to discover weaknesses in the design. They also formed the International Network Working Group in 1972 to coordinate standards. As the number of sites connected to the network steadily grew, moving from the West Coast to the East Coast, new technological innovations kept emerging, some expected and some surprising.

Electronic Mail: The "Killer" Application

In 1971, Ray Tomlinson, a computer engineer at BBN, devised an experimental program, called CPYNET, for sending files between computers. Tomlinson added an additional program to CPYNET for sending and receiving messages, called SNDMSG and READMAIL (Hafner & Lyon, 1996). Within a few months, Tomlinson's application was being used to send messages over the network to machines in different geographical locations. Tomlinson paired each user's name with the user's host machine identification using the symbol "@" (meaning "at"), not knowing he was coining the universal symbol of the soon-to-be wired world (Hafner & Lyon, 1996). The resulting application, "e-mail" for short, was the first killer application for the emerging network. The network e-mail capability spread rapidly, quickly becoming the most widespread use for the network in the early years. Kleinrock noted that "[a]s soon as e-mail came on, it took over the network. We said, 'Wow, that's interesting'. We should have noticed there was something going on here. There was a social phenomenon happening" (Segaller, 1998, p. 105). The original purpose of the network, to share computing resources across distant geographical locations, was surpassed by its use as a communications tool. The instant popularity of e-mail among the scientists came as quite a surprise because the rationale for building the network was to share access to computers, not access to people (Abbate, 1999). Other applications were soon developed that facilitated the creation of "virtual" work groups and news groups, firmly establishing the network as a "virtual community" (Moschovitis, et al., 1999).

ARPANET Makes Its Public Debut

At this point, the use of ARPANET was almost exclusively confined to the networking branch of the computer science community, and the general public was unaware of the existence of it. Roberts knew that if the network was ever going to be widely adopted, it was time for a demonstration. Kahn and a small group of principle investigators were asked to devise a showcase of network capabilities. In 1972, ARPANET debuted to the public at the International Conference on Computer Communications (ICCC) in Washington, D.C. (Moschovitis, et al., 1999). By that time, ARPANET had expanded to 15 nodes with 23 sites.

ARPANET's debut to computer vendors, university representatives, government officials, scientists from a variety of disciplines, and the press at ICCC was a success. The system crashed only once over the weekend conference, and oddly enough, this gave Kahn and his colleagues the opportunity to show that problems were fixable (Hafner & Lyon, 1996). The conference attendees were astonished, curious, and anxious to become a part of the network; the buzz had officially begun. Although access to ARPANET was still under the tight control of ARPA and the U.S. Department of Defense, the demonstration at ICCC ignited curiosity and confidence in other networking projects. The early 1970s proved to be an intense period of network experimentation. Within 3 years of the first ARPANET transmission over telephone lines, different types of networks using packet-switching technology emerged. ARPANET remained under the control of the Department of Defense; thus, rather than expand the existing ARPANET, new networks were created. Networks that linked computers through radio connections, such as the ALOHAnet at the University of Hawaii, and ones that used satellites to link computers, such as SATnet, were popping up all over the world (Segaller, 1998).

A NETWORK OF NETWORKS

The success at the ICCC conference swelled the ranks of computer scientists who were interested in packet-switching networks. But the 1972 conference revealed that there remained a significant obstacle to realizing Licklider's ultimate vision of "... a 'thinking center' that will incorporate the functions of present-day libraries together with anticipated advances in information storage and retrieval and the symbiotic functions." Although computers within a network could be used for time-sharing electronic communication and the sharing of files, these same options were not possible between networks. That is, a computer scientist working at the University of Utah (ARPANET) could not exchange files with a colleague at the University of Michigan (MERIT). In order for the existing computer networks to become connected, there needed to be a new protocol so that computers on different networks could communicate with each other.

Robert Kahn began researching the internetworking problem at ARPA in 1973 (Kahn, 1994). In many ways, the problem of inter-networking was similar to the earlier problem of making dissimilar computers communicate that had been solved by using the IMPs. That is, each network used a different language to control the packet switching, preventing computers on different networks from being able to communicate with each other.

Kahn decided that the network of networks needed an open architecture that would allow for innovation within networks, without requiring existing networks to abandon their software. The problem of connecting networks was solved similarly to the earlier problem of connecting computers in a network. Kahn's main collaborator, Vinton Cerf, proposed the idea of "gateways" between the networks that would employ a common language or protocol. These gateways would allow networks to interconnect in much the same way that the IMPs originally allowed computers to interconnect. Thus, Kahn and his team specified a set of four requirements for the inter-network:

> Each distinct network would have to stand on its own and no internal changes could be required to any such network to connect it to the Internet. Communications would be on a best effort basis. If a packet didn't make it to the final destination, it would shortly be retransmitted from the source. Black boxes would be used to connect the networks; these would later be called gateways and routers. There would be no information retained by the gateways about the individual flows of packets passing through them, thereby keeping them simple and avoiding complicated adaptation and recovery from various failure modes. There would be no global control at the operations level. (Leiner, et al., 1994)

A common language was needed to allow computers from different networks to communicate with one another. Network control protocol (NCP), the software used to oversee packet switching on the ARPANET, lacked end-to-end host error control because packets never traveled outside the ARPANET. Cerf and Kahn wrote the transmission control protocol/Internet protocol suite (TCP/IP) to implement the packet-switching process at gateways (Cerf & Kahn, 1974). It took four incarnations before the final TCP/IP software was considered ready for deployment. TCP would be responsible for service features involved in the packet-switching process, including flow control and lost packet recovery, whereas the IP was used for the addressing and forwarding of individual packets. Standardized 32-bit IP addresses were assigned to each computer on the Internet; the first 8 bits signified the network and the other 24 bits designated the host on that network. Initially, TCP/IP protocol was used to connect the ARPANET (50 kbps) land lines, the Packet Radio Net (PRNET at 400/100 kbps), and the Packet Satellite Net (SATNET at 64 kbps; Kahn, 1994). On January 1, 1983, the ARPANET made the transition from NCP to TCP/IP as the standard protocol for networking on ARPANET. Now, it was possible to interconnect networks and the *Internet* was born.

Standardization of the TCP/IP software led to the development of commercially available gateways and routers for industry workstations, minicomputers, and mainframes. As the number of local area networks (LANs) increased, so did potential engineering incompatibility problems. The Internet Activities Board (IAB) was founded in 1983 to oversee standardization of networks. For example, the IAB designated three network types based on the expansion of network interconnectivity: class A, which were large-scale national networks; class B regional networks; and class C local area networks. With increasing interconnection came difficulty in locating information on the Internet. IP addresses of all host computers were found in a distributed file called HOSTS.TXT. Eventually, the updating of this file became a bottleneck, so Paul Mockapetris of USC/ISI invented the domain name system (DNS) to alleviate this problem (1983a, 1983b). DNS assigns and tracks easy-to-remember names to IP addresses across the Internet (e.g., purdue.edu).

The ARPANET/MILNET Split

By 1984, the computer architecture for the Internet was in place, and a number of computer networks, both governmental and commercial, had emerged. Within a few years, there were a number of networks linking geographically proximate universities. Among these networks were the following.

ARPANET (1969)—UCLA, Stanford University, UCBS, University of Utah, BBN, MIT, RAND Corporation, SDC, Harvard University, MIT's Lincoln Laboratory, University of Illinois—Urbana/Champaign, Case Western Reserve University, Carnegie Mellon University, NASA/Ames.

MERIT (1969)—University of Michigan, Michigan State University, Wayne State University.

THEORYNET (1977)—University of Wisconsin.

USENET (1979)—Duke University and University of North Carolina.

CSNET (1981)—University of Delaware, Purdue University, University of Wisconsin, RAND Corporation, and BBN.

BITNET (1981)—City University of New York and Yale University.

The expansion of the Internet to broader academic and commercial interests raised concerns about the security of Department of Defense information on the Internet. For that reason, the military component of the Internet was separated from the Internet, resulting in two networks: ARPANET and MILNET (the new military portion of the Internet). Computers on the MILNET would still be able to communicate with computers on the ARPANET, but the architecture was designed such that MILNET could be quickly disconnected from ARPANET in case of a threat of a security breach.

The NSFNet

Perceiving the advantage of the Internet for scientific research, the National Science Foundation (NSF) decided to expand use of the Internet beyond the limited number of computer scientists and defense contractors and make it available to all disciplines for research and education. Dennis Jennings was responsible for the initial decision to use TCP/IP to build a network linking the NSF-sponsored super computer sites. In 1986, the NSF brought in Steven Wolff to oversee construction of a new research network (Leiner, et al., 1994). Wolff was charged with developing a computer network that would facilitate cooperative research and that would eventually outgrow its need for NSF funding. In order to do this, Wolff created policy designed to encourage broad-based use of the Internet while encouraging the creation of separate, commercial, long-haul networks. The NSF required universities seeking funding for an Internet connection to make connections available to all qualified users on campus. Further, they encouraged growth of regional networks by urging regional networks to seek commercial, nonacademic customers in order to expand use of the networks, take advantage of increased economies of scale, and ultimately lower subscription costs. They also wrote the Acceptable Use Policy, which restricted use of the national NSF network to academic and research work (National Science Foundation, 1986). It was believed that denying commercial users access to the national network would encourage creation of nationwide commercial networks. This strategy paid off when commercial Internet long-haul carriers UUNET and PSI opened.

In 1986, the NSF launched the NSFNET backbone, a high-speed (56 kbs) network connection between six supercomputer centers running across the United States. It was built by MCI, IBM, and MERIT (the University of Michigan). The supercomputer centers were at Princeton, Cornell, Carnegie-Melon, the University of Illinois, the University of Colorado, and the University of California, San Diego. In under nine years, the NSFNET backbone grew from 6 nodes at 56 kbs to 21 nodes at 45 mbs, connecting 50,000 networks on all seven continents and in space.

FINDING OURSELVES IN VIRTUAL SPACE

A network with the breadth and reach of the Internet quickly surpassed users' ability to keep track of the location of information. With 50,000 networks, keeping track of files and IP addresses quickly became impossible. Soon computer scientists began writing software to help users find and transfer files. One of the most popular was Gopher, released in 1991 by Paul Lindner and Mark P. McCahill from the University of Minnesota. Gopher, and its descendents, allowed users to search through "Gopher space" for files to transfer via file transfer protocol (FTP). Gopher was an Internet protocol that provided menu-driven file-and-data retrieval from remote computer servers (Anklesaria, et al., 1993). Gopher sites organized files for retrieval and were set up to be searched by users. Although they were text based, these early versions of internet searching software greatly simplified the problem of finding files on the Internet.

All this was soon to change. Desktop computers were becoming easier to use because of graphical user interfaces, developed at the Xerox Palo Alto Research Center (PARC), and the commercially distributed Apple Macintosh operating system, in 1984. One year after the introduction of the Macintosh operating system, Microsoft released the Windows 1.0 operating system in an attempt to replace MS-DOS text commands with an intuitive point-and-click graphical user interface similar to the Macintosh operating system, although the Windows operating system would not become commercially popular until the introduction of version 3.0 in 1990 (Microsoft, 2002). Computing power that the early Internet pioneers could only dream of had found its way to business desktops, elementary school classrooms, and home recreation rooms. Access to the Internet was increasingly supplied to the general public through commercial gateways and public access points, such as Michigan's MERIT network and Cleveland's Freenet. The number of hosts on the Internet had grown exponentially through the 80s (see Figure 3). But the biggest change in the Internet was still to arrive.

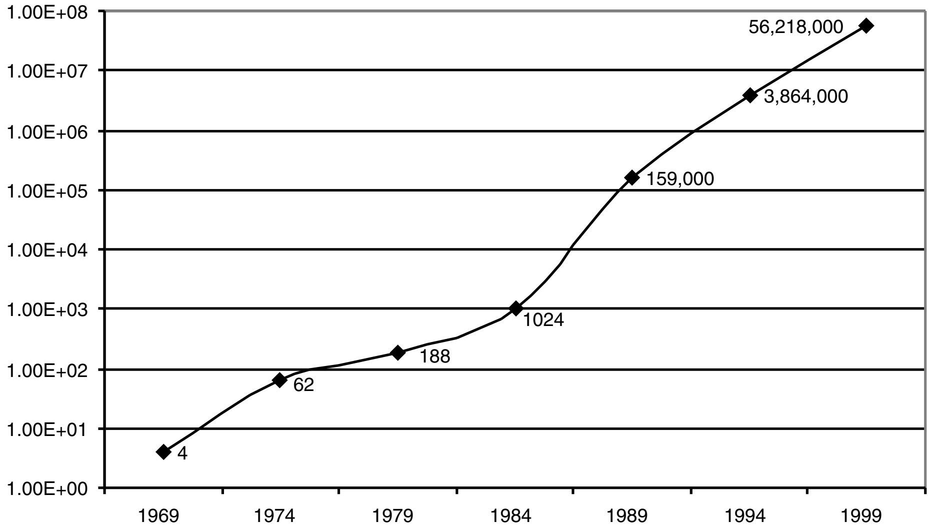


Figure 3: Exponential growth of Internet hosts.

The World Wide Web

In the late 1980s, Tim Berners-Lee was a research fellow at the European Laboratory for Particle Physics (CERN) in Switzerland. He had become intrigued with using computers to organize information via hyperlinked connections. As his ideas developed, his vision fixed on a scalable system for linking related documents via the Internet. It was important that such a system not use a central database to keep track of files, as updating the database would limit its speed and size to available resources (because of computer storage limits and the human resources needed for updating the database). Instead, Berners-Lee wanted all system users to be able to both find and contribute information. Administrators at CERN had a difficult time understanding what use they could make of Berners-Lee's invention and would not provide financial support for him to pursue the idea. Nonetheless, he persevered and finally received financial support in the form of a new NeXT computer.

Steve Jobs was forced out of Apple in 1985 and started up NeXT computers. Jobs felt his computers were the "next" great advance after the Macintosh for desktop computers. Berners-Lee found that much of what he needed to create his World Wide Web (WWW) hyperlinking software was available on his NeXT computer. He developed an early prototype WWW browser to make the CERN phone directory available online. From this modest success, he continued to promote the use of the WWW both inside and outside of CERN. He traveled to the United States to display the WWW at a conference of hypertext developers, posted information about the WWW on Internet user groups, and provided telnet access to the browser on a server at CERN so that people could try it out.

Berners-Lee's (1999) vision was that the WWW would maximize flexibility while accommodating standards that were already in play on the Internet. He wanted the WWW to be open to existing Internet protocols, so he designed it to read FTP, Gopher, network news (NNTP), wide area information server (WAIS), and other documents. This was accomplished, in part, by specifying the protocol at the beginning of the link address or uniform resource locator (URL). For example, links to FTP files began with "ftp://. . . " whereas links to Gopher sites began "gopher://. . . ." His own protocol for linking hypertext documents is the now familiar hypertext transfer protocol (HTTP). This software provided instructions on how to download and connect hypertext documents via the Internet. Finally, Berners-Lee standardized the language that Web page designers would use, called hypertext markup language (HTML).

The First Browsers and the Rise of Popular Internet

Despite the obvious advantages of the WWW, which we see in retrospect, few in the computing world were eager to adopt this new technology (Berners-Lee, 1999). Although Web browsers were developed for operating systems commonly used by computer programmers (e.g., UNIX), they were not available for the most commercially popular operating systems (PC, Mac), and Web editors were even more rare. Without a browser that was easy to install and without abundant content to discover, the Web was unlikely to take off. In 1993, Marc Andreessen and colleagues at the University of Illinois National Center for Supercomputing Applications (NCSA) released Mosaic, a web browser that was eventually available for UNIX, Macintosh, and PC. The immediate advantage of Mosaic was that it was graphical, user friendly, and easy to install. Mosaic caught on quickly, thanks in part to an aggressive public relations campaign by the University of Illinois. WWW use increased at a rate of over 340,000% in the first year. Soon new technologies proliferated on the WWW. By the end of 1995, real-time audio streaming was made available by RealAudio, RT-FM had begun netcasting from Las Vegas, the Arizona law firm Canter & Siegel introduced "spam" to the Internet by mass e-mailing of advertising for immigration services, banner ads were introduced, Sun Microsystems introduced JAVA, and virtual reality markup language (VRML) made its appearance on the Web.

In 1994, Mark Andreessen left the NCSA to form Netscape Communications with Jim Clark. Netscape soon produced its own version of the Mosaic browser, making it available for free to educational institutions and for a small fee to others wanting the use the WWW. Microsoft was slow to realize the potential of the WWW but aggressively joined the Web when Bill Gates decided to bundle a Web browser with the new release of Windows 95. Rather than write their own code, Microsoft bought the rights to a NCSA spin-off browser from a small company called Spyglass for $2 million (Berners-Lee, 1999). This marked the beginning of what became known as the "browser wars" between Netscape and Microsoft. Sensing competition in the market that they thought they controlled, Netscape urged the Justice Department to pursue antitrust action against Microsoft for bundling the browser with the most popular desktop operating system. Versions of this case, particularly in an antitrust suit brought by the U.S. Department of Justice, continue to be tied up in the courts (CNN, 2002).

In the years following 1995, the Internet, WWW, and related technologies became part of the public consciousness. Magazines, newspapers, and television programs increased coverage of this new technology. Debates erupted on Capitol Hill in Washington, D.C., about what content could be allowed on the Web. Maybe most important for the continued development and commercialization of the Internet, Congress passed the Telecommunication Act of 1996, opening up media and telecommunications companies to increased competition. People became troubled by such security issues as "hacks" and "viruses" even as they slowly began purchasing products from "e-tail" outlets, such as Amazon.com, and auctions sites, such as eBay.com.

CONCLUSION

The precursors of the Internet have outlived their usefulness. The ARPANET shut down in 1990 and the original NSFNET backbone followed in 1995. The NSF created the vBNS (very high performance backbone network service) through MCI in 1994 and it was used to link supercomputer centers sponsored by NSF. Nonetheless, there remains a place on the Internet restricted from trade and dedicated to researching and developing new communication technologies. Internet2 is a project of the University Consortium for Advanced Internet Development (UCAID) and it uses both the vBNS network and the Abilene Network from QWEST to achieve desired flexibility. Internet2, the very-high-speed backbone network, was created in 1996 and limited to research only. Internet2 continues to expand the horizons of new technologies, introducing richer communication experiences via three-dimensional immersive environments.

The Internet was born of the excitement of solving complex engineering problems in an open and collaborative manner. This led to an open-architecture design that has been flexible enough to support exponential growth in hosts and users as well as technological innovation in software. With a minimum of governance, the Internet has grown to be the information infrastructure for the 21st century, supporting exchange of audio and video files, animations such as Shockwave Flash, three-dimensional interfaces such as VRML, media wrappers such as QuickTime, chat rooms, and instant messaging. It is difficult to imagine where innovations on the Internet will take us, but it is clear we will never see the Internet in its final form. If future designers adhere to the simple rules of the Internet pioneers—an open architecture built through cooperation—we will likely never see the Internet stop evolving.

GLOSSARY

Advanced Research Projects Agency Network (ARPANET) Created by the U.S. Advanced Research Agency (ARPA) in 1969. ARPANET was a wide area network linking computers at various university research centers. ARPANET is considered the precursor to the modern-day Internet.

Analog Electronic transmissions that represent data by continuous signals of varying frequency (versus being sent as an "on or off" data transmission; see *Digital*).

Circuit switching A communication technology in which a dedicated path is established between the source and destination for the duration of the transmission. The telephone network is an example of a circuit-switching network.

Digital Electronic transmissions that represent data in various combinations or strings of 1's and 0's, with 1 representing "on," or "positive, and 0 representing "off," or "nonpositive."

Host A computer system containing data that is accessed by a user at a remote location.

Interface message processors (IMP) Smaller, intermediate computers connected to the mainframes involved in ARPANET. The IMPs were packet switches dedicated to the communications "subnet."

Internet A vast system of linked computer networks, international in scope, that facilitates data transfer and communication services, such as remote log-in, file transfer (FTP), electronic mail (e-mail), newsgroups, and the World Wide Web (www); always capitalized.

Intranet Internal network utilized within organizations allowing file sharing and collaboration. Intranets cannot (generally) be accessed from outside the organization.

IP address A numeric address that identifies host computers on the Internet.

Local area network (LAN) A network that connects computers in a relatively small geographical area (generally no more than 2 miles). LANs are used in offices, buildings, or a set of buildings.

Network A collection of two or more computers (and/or printers, routers, switches, or other devices) connected to one another allowing data to be shared and used over communication paths.

Network control protocol The host-to-host protocol of the ARPANET that operated above the layer of packet switching implemented in the IMPS. Once ARPANET transitioned from NCP to TCP/IP, it was possible to connect to other independent networks.

Node A connection point in a network, either for redistribution of data or as an end point for data transmissions.

Packet switching Method used to efficiently move data around on a network. Data is divided into smaller pieces called packets, assigned a number and destination, sent independently through the network, and reassembled at the other end.

Protocol Formal set of rules that govern how devices on a network exchange information; the "language" of a network.

Request for Comment (RFC) Document series used among researchers as the primary means of communicating ideas and information about the Internet.

Router The device that forwards packets between networks.

Transmission control protocol/Internet protocol (TCP/IP) The common language that allows existing networks to be connected, or "inter-networked." TCP implements flow control and recovery in packet switching, and IP is used for addressing and forwarding individual packets. TCP runs over IP and the two form part of a layered protocol stack.

Uniform resource locator (URL) A character string describing the location and access method of a resource on the Internet.

World Wide Web Invented in 1991 by Tim Berners-Lee, the World Wide Web is an application that facilitates exchange of hypertext documents via the internet. It uses a variety of protocols including HTTP, HTML, FTP, VRML, XML, and SSL.

CROSS REFERENCES

See *Circuit, Message, and Packet Switching; Internet Literacy; Internet Navigation (Basics, Services, and Portals); Standards and Protocols in Data Communications; TCP/IP Suite.*

REFERENCES

Abbate, J. (1999). *Inventing the Internet*. Cambridge, MA: MIT Press.

Anderberg, A. (2002). *History of the Internet and Web*. Retrieved August 1, 2002, from http://www.anderbergfamily.net/ant/history.

Anklesaria, F., McCahill, M., Lindner, P., Johnson, D., Torrey, D., & Alberti, B. (1993). *RFC 1436—The Internet Gopher protocol*. Retrieved August 30, 2002, from http://www.ietf.org/rfc/rfc1436.txt

Baran, P. (1964). *On distributed communications networks*. Retrieved August 4, 2002, from http://www.rand.org/publications/RM/baran.list.html

Berners-Lee, T. (1990). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor.* San Francisco: Harper.

Bush, V. (1945). As we may think. *The Atlantic Monthly, 176* (1), 101–108. Retrieved August 14, 2002, from http://www.theatlantic.com/unbound/flashbks/computer/ bushf.htm

Cerf, V. G., & Kahn, R. E. (1974). A protocol for packet network interconnection. *IEEE Transactions on Communication Technology, 5*, 627–641.

CNN (2002). *Netscape sues Microsoft*. Retrieved August 14, 2002, from http://money.cnn.com/2002/01/22/technology/netscape

Crocker, S. (1969, April 7). *Host software: Network working group request for comment 1*. Retrieved August 6, 2002, from http://www.faqs.org/rfcs/rfc1.html

Davies, D. (1966). *Proposal for a digital communication network*. Unpublished manuscript.

Davies, D. W., Bartlett, K., Cantlebury, R., & Wilkinson, P. (1967). *A digital communications network for computers giving rapid response at remote terminals*. Paper presented at the Association for Computing Machinery Symposium on Operating Systems Principles, Gatlinburg, Tennessee, October.

Hafner, K., & Lyon, M. (1996). *Where wizards stay up late: The origins of the Internet*. New York: Simon & Schuster.

Internet Society (1999). *Request for comments 2555: 30 years of RFCs*. Retrieved August 3, 2002, from ftp://ftp.isi.edu/in-notes/rfc2555.txt

Kahn, R. E. (1994). The role of the government in the evolution of the Internet. In *Revolution in the U.S. information infrastructure*. Washington, DC: National Academy of Science. Retrieved February 17, 2003 from http://www.nap.edu/readingroom/books/newpath/chap2.html

Kleinrock, L. (1961). Information flow in large communication nets. Unpublished doctoral dissertation, Massachusetts Institute of Technology. Retrieved February 17, 2003 from http://www.lk.cs.ucla.edu/LK/Bib/REPORT/PhD/part1

Kleinrock, L. (1964). *Communication nets: Stochastic message flow and delay.* New York: McGraw–Hill.

Kleinrock, L. (1976). *Queueing systems. Vol II. Computer Applications*. New York: John Wiley & Sons.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (1994). *A brief history of the Internet*. Retrieved May 17, 2002, from http://www.isoc.org/internet/history/brief.shtml

Licklider, J. C. R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics, HFE-1*, 4–11. Retrieved May 10, 2002 from ftp://gatekeeper.research.compaq.com/pub/DEC/SRC/research-reports/SRC061.pdf

Licklider, J. C. R., & Clark, W. E. (1962). On-line man-computer communication. *Proceedings of the AFIPS SJCC, 21*, 113–128.

Marill, T., & Roberts, L. G. (1966). Toward a cooperative network of timeshared computers. *AFIPS Conference Proceedings FJCC, 29*, 425–432.

Microsoft (2002). *Windows operating systems family history*. Retrieved August 30, 2002, from http://www.microsoft.com/windows/WinHistoryIntro.mspx

Mockapetris, P. (1983a). *RFC 882—Domain names: Concepts and facilities*. Retrieved August 30, 2002, from http://www.ietf.org/rfc/rfc882.txt

Mockapetris, P. (1983b). *RFC 883—Domain names: Implementation and specification*. Retrieved August 30, 2002, from http://www.ietf.org/rfc/rfc883.txt

Moschovitis, C. J. P., Poole, H., Schuyler, T., & Senft, T. M. (1999). *History of the Internet: A chronology, 1843 to present.* Santa Barbara, CA: ABC-CLIO.

National Science Foundation (1986). *NSF network news*. Retrieved August 30, 2002, from http://www.cni.org/docs/infopols/NSF.html

RFC-Editor Homepage (n.d.). *The request for comments*. Retrieved August 2, 2002, from http://www.rfc-editor.org

Segaller, S. (1998). *Nerds 2.0.1: A brief histroy of the Internet*. New York: TV Books.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460.

Zakon, R. H. (2002). *Hobbes Internet timeline v. 5.6*. Retrieved May 10, 2002, from http://www.zakon.org/robert/internet/timeline

HTML/XHTML (HyperText Markup Language/ Extensible HyperText Markup Language)

Mark Michael, *King's College*

INTRODUCTION

The World Wide Web Consortium's (W3C) claim that "HTML is the *lingua franca* for publishing hypertext on the World Wide Web" (W3C, 2003b) is no idle boast. In fact, by definition, the Web (a specific aspect of the Internet) did not exist until the introduction of the Internet's most famous duo: *HyperText Markup Language* (*HTML*) for structuring and formatting documents without the use of special characters and the *Hypertext Transfer Protocol* (*HTTP*) needed to transport HTML documents across the Internet.

HTML rapidly outgrew its original, limited purpose, viz., to display the text of Web pages and provide the ability for the viewer of one document to link to another. A great number of capabilities have been added to its repertoire. This growth was more chaotic than that for most languages. While Web development is not in quite the state of disarray that it once was, it still suffers from having to placate browsers with different methods of rendering HTML documents.

HTML has even outgrown the Web. It is also used as a format for e-mail messages and as a vehicle for displaying information about a computer's file directory. At the same time, the Web has outgrown HTML. New technologies have outstripped HTML's ability to adapt.

W3C is vigorously attempting an HTML renaissance. The reborn edition, the *Extensible HyperText Markup Language* (*XHTML*), is actually a family of languages and modules. Recommendations on these and supporting applications are coming out at such a fast clip that it is impossible to predict what the state of affairs will be by the time this is in print. What is clear is that HTML has reached the end of its usefulness as an all-purpose markup language for the Web, and the torch has been passed to XHTML.

Consequently, we will focus on XHTML. What we say about HTML is intended to provide a better understanding of XHTML as it stands at the moment this is written. Likewise, what is written here is merely a foundation for understanding XHTML in whatever incarnation it exists when this is read.

THE HISTORY OF HTML AND XHTML

To the extent that a sequence of distinct versions of a language can be considered a continuum, XHTML must clearly be seen as part of the continual evolution of HTML. W3C views XHTML as a family of document types; the HTML document types are in the extended family. W3C acknowledges this continuity by giving an overview of the entire HTML/XHTML lineage (except HTML 1.0) in a single Web page (W3C,2003b).

In recent years, the XHTML family tree—and the larger neighborhood—has begun to look more like a complex braiding of threads, with modularization coming between versions; extensions such as MathML preceding that; Document Object Model (DOM), Cascading Style Sheets (CSS), and Extensible Markup Language (XML) evolving in parallel; and alternative representations (XML Schemas) being thrown in as well.

The Genesis and Evolution of HTML

The inspiration for HTML can be found in the Standardized General Markup Language (SGML), which had been used for coding the structure of electronic documents since 1986. Tim Berners-Lee led work at the Conseil Européen pour la Recherche Nucléaire (CERN) in Switzerland to develop a simpler scheme—HTML—which also had the ability to include hyperlinks. The protocol—HTTP—needed to send HTML documents across the Internet was developed concurrently. CERN launched the markup language and the protocol in 1991, thus creating the World Wide Web.

HTML 1.0 was designed purely for rendering text-based documents in text-only browsers. Supported elements included: `title`, paragraph (`p`), headings (`h1` to `h6`), definition lists, unordered lists, and, most importantly anchor (`a`), for linking between documents. The characters `<`, `>`, and `&` were also part of the alphabet for the language.

The World Wide Web Consortium was formed in 1994 to provide guidance with regard to Web-related matters. Its early life saw its leadership role usurped by others, perhaps due to its initial slowness to act.

It was not until 1996 that W3C approved HTML 2.0, despite the fact that a successor to that version was proposed in 1993! It was only at this point that the `html`, `head`, `body`, `base`, `link`, `meta`, and `form` elements were officially added to the language. Ordered lists, inline elements (`b`, `cite`, `code`, `em`, `kbd`, `samp`, `strong`, `u`, and `var`), block elements (blockquote and pre), and breaks (`br` and `hr`) were also added. *Character entities* (e.g., `&copy;`) could now be used to render special characters (e.g., ©). Perhaps the biggest improvement was that graphic files could be rendered by browsers such as Mosaic and Cello thanks to the `img` element. This version possessed much of the functionality required of static Web pages.

The belated endorsement of HTML 2.0 was actually at the end of the fiercest period (1993–1996) of the "browser wars." The competition between newcomer Microsoft Internet Explorer and Netscape Navigator (later Communicator), which had replaced Mosaic as the dominant browser, created both chaos and a slew of innovations (some of which were eventually adopted by W3C) as each browser supported features not in the W3C recommendations. Despite the predicament this caused, Web developers nevertheless frequently chose to "optimize" their code for one or the other browser rather than code for the least common denominator. Thus, HTML documents contained elements, attributes, or values supported by only one browser, and "best viewed with" messages appeared at the bottom of many pages.

HTML 3.2, the next official W3C recommendation, came in January 1997. (HTML 3.0 was a proposal submitted to W3C but not endorsed.) This represented W3C's mostly successful attempt to regain control of the situation. Important new elements were `style`, `script`, and `div.` Inline elements added were `big`, `small`, `sub`, and `sup.` Other new elements (which are now on the way out) were `applet`, `center`, and `font`. *Image maps* were introduced to allow *hotspots* within images to act as hyperlinks. Another useful addition was the ability to specify colors in the `#RRGGBB` format, in which three pairs of hexadecimal digits represent the intensities (0–255, in decimal) of the red, green, and blue components of an additive color.

HTML 4: The Final Version

HTML 4.0 was released by W3C in December 1997. New elements were `object` (in place of Internet Explorer's `embed`), `abbr`, `acronym`, `colgroup`, `col`, `tbody`, `tfoot`, and `thead`. The `font` element was deprecated while support for style sheets was improved. The addition of the `<!DOCTYPE>` statement allowed specifying which Document Type Definition (DTD), i.e., version, of HTML 4 applied to the document. Documents could also be written in any language.

An important change was the introduction of event attributes, which could, for instance, invoke JavaScript code. These can be broken down into four categories:

1. window events (for `body` and `frameset` elements)—`onload` and `onunload`;
2. events only for `form` elements or their children—`onchange`, `onsubmit`, `onreset`, and `onselect`;
3. keyboard events (for `block` and `inline` elements)—`onkeydown`, `onkeypressed`, and `onkeyup`; and
4. mouse events (for `block` and `inline` elements)—`onclick`, `ondblclick`, `onmousedown`, `onmousemove`, `onmouseover`, `onmouseout`, and `onmouseup`.

The final W3C recommendation on HTML, HTML 4.01, was released on December 24, 1999. This was merely a clarification of some attributes and the removal of `mailto` (properly a protocol to be included in a URI) as a possible value of a `form` element's `action` attribute. ISO HTML, the International Organization for Standards/International Electrotechnical Commission's standardized subset of HTML 4.01, was released as ISO/IEC 15445:2000 in May 2000 (ISO/IEC, 2000). The first nail had already been hammered in HTML's coffin by January 2000, when W3C released its first proposed recommendation for XHTML.

In looking back at its meteoric career, we see that HTML was originally conceived to manage content. It evolved to manage presentation, but a better approach arose: Cascading Style Sheets (the subject of a separate chapter).

The lack of clear separation of content and presentation was accompanied by a lack of well-defined structure. These failings prevent applications from searching and referencing documents based on their semantics.

More important, HTML does not have the flexibility to accommodate future or even current technologies. On one hand, it does not allow high-powered platforms to realize their potential to render complex content. On the other hand, it is too bloated for platforms with limited (if any) screen space and modest computing power. Ultimately, it was the inability of HTML to further adapt that forced its dramatic metamorphosis.

XML: A New Direction

To understand XHTML, it is necessary to first take a step back.

In 1996, even while W3C was struggling to catch up with the browsers, it began a separate project: the *Extensible Markup Language,* the subject of a separate chapter. This was to be a markup language as powerful as but simpler than SGML.

XML is a *metalanguage* in that it provides a general structure for defining markup languages. In particular, XML defines what are elements and attributes, but leaves it to a user agent (a browser, e-mail program, etc.) to interpret what each element and attribute means.

Superficially, the syntax of XML resembles HTML in that there are tags (the identifiers for an element enclosed in angle brackets) and attributes, which are assigned values using an equal sign.

Structurally, XML, like a programming language, demands perfect grammar. Browsers had a long tradition of rendering imperfect HTML pages, often by correctly reading the mind of the page's author. This mind reading was inconsistent across browsers: one would render `&copy` as the copyright symbol despite the missing semicolon, but another would display the characters as text. Yet developers got away with a great deal of sloppiness. An XML-based user agent is expected to reject a malformed XML document rather than attempt to mind read.

The two versions of XHTML that have reached Recommendation status as this is written (1.0 and 1.1) are based on XML 1.0, which is defined in (W3C, 2000a). (Later, we will see this reflected in the first line of code in Table 1.)

The Advent of XHTML

XHTML has been variously described as

1. HTML + XML;
2. a merger, marriage, or hybrid of HTML and XML;
3. a step toward or transition to XML;
4. the next phase or generation of HTML.

Since XHTML "is both XML-conforming and, if some simple guidelines are followed, operates in HTML 4 conforming user agents" (W3C, 2002a), it is tempting to think of XHTML as a kind of intersection of HTML and XML.

However, XML does not directly provide a language specification for XHTML. Rather, XML is a general framework for defining an unlimited number of different languages, of which XHTML is just one. The most accurate statement is that made by W3C: "XHTML is a reformulation of HTML as an XML application" (W3C, 2002a). This does not mean that XHTML is software that uses XML (as a browser might). Rather, it is a set of tags (specified in a Document Type Definition or an XML Schema) that applies XML to a particular domain. Some domains might be highly specific, such as vector graphics or mathematical notation for SVG and MathML, respectively (discussed later). XHTML is a general-purpose markup language whose domain encompasses much of what needs to be done in Web pages.

Flexibility is the primary impetus for using XML as a foundation for specifying how Web content should be marked up. The Web needs to be able to adapt to ever-changing technology, which is developing greater capabilities while also branching off into more different platforms, which often have very restricted resources. Thus, the needed flexibility comes in two forms: (1) the ability to contract XHTML and (2) the ability to extend it by allowing its integration into other electronic documents or the inclusion of other XML applications in XHTML documents. The initial W3C recommendation, XHTML 1.0 (which, we will see, is actually three different recommendations), is just a first step toward modularizing the language specification so that parts can be subtracted from or added to the language to fit a particular platform. It is a bridge to the XHTML 1.1 recommendation, which incorporates modularization. We will concentrate on XHTML 1.0 and comment on the changes in XHTML 1.1.

A second motivation for moving to XHTML is to separate more sharply content from the presentation of that content. To that end, a number of features of HTML 4 have been deprecated. In some cases, an entire element has been eliminated; in other cases, an attribute of an element has been eliminated, with equally profound consequences.

A third goal of XHTML is to enforce the rules of well-formed code. Traditionally, browsers have simply ignored tags they did not recognize; no error messages would display. If a tag was malformed, e.g., `<html<`, the browser would attempt to render it as text (in whatever the current font, size, and color was). In the extreme case, an "HTML document" might contain no HTML code at all! The smallest Web page that can be displayed is a single ASCII character, which a well-meaning browser simply displays as text in a default font. Years of living in this somewhat lawless environment have encouraged poor habits, which must now be unlearned. An XHTML-conforming user agent must verify the well-formedness of a document.

XHTML 1.0 FUNDAMENTALS

Despite W3C's thoughtful work (already several years old), the message to migrate to XHTML has not been widely embraced. A vast amount of new HTML code is written or generated each day. Here we take the forward-looking approach: presenting the rules for writing valid XHTML code and dealing in a later section with updating HTML legacy code.

Syntax

Since XML is case-sensitive and XHTML is an XML application, all XHTML identifiers must appear in all lowercase letters. This is the opposite of the all-uppercase tradition in writing HTML code. HTML is actually case-insensitive, so that`<BR>`, `<Br>`, `<bR>`, and `<br>` should be interpreted equivalently; but only the last version is valid in XHTML.

Another implication is that the ampersand (`&`), left angle bracket (`<`), and right angle bracket (`>`) characters cannot be used as character data. When not intended as markup characters, they must be encoded as `&amp;`, `&lt;`, and `&gt;`, respectively. Likewise, the apostrophe/single quotation mark (`'`) and double quotation mark (`"`) should be encoded as `&#39;` (XML's `&apos;` is not supported by older browsers) and `&quot;`, respectively, within attribute values using the same delimiter. Furthermore, the form feed character (U#000C in Unicode, decimal code 12 in ASCII) is illegal.

A variety of special symbols can be displayed by means of mnemonic or numerical codes. One of the most useful is `&nbsp;` for a *nonbreaking space,* a blank space which will be rendered despite a browser's *normalization* of character data by ignoring excess whitespace; this is an essential tool for forcing an empty cell in a table to be outlined in the same way as a nonempty cell.

An HTML/XHTML *element* has an *opening tag*, which is the element's type enclosed in angle brackets (< >). Elements must be closed in the sense that an opening tag must be matched by a *closing tag*. The latter is signified by a forward slash followed by the element's type, as in

```
<p>The text of a paragraph.</p>
```

An exception is empty elements. Certain elements, such as `br` (a forced line break), would never contain any content. The opening and closing tags of such elements (`<br></br>`) can be combined into a single tag, which contains a forward slash after the element's identifier, as in `<br/>`. A blank space should be provided before the forward slash (`<br />`) so that older browsers will correctly recognize the element. (On the other hand, the intentional use of `<br/>` with no blank space is a kluge that can sometimes compensate for variations in how different browsers provide vertical spacing, for example, after tables.)

All elements may have certain *attributes*. Some elements have required attributes. Attributes reside in the opening tag, following the element's type and a blank space. Multiple attributes are separated by blanks. Unlike in HTML, attributes in XHTML must always be followed by an equals sign (=) and a value, which is enclosed in matching single or double quotation marks. Some attributes can take on an unlimited number of values, some can only take on a small number of specified values, and some are termed *Boolean* (an abuse of terminology). In HTML, an attribute of this last type would either be present or absent in the code; for example, a check box on a form is either `checked` or it is not. This was referred to as *attribute minimization*. It might seem logical for XHTML to demand a value of `"true"` be assigned to any Boolean attribute present. Instead, the required value is the very same reserved word as the attribute, but enclosed in quotation marks (e.g., `checked = "checked"`). Thus, a name may be both an attribute and a value.

In a few cases, the same identifier may represent an element in one context and an attribute in another. An example is `title`. As an element, it provides text that a browser may display in its title bar. As an attribute of a block or inline element, it can provide text that a newer browser may display in a transient balloon when the user's mouse cursor pauses over the element so attributed.

Comments in both HTML and XHTML are stored within a special kind of tag, as in

```
<!-- These comments will be ignored. -->
```

The enclosed text will not be processed by the user agent interpreting the document. The delimiters can enclose forced line breaks, so a single comment may encompass multiple lines. A comment should not, however, contain consecutive hyphens (`--`).

A traditional (optional) use of comments is to shield scripts from applications that do not handle them. An example of this format is

```
<script type="text/javascript">
<!--// start hiding script from JavaScript-
  challenged browsers
   // JavaScript code goes here
// end of hidden script -->
</script>
```

Applications that do interpret scripts will disregard the apparent cue (`<!--`) to ignore the code between the opening and closing script tags. JavaScript implements the end-of-line (`//`) form for comments, used also in C++. The final JavaScript comment hides from the JavaScript interpreter the closing of the XHTML comment. Note the recommended blank space before `-->`.

This technique is now discouraged for XHTML. Instead, scripts (including internal style sheets) should be contained in a CDATA (character data) section to indicate to the user agent that it should not be parsed as XHTML. A sample section is

```
<![CDATA[ character data here is not parsed
  as XML ]]>
```

But the delimiters of the character data section (`<![CDATA[ and ]]>`) should in turn be hidden as comments from the viewpoint of the application interpreting the character data; C-style comments (beginning with `/*` and ending with `*/`) will work with both JavaScript and CSS. The safest course of action is to use a `link` element to reference an external file that contains the script.

Document Type Definitions (DTDs) and XML Schemas

Any XML document conforms to the specifications within a *Document Type Definition* (*DTD*). This is a declaration, using an Extended Backus-Naur Form (EBNF) grammar, of the valid elements, the appropriate attributes for these elements, and the acceptable range of values for attributes.

Corresponding to the three HTML 4 DTDs are three XHTML DTDs with the same names: XHTML 1.0 Strict, XHTML 1.0 Transitional, and XHTML 1.0 Frameset. For simplicity, we will henceforth refer to these DTDs as *Strict, Transitional,* and *Frameset,* respectively. As the names suggest, the first of these is the most restrictive, the second is seen as temporary in nature, and the third supports frames (which, by implication, are not endorsed for the long term). We will restrict our attention almost entirely to Strict since it is the basis for the only (admittedly modularized) XHTML 1.1 DTD.

Unlike the situation with standards organizations, all publications of the W3C are technically considered to be recommendations. Important in the history of the Internet has been the disregard for these recommendations. In particular, competition between Microsoft Internet Explorer and Netscape Navigator saw each browser supporting extensions to the recommendations then in effect; some of those innovations died on the vine (such as Netscape's `layer` element) while others were eventually incorporated in future W3C recommendations.

An alternative to DTDs is XML Schemas. Unlike a DTD, an *XML Schema* is itself an XML document. Consequently,

Table 1 Sample Minimal XHTML Documents with No Content

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
  <head>
    <title></title>
  </head>
  <body>
  </body>
</html>
```

(a) for XHTML 1.0 Strict

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Frameset//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml-frameset.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
  <head>
    <title></title>
  </head>
  <frameset>
  </frameset>
</html>
```

(b) for XHTML 1.0 Frameset

it can be manipulated like any other XML document. Each approach requires validating parsers (so well-formedness will be enforced). XML Schemas have the advantage of being able to validate the contents of an element to verify that it consists of the expected type of data. On the other hand, an XML Schema cannot accommodate character entities. However, a Web page can make use of both a DTD and an XML Schema (see W3C, 2002b). Some expect XML Schemas to eventually supplant DTDs.

Document Structure and the Document Object Model (DOM)

Table 1 contrasts sample valid Strict and Frameset documents with no content.

In each case, the first line of code is an XML declaration indicating the XML version and the encoding used. The encoding is a particular mapping of characters in a natural language to a sequence of zeroes and ones. UTF-8, the default XML encoding, is highly recommended for the Internet since it enjoys the widest browser support. Using 1–4 bytes per character, it includes the standard set of 128 ASCII characters as well as Unicode (a registered trademark of the Unicode Consortium), which encodes the characters of most of the world's major languages.

Technically, an XML document does not require this XML declaration if encoding is UTF-8 or UTF-16. However, the inclusion of the XML declaration may cause problems for some browsers parsing a document as HTML; if this is expected to be a problem, the `<?xml>` tag should be omitted, and the character encoding should instead be indicated in a `meta` element (described later).

The second line of code is a `DOCTYPE` declaration that specifies the root element of the DTD (see later), information about the DTD and its owner (W3C), and finally the DTD's URI. See Sauers & Wyke (2001) for more details.

The third line of code is the opening tag for the HTML document. Its first attribute specifies an XML namespace by means of the URI of the particular file which defines the allowable element types and attribute names. In this case, the namespace applies to the entire document. Different portions of an XML document could have different namespaces, allowing the same name to have different meanings in each portion. In such cases, a namespace prefix can be used to allow unambiguous use of a name and its homonym from a different namespace. For details on XML namespaces and how they differ from conventional namespaces, see W3C (1999).

The second and third attributes in the opening HTML tag specify (in two different ways) a two-letter code for the natural language in use (e.g., `"en"` for English).

The *root element* of a well-formed XHTML document must be an `html` element. This must have two child elements. The first, a `head` element, will hold information about the document. The second part of the document will be one of two types. A typical document, one conforming to Strict or Transitional, will have a `body` element to hold all the content to be rendered. An occasional document will merely serve to set the stage for several documents to be rendered at once. It will have a `frameset` element and conform to Frameset. (As explained later, this element may have a grandchild `body` element and all that it could ordinarily contain.)

The indentation shown is not required, as both HTML and XHTML ignore excess whitespace, e.g., the second and subsequent blanks in a sequence of consecutive blanks. (However, nonessential whitespace within a tag may have unexpected consequences, such as preventing the pieces of a subdivided image from being reassembled snuggly in a table.) It is used here to indicate the structure of the document, that is, how elements are nested within one another. This time-honored tradition from programming was ignored by most Web authors, but is now making a comeback, perhaps encouraged by XHTML's demand for proper structure.

While well-formed HTML should also possess proper structure, XHTML has higher standards. For one, all elements must nest properly: if two elements overlap, one must be entirely contained within the other. Suppose one wanted to have a combination of italicized and boldfaced text of the form:

This is italicized ***and*** **boldfaced.**

It would be illegal in XHTML to script this with two inline elements as in

```
<i>This is italicized <b>and</i>
  boldfaced.</b>
```

Instead, one would need to use three inline elements, with one nested in another:

```
<i>This is italicized <b>and</b></i>
  <b>boldfaced.</b>
```

There are also strict limitations on which elements can nest in which. XML-based languages cannot specify in their DTDs exclusions of certain nestings. However, the following rules must be observed:

1. An a element cannot nest within another a element.
2. A pre element cannot contain any big, img, object, small, sub, or sup element.
3. A button element cannot contain any button, fieldset, form, iframe, input, isindex, label, select, or textarea element. (The iframe and isindex elements are not included in Strict in any case.)
4. A label element cannot nest within another label element.
5. A form element cannot nest within another form element.

Table 2, which recasts a table from Graham (2000) in a format that corresponds to how code would be written, contrasts the global structure of Strict and Frameset documents. Note that the latter type may have a body element as a great-grandchild element, but not as a child element of the root (html) element.

Table 2 also illustrates the nesting of block elements (discussed later) within the body element. It also shows the elements which can appear within the table, form, and list (dl, ol, ul) elements. Tables would seem to be self-explanatory, but their use and misuse are sufficiently important to merit a separate section later. Details about the form and list blocks' child and grandchild elements will not be included here. However, they are important blocks. Forms provide a way for users to interact with a Web page by submitting data. Lists may consist of unordered (bulleted) or ordered (numbered or lettered) items or may consist of terms and their definitions.

Note that there is a recursive aspect to these nestings. Explicit is that a frameset element may contain another frameset element (so that a screen can be subdivided *ad nauseum*). Most of the possible nestings are not illustrated. Notably, lists may contain other lists, though this is trickier than before; XHTML will demand that an inner list have a list *item* (li) as its direct parent. Likewise, a table may be the child of a *cell* (td) of an outer table. (As we shall see, nested tables have historically been very important for arranging content.)

All content that is to be rendered—text, images, and objects—must appear in the body element of a document. Most elements that appear within the body element can be classified as being either block elements or inline elements. Roughly speaking, *block elements* are larger sections, such as paragraphs (p). Following the analogy, an *inline element* would correspond to a portion of a paragraph. "Block" and "inline" are different levels in the structure of a document. In particular, no inline element may have a block element as a child, but every inline element must be embedded within a block element. A few elements (del, ins, script, and noscript) can be used at either level. As with paragraphs, block elements always begin on a new line and are often separated by vertical space. This is not the case with inline elements.

The div element has a unique role among block elements. Its opening and closing tags delimit a section of a document (without illegally crossing the boundaries of any other block element) so that it can be identified for formatting using CSS. An analogous inline element for delimiting a portion of a block element is span. These two elements are called *grouping elements*.

Associated with each element is the set of attributes it may (and, in some cases, must) have. Some attributes are unique to a particular element. Other, so-called generic attributes are shared by many elements. These are categorized as follows:

1. *Core attributes* (class, id, style, and title) identify an element.
2. *Internationalization (i18n) attributes* (dir, lang, and xml:lang) pertain to the natural language used in a document and the direction in which text is rendered.
3. *Event attributes* (onclick, ondblclick, onkeydown, onkeypress, onkeyup, onmousedown, onmousemove, onmouseout, onmouseover, and onmouseup) allow dynamic interaction by the user.
4. *Focus attributes* (accesskey, tabindex, onblur, and onfocus) apply to elements that could be the focus of the user's attention as indicated by the mouse or keyboard.

We make no attempt in the tables that follow to list event or focus attributes or any of the other attributes related to user interaction with an element, which involves Dynamic HTML, the subject of a separate chapter.

To better understand the hierarchical structure of a well-formed XHTML document, we illustrate a sample parse tree in Table 3.

Related to the issue of document structure is the *Document Object Model* (*DOM*) concept. This is a way of

Table 2 Sample Nestings of Block Elements in XHTML Documents

```
<html>[a]
   <head>[a]
      <title></title>[a]
      <base></base>
      <link></link>
      <meta></meta>
      <script></script>
      <style></style>
   </head>
   <body>
      <address></address>
      <blockquote></blockquote>
      <div></div>
      <hn></hn> (n = 1 to 6)
      <hr />
      <p></p>
      <pre></pre>
      <form>
         <button></button>
         <fieldset>
            <legend></legend>
         </fieldset>
         <label></label>
         <input/>
         <select>
            <optgroup>[c]
               <option></option>
            </optgroup>
         </select>
         <textarea></textarea>
      </form>
      <dl>
         <dt></dt>
         <dd></dd>
      </dl>
      <noscript></noscript>
      <ol>
         <li></li>
      </ol>
      <table>
         <caption></caption>
         <colgroup>
            <col></col>[b]
         </colgroup>
         <thead></thead>[cd]
         <tfoot></tfoot>[cd]
         <tbody></tbody>[c]
            <tr>[b]
               <th></th>
               <td></td>
            </tr>
      </table>
      <ul>
         <li></li>
      </ul>
   </body>
</html>
```

(a) Nesting in XHTML 1.0 Strict

```
<html>[a]
   <head>[a]
      <title></title>[a]
      <base></base>
      <link></link>
      <meta></meta>
      <script></script>
      <style></style>
   </head>
   <frameset>[a]
      <frame></frame>
      <frameset></frameset>
      <noframes></noframes>
         <body></body>
   </frameset>
</html>
```

(b) Nesting in XHTML 1.0 Frameset

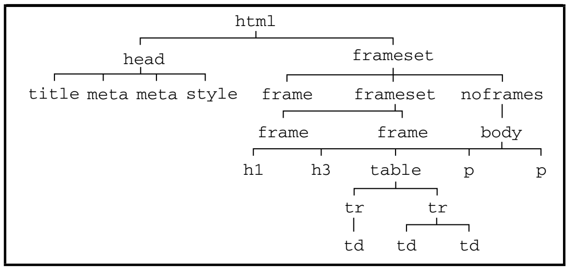
[a] Required element.

[b] Required child element if parent is present.

[c] Optional parent element.

[d] Same (grand)child elements as for `tbody`; cannot appear after or without a `tbody` sibling.

Table 3 Sample Structure of an XHTML 1.0 Frameset Document



referencing objects and subobjects related to a document. These could be elements, attributes of elements, or content within elements. An object might be information about a document (and not necessarily listed in the `head` element, as described next) or about the user agent (e.g., which version of which browser is being used). Having a way to reference items allows a client-side script to change those that are not read-only. For example, a script can be used to change simultaneously the URI of the documents loaded into two different frames.

As of this writing, W3C is finalizing DOM Level 3. The ultimate reality, however, is the DOM implemented by each individual browser. These have varied widely, even between versions of the "same" browser. (Netscape 4 and 6 offer a striking contrast.)

Obviously, having a well-defined, hierarchical structure within a document as demanded by XHTML can only aid any DOM in referencing objects. See the separate chapter on DHTML, where the DOM is of critical importance.

Document Information

The required `head` element of an XHTML document contains information about the document. The information is not rendered on the screen. However, the contents of the only required element within the document heading, the `title` element, will typically be displayed in a browser's title bar. The elements that can reside in the `head` are described and their attributes listed in Table 4.

The `base` element can be used to specify the URI on which all relative URIs within the document are based. A URI within a document can be stated either as a complete URI (starting with the protocol, e.g., `http`) or as a relative URI (with addresses being relative to the address of the current document, e.g.,`/images/shim.gif`). The use of `base` facilitates the moving of files to another server.

While a document can have at most one `title` or `base` element, it can have any number of the other types of elements listed in Table 4.

The `script` element allows one to provide code (perhaps in JavaScript) for a dynamic page, which reacts to the user's actions. A `noscript` element (within the `body` element) can hold alternative content for a browser that is unable or currently unwilling (according to its user's wishes) to execute scripts.

The `style` element can be used to include style information. This is becoming increasingly important as XHTML pushes us to better separate content and presentation.

Either script or style information can be included from an external file by means of a `link` element. This is the elegant way to use the same JavaScript code or CSS specifications in multiple documents, but it is now recommended practice even for use in a single document. See the separate chapters on DHTML and CSS.

Special mention should be made of `meta` elements because of the vast array of jobs they can perform. The `content` attribute is always required. Arguably the most useful attribute is `http-equiv`. A recommended usage is to specify for an HTML user agent the same character set as was specified for an XML user agent. (Or, in light of previous comments, the XML declaration might be omitted, and this would partially play its role.) A sample is

```
<meta http-equiv="Content-Type"
  content="text/html;
  charset=UTF-8" />
```

It is well known that `meta` elements are used by Web page authors to attract the attention of search engines. This is accomplished by code such as

Table 4 Document Information Elements for XHTML 1.0 Strict

Element	Description	Attribute(s)
`base`	Contains the document's base URL.	`href`
`link`	Specifies URI and type of an external script or style file.	`charset`, `class`, `dir`, `id`, `href`, `hreflang`, `lang`, `media`, `rel`, `rev`, `style`, `title`, `type`, `xml:lang`
`meta`	Contains information about content.	`content`[a], `dir`, `http-equiv`, `lang`, `name`, `scheme`, `xml:lang`
`script`	Contains code for dynamic content.	`charset`, `defer`, `src`, `type`[a], `xml:space`
`style`	Contains code to define formatting.	`dir`, `lang`, `media`, `title`, `type`[a], `xml:lang`, `xml:space`
`title`[b]	Contains text for browser title bar.	`dir`, `lang`, `xml:lang`

[a] Required attribute.
[b] Required child element within `head`.

Table 5 Nonempty Block Elements for XHTML 1.0 Strict

Element	Description	Attributes
address	Holds contact information for the document's author.	dir, id, class, lang, style,title, xml:lang
blockquote	Usually indents on left and right.	cite, dir, id, class, lang, style, title, xml:lang
div	Division: delimits content for specific formatting.	class, dir, id, lang, style, title, xml:lang
dl[b]	Holds a list of definitions.	class, dir, id, lang, style, title, xml:lang
form[b]	Form to be filled in.	accept, accept-charset, action[a], class, dir, enctype, id, lang, method, style, title, xml:lang
h1 to h6	Heading (largest to smallest): sets font-size.	class, dir, id, lang, style, title, xml:lang
ol[b]	Holds an ordered (numbered or lettered) list of items.	class, dir, id, lang, style, title, xml:lang
noscript	Contains alternative content when scripting is not supported.	class, dir, id, lang, style, title, xml:lang
p	Provides vertical whitespace around a paragraph.	class, dir, id, lang, style, title, xml:lang
pre	Text is "preformatted" (all whitespace preserved) and monospaced.	class, dir, id, lang, style, title, xml:lang, xml:space
table[b]	Arranges content in rows and columns.	border, cellpadding, cellspacing, class, dir, frame, id, lang, rules, style, summary, title, width, xml:lang
ul[b]	Holds an unordered (bulleted) list of items.	class, dir, id, lang, style, title, xml:lang

[a] Required attribute.
[b] Parent element (see Table 2 for possible child elements).

```
<meta name="keywords" content="cheap books"
  />
<meta name="description" content="Our
  discontinued items" />
```

The values of the `content` attribute have sometimes been used deceptively or unfairly. This has had legal consequences, such as successful lawsuits charging infringement on trademark rights.

There are also ways to attempt (!) the opposite—intentionally steer people away from a page. One way is to provide the *Platform for Internet Content Selection* (*PICS,* a W3C trademark) rating for a page; this system was originally designed as a filtering aid for parents and teachers, but is now in wider use (Boumphrey et al., 2000). Another technique is to give metainformation to indicate to *spiders* (a.k.a. *Web crawlers* or *bots*) that a page should not be indexed by a search engine and/or that links from that page are not to be followed for that purpose either.

A completely different use of a `meta` element is to control the loading of a page. An `"expires"` value for `http-equiv` can force a browser to load a new (rather than cached) copy of a page. A value of `"refresh"` causes a page to be reloaded after a specified time delay (10 s in the following example). Alternatively, the browser can be redirected to a different URI, perhaps to the new host server of the page. Often done primarily for visual impact, the splash page technique (illustrated here) was once popular.

```
<meta http-equiv="refresh" content="10;
  URL='page2.html'" />
```

Block Elements

Unlike in HTML, all content must reside within a block element. An easy mistake to make is to create vertical whitespace using `br` elements that are directly contained in the `body` element without being nested within a block element, such as the `p` element.

The allowable nonempty block elements are described and their attributes listed in Table 5.

Inline Elements

Inline elements fall into two categories based on the role they play.

Style elements or *physical style elements* exist solely to control the physical appearance, i.e., presentation of content; therefore, they are also referred to as *presentation elements*. (In the XHTML DTDs, they are called *font style elements*.) The elements in Table 6 are "reliable" in that they do what one would expect. Curiously, despite W3C's campaign to separate content and presentation, the only style elements eliminated from Strict are `s(=strike)` and `u`, which specified strikethrough and underlined text, respectively.

By contrast, most of the *descriptive elements* (called *phrase elements* in the XHTML DTDs) seem superfluous at first glance. As Table 7 illustrates, only `sub` and `sup`

Table 6 Inline Style (or Physical Style or Presentation) Elements for XHTML 1.0 Strict

Element	Description	Rendering
`b`	Boldfaced	**Boldfaced** text.
`big`	Bigger	Bigger relative to current font size.
`i`	Italicized	*Italicized* text.
`small`	Smaller	Smaller relative to current font size.
`tt`	Teletype	`Monospaced` text

do something that could not be accomplished by style elements. Some elements' contents are not even rendered with any distinctive visible characteristic. And the appearance of the content may vary from browser to browser. Yet these are precisely the elements good practice dictates that we use. The reason is that they convey the roles they play. They are sometimes called *logical* or *content-based elements* because they reveal the semantic structure of the document. An XML-aware application could search, reference, or modify a document based on the meaning associated with content that is wrapped in, for example, a `code` element.

Special Elements

Some of the most famous elements do not fit into any of our previous tables, as they play special roles. They are described and their attributes are listed in Table 8.

The anchor element (`a`) is the most important element of all, for without it we would have no hypertext! It may have content, which the viewer clicks on to link to another document or another location within the document. Often this content is text. The traditional, default rendering of such content in many browsers is blue, underlined text. If the contents of an anchor element is an `img` element (see later), the analogous default rendering adds a blue border to the image on which the user would click. Increasingly, CSS is being used to override these default renderings. If the anchor tag is empty and has a `name` attribute, it can be used as the destination of a link so the user can be brought to a specific location within a document. Clearly, such a tag would not need an `href` attribute, which specifies the URI of a destination.

Table 7 Descriptive (or Logical or Content-Based) Inline Elements for XHTML 1.0 Strict

Element	Description	Rendering
`abbr`	Abbreviation	No distinction in most browsers.
`acronym`	Acronym	No distinction in most browsers.
`cite`	Cited text	*Italicized* in most browsers.
`code`	Code	`Monospaced` in most browsers.
`dfn`	Definition	*Italicized* in most browsers.
`em`	Emphasis	*Italicized* in most browsers.
`kbd`	Keyboard	`Monospaced` text.
`q`	Quotation	No distinction in most browsers.
`samp`	Sample	`Monospaced` in most browsers.
`strong`	Strong	**Boldfaced** in most browsers.
`sub`	Subscript	$_{\text{Lowered}}$ relative to baseline.
`sup`	Superscript	$^{\text{Raised}}$ relative to baseline.
`var`	Variable	*Italicized* in most browsers.

The `img` element inserts a JPEG, GIF, or PNG image in a document. (The first two formats are well established and supported. The last format, introduced by W3C in 1996, is similar to GIF, but not reliant on a licensed compression algorithm.) There are two required attributes. The value of `src` is the URI of the image file. XHTML also enforces courtesy, i.e., specifying with the `alt` attribute what text should be displayed (or spoken by a voice synthesizer) if the image is not. It is always good practice to specify the `height` and `width` of an image, even if the intention is to load the image at its natural size. For if the browser knows the footprint in advance, it will not have to do more computations and rearrange content once the image is completely downloaded. The value of a `usemap` attribute can be used to point within the document to a `map` element whose child `area` elements specify hotspots within the image that can act as links; this situation is called a client-side image map. The `map` and `area` elements are not detailed here. See Boumphey et al. (2000) for more information on image maps.

Two special elements cause breaks in the presentation. The **`<br/>`** tag forces a line break, while the **`<hr/>`** tag creates a horizontal rule. These two elements differ greatly from the standpoint of XHTML, as `hr` is a block element, while `br` is an inline element.

The `object` element is an inline element that can be used to embed a wide variety of documents in an XHTML file. One such document is a Java Applet, which previously could be embedded using the now deprecated `applet` element. An `object` element's `param` child element (not listed in the tables) can be used to specify values for parameters used by the embedded document.

Two other elements, `ins` and `del`, are used to signify the insertion or deletion of content since a previous version of the document.

Finally, the highly specialized `bdo` element allows an advanced programmer to take control of character sequencing. The required `dir` attribute takes either `"ltr"` or `"rtl"` as a value to override the default (bidirectional) rendering algorithm.

STRATEGIES FOR WRITING XHTML CODE

Knowledge of the rules of a game does not automatically imply knowledge of how to play the game effectively. Likewise, knowledge of the syntax rules of a language does not imply understanding of how to communicate (write code) effectively. The issues that influence what constitutes good practice in writing XHTML code are often beyond the control of the programmer, for instance, the capabilities of a user's browser or the options the user has chosen for that browser's behavior.

Another issue, already discussed, has to do with Web authoring philosophy. From a superficial standpoint, inline style elements seem to be more manageable since

Table 8 Special Elements for XHTML 1.0 Strict

Element	Description	Attributes
a	Anchor: source of one link or destination of another link.	charset, class, classid, codetype, coords, dir, id, href, hreflang, lang, name, rel, rev, shape, title, type, xml:lang
bdo	Specifies left-to-right or right-to-left (rather than default, bi-directional) rendering.	dir[a], lang, xml:lang
br	Forces a line break.	class, id, style, title
del	Identifies content deleted since an earleir version	cite, class, datetime, dir, id, lang, style, title, xml:lang
hr	Horizontal rule.	class, dir, id, lang, name, style, title, xml:lang
img	JPEG, GIF, or PNG image.	alt[a], class, dir, id, ismap, height, lang, longdesc, name, src[a], title, usemap, width, xml:lang
ins	Identifies content inserted since an earleir version.	cite, class, datetime, dir, id, lang, style, title, xml:lang
object[b]	Specifies URI of external content file to be embedded.	archive, class, classid, codebase, codetype, data, declare, dir, id, height, lang, name, standby, style, title, type, usemap, width, xml:lang

[a] Required attribute.
[b] May have a param child element.

they deliver what they promise and have mnemonic names. In the long term, the document itself will be more manageable if the elements convey the role of the content. In other words, the descriptive elements in Table 7 are preferable to those in Table 6.

Frames

One of the most famous and controversial techniques in the development of Web pages is the use of *frames* to partition what superficially appears as one page into separate pages, which may or may not share some kind of linkage. In XHTML, this is accomplished in Frameset using the frameset element to replace the body element of a Strict or Transitional document. (Frameset is not simply a superset of Strict; the former does not support the body element's being a direct child of the html element, though it can be a great-grandchild element with a noframes parent element.)

If two frame elements are nested within the frameset element, the screen display will be subdivided into two independent areas, each displaying its own XHTML document. Whether the screen is divided horizontally or vertically depends on whether the frameset element has a rows or cols attribute. The value (actually a set of values) of that attribute determines how much of the screen is used by each frame. The screen can be subdivided in both directions at once or can be successively subdivided with nested frameset elements, though the disadvantages of screen fragmentation are obvious.

Thus to display frames, there is, in addition to the documents rendered within the individual frames, a controlling XHTML document that contains the frameset element that determines how the screen will be subdivided. Even though it may not provide any of the content rendered, it is the primary document in several senses. When a user requests a browser to display the source code being rendered, it is the sparse code for setting up the frames that will be displayed, unless the user somehow selects a particular frame (say, by clicking a mouse over it). The browser will display in its title bar the text specified in the title element within the frameset document; this is logical from the standpoint of the browser, which can only display one title.

Frames offer the opportunity for a combination of constancy and changeability. A classic application is to have a *navigation bar* on the left or top of the screen to maintain a set of links to other pages. It stays unchanged while the larger portion of the screen loads a new page. The code for the "navbar" will specify the initial URIs of the documents for both frames, but the current URIs will be determined (for at least one frame) by the viewer's interaction.

The original objection to frames was that not all browsers supported them. While fewer and fewer people with full-size screens are stuck using such browsers, many sites still provide alternative, nonframes versions of their pages. A rising issue is that of very small screens, say in hand-held devices. This will be dealt with in greater detail later, as we discuss the true potential of the XML approach to marking up hypertext.

For users wishing to cite Internet resources as formal references, perhaps in the bibliography of a report, frames present two problems. The title of the frameset document and therefore the title displayed does not change even when the page loaded into a frame does. The same is true of the address bar. Consequently, a naïve user who does not know how to "isolate" a frame to determine its precise source may provide a URI that is not specific enough to home in on the desired page; a reader of the inaccurate citation will have to navigate (and possibly search) further.

For users who want a hardcopy of a Web page, printing from a screen that actually displays two or more frames can be problematic.

Strict does not support frames. Even if one uses Frameset only for the controlling document and Strict for all documents loaded into individual frames (as some authors suggest), the purpose of frames has been defeated. This is because to load a document into a specific frame, the anchor (a) element commanding the load must specify the value of a target attribute; the value might be the value of a name/id attribute (id is preferred by XHTML, but name is needed for compatibility with older browsers) previously assigned to a frame, or it may be a relative value, such as "_top" (for the entire screen) or "_self" (for the frame making the call). But Strict does not support the target attribute! The alternative is to use JavaScript or server-side code to change the contents of a frame.

Clearly, W3C has put the handwriting on the wall for frames. The sole XHTML 1.1 DTD is based on Strict and, therefore, does not support frames.

Already, even among those sites using HTML rather than XHTML, the trend is to develop navigation schemes that superficially resemble frames-based pages, but actually use DHTML or server-side code to change the "main event" on a page while keeping the navigational area constant (or, better yet, modifying the navigational area slightly to indicate "where you are" or to expand or pop up submenus of links).

Tables

Another controversial component of Web pages is tables. Their purpose is to arrange content into rows and columns. Structurally, tr elements define rows, and they may have as children td (normal) and/or th (heading) cells. The latter have default formatting that usually entails boldfaced and horizontally centered text. A row may have rowspan and/or colspan attributes to merge cells into larger, rectangular cells. The formatting of entire columns can be controlled by means of col elements within a colgroup element (recall Table 2). The table element may have a caption element, and rows may be grouped within optional thead, tfoot, and/or tbody elements; if either or both thead and tfoot elements are used, they must precede a tbody element, although a tfoot element will be rendered below the tbody element.

By default, tables will display a "border," which actually includes vertical and horizontal walls around nonempty cells. If a consistent look is desired, an empty cell must be given some invisible content, such as a nonbreaking space (&nbsp;); an alternative, which is also used to ensure a minimum size for a cell, is an invisible GIF, in which every pixel is transparent.

The vast majority of tables in use in Web pages have a value of "0" for the border attribute. Thus, the viewer sees no walls between rows and columns. The page designer is using tables not to create visible boxes for content (the original intent), but to control the placement of content. This is a natural way to overcome browsers' reluctance to place items side by side. While browsers typically separate consecutive block elements with vertical whitespace, tables can be nested without extra space being created between a table and the walls of its parent td element. Web designers have reportedly nested tables as much as seven deep to obtain desired arrangements.

A sample application of tables is to slice an image into quadrants (for instance) and nest each in an anchor element in a td element. If the pieces fit together seamlessly, the user perceives one image as having four hotspots linking to different pages. This is an alternative to image maps.

One problem with tables is that browsers must compute how to fit things in a table, and this takes time—more time than if CSS had been used to position items. Perhaps a bigger problem is that each browser has its own algorithm for deciding *how* to fit content into cells. The way in which the total width of a table is distributed among columns can seem arbitrary and sometimes esthetically distressing. In reaction, Web designers have often "hardcoded" the width of tables and columns, which in turn has forced some viewers to scroll horizontally if the table does not fit their current window.

While the use of CSS is recommended over the use of tables for (transparently) arranging content, the fact remains that CSS support is irregular, especially among older browsers. The reality of the current browser climate pressures many developers to cling to tables.

Living without Deprecated Features of HTML

One of the significant deprecations in XHTML 1.0 is the center element. The simultaneous deprecation of the align attribute for most elements means that something as simple as placing a table in the middle of a page is now more complicated than it used to be. One solution is to use CSS. Another is to nest the target table within the sole cell of an outer table element with "100%" as the value of the width attribute. Since the td element still is allowed to have "center" as the value of the align attribute, the target table can be centered within the cell and, therefore, the page.

Another deprecated element is font, used to alter the size, font face, and/or font color of text contained in the element. The only substitute for this is to use CSS to define a style. The span or div element could be used to delimit the section of text to be altered. The latter element would be used to encompass one or more block elements, while the former would stay within a single block element. In either case, the element must have a class attribute whose value would be referenced by the style sheet. The situation is simpler if the target text is precisely the contents of an existing element, for then that element can carry the class attribute without need for a span or div element. "Inlining" the style with a style attribute for an element will not have the desired effect in Netscape 4 and may cause truly strange side effects. In any case, the style attribute of elements is removed from XHTML 1.1.

In a similar vein, Strict removes all formatting attributes of the hr element, forcing one to use CSS to exert even a modicum of control over the appearance of horizontal rules.

Conversion and Validation Tools

A number of tools exist for transforming legacy HTML code to XHTML. The most famous is Dave Raggett's HTML Tidy, originally hosted at W3C. The continued development of this application is now maintained by Source Forge (Source Forge, 2002). The platforms supported include many flavors of UNIX, Windows, and Mac

OS operating systems. Each version has a command-line interface. See St. Laurent & DeLong (2000) for a sample before-and-after look at Tidy's handiwork.

For the Windows platform, HTML Tidy can be augmented with a graphical user interface, Tidy UI, by Charles Reitzel (2003). Another option is to run HTML Tidy from within the much more comprehensive HTML-Kit (Chami.com, 2002), which also allows multiple files to be converted at once.

Despite the many options offered by these conversion tools, there is no way to automate the conversion process completely. Aside from the usual bugs, these programs have inherent limitations when it comes to imposing proper structure. An extreme example is when the HTML code is lacking a closing delimiter, such as a right angle bracket (>) or closing quotation mark. Even a less drastic flaw may cause a tool to issue a warning or error message, leaving the program's user to decide how to fix the problem by hand. There are times, however, when a tool does make a decision, and the result is inconsistent with the author's intentions or with a browser's rendering of the original HTML code. One situation would be incorrectly nested elements. Conversion tools must be viewed as potentially imperfect aides.

Whether authoring a page from scratch, converting legacy code, or generating code with an authoring system, purported XHTML code should be validated. The W3C MarkUp Validation Service (see W3C, 2002c) has recently been improved so that local files can be checked. When errors are found in a file, the source code can be listed with line numbers to quickly locate the problem. Predictably, however, a single error often results in multiple error messages. A parse tree can also be generated for a valid document. Furthermore, one can embed in a page a referrer "button" (actually `<p><a...><img.../></a></p>`) to allow a viewer to test the validity of the page's XHTML code.

The value of validating documents is in preparing for a future when (it is hoped) browsers no longer render a structurally defective document. A strictly conforming XHTML user agent must parse a document and check it for well-formedness (proper nesting of elements, etc.). Increasingly intelligent indexing and manipulation of Web resources (the "Semantic Web") will rely on the well-defined structure of documents. A validating user agent must additionally validate a document against its referenced DTD (to see that it is using only the allowed names, etc.). The specifics of user agent conformance can be found in section 3.2 of W3C (2002a).

Updating HTML-Generating Applications

As difficult as it may be to convert existing HTML code to XHTML code, there is a more vexing problem: converting the code that generates HTML. It could be PHP, Perl, ASP (Active Server Pages), or JSP (Java Server Pages) programs on a server constructing Web pages on the fly. Or it could be a Web authoring tool such as Microsoft FrontPage, Adobe GoLive, Macromedia Dreamweaver, Mozilla, or even a word processor. Or it could be JavaScript embedded in a page's code to dynamically create additional HTML code.

An excellent discussion of the issues involved in this aspect of the switchover to XHTML can be found in Chapter 11 of St. Laurent & DeLong (2000, p. 201), which claims that "the cleanup process is nearly as complicated as the work that was done on Y2K."

A glimpse of the difficulties can be seen by analyzing how developers have hand-coded HTML. Those behavior patterns translate into analogous authoring algorithms.

One flaw is to think of HTML tags as nonprinting characters in a word processor. An example is putting `<p>` tags at the end of lines as a kind of line break, rather than as an opening tag (which must later be matched by a closing tag). Another flaw is, thinking of `<li>` as a "bullet symbol" at the beginning of a list item, with no closing tag at the end of the item.

The importance of updating authoring tools and page generating code cannot be overstated. A substantial fraction of Web pages are not written "by hand," but are the product of Web authoring tools or are generated on demand, perhaps in response to a user's query.

Two XHTML authoring tools currently available are Macromedia's Dreamweaver MX and W3C's own Amaya, a combined browser and WYSIWYG editor (W3C, 2003c).

FUTURE DIRECTIONS

To date, many organizations have failed to see the need to convert old HTML code or even begin using XHTML code for new pages. They feel the move is not justified by the small return on investment they anticipate. In addition to updating commercial or in-house products for generating code, some companies face another problem. In environments where employees with minimal training have been contributing HTML code, those contributors must be retrained to meet the more demanding standards of XHTML.

Even while much of the Web development community drags its feet in coming into conformance with XHTML 1.0, W3C marches on. Just as some movie scripts are written with a sequel in mind, XHTML 1.0 was merely the beginning of a very deliberate trek toward XML. XHTML 1.1 represents another quantum leap in HTML/XHTML evolution. As of this writing, the Working Draft of the XHTML 2 specification breaks from tradition by foregoing backward compatibility (W3C, 2003a).

Extensions of XHTML

The potential for extensions of XHTML is unlimited. This is done with what W3C terms "modules." These are not technically part of XHTML, but additional XML applications with which XHTML can work. While the expectation is that new, specialized modules designed by individual vendors will proliferate, W3C began undertaking the development of its own modules even before XHTML was released. Two of these are the Mathematics Markup Language (MathML) and Scalable Vector Graphics (SVG), both W3C trademarks. These have been slow to gain acceptance, but some foresee SVG as becoming a serious open-source challenger to Macromedia's well-entrenched Flash.

The MathML module (W3C, 2001a), first released in 1997, provides a way of formatting mathematical expressions so that they will be rendered in the expected fashion. Elements, for example `mfrac` for fractions, specify the semantics of subobjects within an expression. Expressions can be dynamically manipulated by client-side or server-side script. Numerous applications—WebCT, to name just one—are MathML-aware. Yet support among browsers is not as simple or rosy a story as one might first believe. For example, in order for Internet Explorer 5.5+ to find where the MathPlayer plugin is installed locally, an `object` element must be added to the code (see Design Science, 2003). MathML content resides within this element:

```
<math xmlns="http://www.w3.org/1998/Math/
  MathML"> </math>
```

The SVG module carries even greater potential. Traditional graphical images for the Web have been in three formats: JPEG, GIF, and PNG. Each of these is a (compressed) bitmapped image, i.e., an array of pixels. Such images tend to be large and slow to download. Moreover, displaying an image at different sizes (to better accommodate a range of screen resolutions) results in either a blocky, "pixelated" appearance if an image is enlarged beyond its natural dimensions or a waste of time if an image is downloaded only to be displayed in shrunken form. Vector graphics encapsulate directions for how to draw an image. For images without photo-like detail, these directions result in smaller files and allow images to be drawn at any scale without loss of detail. Perhaps more important, an image can have parameters (colors, for instance) manipulated interactively. Obviously, the need for bitmapped images will remain.

For other extensions of XHTML, see the separate chapter on XML or Chapters 20-21 of Deitel et al. (2001).

Modularization and XHTML 1.1

The XHTML family began with XHTML 1.0 as a collection of DTDs. This was itself a transitional stage aimed at providing for (relatively) easy migration of legacy HTML code to well-formed XHTML code. The most important feature of the single XHTML 1.1 DTD is its XHTML modular framework. The purpose is to allow developers to use only the modules they need and to add suitable custom modules.

Starting with Strict, W3 made the following changes to derive XHTML 1.1:

1. All deprecated features were removed, not just those contained in Frameset and Transitional. The `name` attribute of the `a` and `map` elements, retained solely for compatibility with older browsers, was removed in favor of `id`. Similarly, the `lang` attribute was dropped for all elements in favor of `xml:lang`.
2. Ruby annotation was added. A *ruby* is a collection combining content and an annotation to be displayed above (and below, if there is a second annotation) the content, e.g., to provide the pronunciation of a foreign term. These are most often used with eastern Asian languages. For details on the use of ruby annotation, consult W3C (2001d).
3. The DTD was partitioned into modules as shown in Table 9.

Table 9 XHTML 1.1: Modularization

Module	Elements or Attributes
Structure	`body`, `head`, `html`, `title`
Text	`abbr`, `acronym`, `address`, `blockquote`, `br`, `cite`, `code`, `dfn`, `div`, `em`, `h`*n*(*n* = 1 to 6), `p`, `pre`, `kbd`, `q`, `samp`, `span`, `strong`, `var`
Hypertext	`a`
List	`dl`, `dt`, `dd`, `li`, `ol`, `ul`
Object	`object`, `param`
Presentation	`b`, `big`, `hr`, `i`, `small`, `sub`, `sup`, `tt`
Edit	`del`, `ins`
Bidirectional Text	`bdo`
Forms	`button`, `fieldset`, `form`, `input`, `label`, `legend`, `select`, `optgroup`, `option`, `textarea`
Table	`caption`, `col`, `colgroup`, `table`, `td`, `tfoot`, `th`, `thead`, `tr`
Image	`img`
Client-Side Image Map	`area`, `map`
Server-Side Image Map	attribute `ismap of img`
Intrinsic Events	event attributes
Metainformation	`meta`
Scripting	`noscript`, `script`
Stylesheet	`style` element[a]
Link	`link`
Base	`base`
Ruby Annotation	`ruby`, `rbc`, `rtc`, `rb`, `rt`, `rp`

[a] The `style` *attribute* is deprecated.

Note that the partitioning is not along the lines by which we categorized elements earlier. A module may contain both block and inline elements, for example.

The appropriate `DOCTYPE` declaration for an XHTML 1.1 document becomes

```
<!DOCTYPE html PUBLIC "//W3C//DTD XHTML
  1.1 //EN""http://www.w3.org/TR/xhtml11/
  DTD/xhtml11.dtd">
```

The opening `html` tag is the same as that for XHTML 1.0, except that the `lang` attribute is deleted. The default XML namespace remains the same.

XHTML Basic

The purpose of XHTML Basic is to provide a very small DTD for very small platforms, including mobile phones, PDAs (personal digital assistants), pagers, vending machines, set-top interfaces for televisions, digital book readers, smart watches, mobile game machines, and automobile navigation systems. However, it presents a minimal core which can provide bare-bones functionality on all platforms and which can be augmented by additional modules. The outline of XHTML Basic can be seen in Table 10.

XHTML Basic was released even before XHTML 1.1. It is based on an early version of the "Modularization of XHTML" recommendation, the latest version of which is W3C (2001b). W3C recommendations go through a progression of stages: Working Draft, Last Call Working Draft, Proposed Recommendation, Candidate Recommendation, and W3C Recommendation. Since the modules of XHTML Basic are based upon another, mature recommendation, W3C's HTML Working Group considers XHTML Basic to be stable as well (W3C, 2000b).

Comparing Table 10 with the earlier Table 9 shows which XHTML 1.1 modules are missing from the Basic DTD: Presentation, Edit, Bidirectional Text, Client-Side Image Map,Server-Side Image Map, Scripting, Stylesheet, and Ruby Annotation.

Table 10 XHTML Basic: The Least Common Denominator

Module	Elements
Structure[a]	`body`, `head`, `html`, `title`
Text[a]	`abbr`, `acronym`, `address`, `blockquote`, `br`, `cite`, `code`, `dfn`, `div`, `em`, `h`*n* (*n*=1 to 6), `kbd`, `p`, `pre`, `q`, `samp`, `span`, `strong`, `var`
Hypertext[a]	`a`
List[a]	`dl`, `dt`, `dd`, `li`, `ol`, `ul`
Object	`object`, `param`
Basic Forms[b]	`form`, `input`, `label`, `select`, `option`, `textarea`
Basic Tables[b]	`caption`, `table`, `td`, `th`, `tr`
Image	`img`
Metainformation	`meta`
Link	`link`
Base	`base`

[a] Required HXTML Host Language module.
[b] Abridged version of an XHTML 1.1 module.

In addition, the Forms and Tables modules have been trimmed considerably. An additional restriction is that tables cannot be nested. This common technique makes little sense in this context given the lack of screen space on the target platforms and the computational expense associated with rendering tables.

Particularly notable is the elimination of both the inline style elements and the `style` element. This means that formatting of displays is to be accomplished by using the `link` element to include an external style sheet. Such formatting is still possible, as the `span` and `div` elements for delimiting content and the `class` attribute are still supported.

Potentially even more significant is the lack of support for scripting, which is matched by the absence of event attributes. The intention is not to prohibit interactivity. Rather, events, e.g., an incoming phone call, are seen as being highly device-dependent. Consequently, platform-specific modules must be designed and implemented to achieve dynamic content.

CONCLUSION

HTML served the World Wide Web well during much of the Internet Age. In the previous millenium, it became forever frozen at Version 4.01.

The early history of HTML could be characterized by two words: recommendations ignored. Competition between two particular browsers, Microsoft Internet Explorer and Netscape Navigator/Communicator, produced extensions to W3C's recommendations, which influenced the expansion of the officially sanctioned HTML lexicon. At the same time, browsers chose not to support certain features contained in recommendations or supported them poorly or inconsistently. In some cases, the blame can be placed on vagueness in W3C's specifications. In extreme cases, browsers exhibited bizarre behavior which can only be described as major bugs.

Perhaps worst of all, browsers tolerated (and continue to tolerate) malformed HTML documents, encouraging the Web community to develop poor code-writing and code-generating habits. In particular, an HTML document is commonly viewed as text to be processed, not a tree to be parsed. Much existing code does not exhibit a well-defined structure that could be processed based on semantic content.

HTML itself is a hybrid product. In part, it manages content, in the spirit of SGML. But it has, over time, accrued formatting capabilities. This has contributed to the development of another bad habit: using HTML to manage the presentation. Admittedly, CSS arrived late on the scene, and browsers' irregular support for CSS has forced many developers to view HTML itself as a more reliable formatting tool.

HTML's fatal flaw, however, is its inability to adapt to future or, for that matter, present technology. On the one hand, it does not allow high-powered environments to realize their full potential to render complex or specialized content. On the other hand, it cannot accommodate the special requirements of small devices with limited (if any) screen space and modest computing power.

W3C's ultimate vision for the Web is to have a framework that can grow (possibly from a very tiny core) to fit the specific needs of any current or future platform. The basis for such a flexible tool is XML. XHTML, one particular XML application, is a transitional tool for ultimate conversion to XML; as such, it will likely persist for years.

The major contribution of XHTML 1.0 is the enforcement of structure: to be a valid XHTML (and, therefore, XML) document, it must be well-formed. Of the three DTDs or XML Schemas, Strict foreshadows the demise of frames—always controversial, once popular, but now fading in importance—and the further separation of content and presentation in XHTML 1.1.

XHTML 1.1's only DTD is fragmented into modules, the key to extensibility. Designed for a variety of small platforms, XHTML Basic prunes the modules in both number and (for two modules) size. Code from many special-purpose XML applications, such as MathML and SVG, can be incorporated in XHTML documents.

At each step in the evolution of HTML/XHTML, moving forward has entailed leaving some things behind. Controversy is sure to follow this continued evolution. So far, the conversion to XHTML (even on pages at W3C's Web site) has been slow. While this conversion does not have the urgency or criticality of Y2K—power plants will not turn into pumpkins when the clock strikes 12—failing to convert to XHTML translates into not realizing the full potential of the Internet.

GLOSSARY

Attribute Parameter associated with an *element;* located, along with its value, in the element's opening *tag*.

Block element *Element* that can contain data (text, images, etc.), *inline elements*, and/or other block elements; typically *rendered* with intervening whitespace (a line break and often vertical space) between it and neighboring block elements; "larger" than an inline element (as a paragraph is "larger" than a sentence).

Cascading Style Sheets (CSS) A method for formatting and/or positioning content within an *HTML* or *XHTML* document; a *W3C* trademark.

Deprecated Slated for deletion from future versions; not recommended for usage.

Document Object Model (DOM) A system for viewing the content of an *HTML* or *XHTML* document as being organized into objects, which can be programmatically accessed in an object-based fashion; a *W3C* trademark.

Document Type Definition (DTD) (for a family of documents) Document that specifies, using an Extended Backus-Naur Form grammar, the features of a valid document; for a particular version of *HTML* or *XHTML,* these features would include the valid element types, the names of attributes that can (or must) be associated with elements, and the suitable ranges of values that can be assigned to attributes.

Element The unit of structure in an *HTML* or *XHTML* document.

Extensible HyperText Markup Language (XHTML) A highly adaptable *markup* language to allow Web content to be *rendered* by a *user agent;* a *W3C* trademark.

Extensible Markup Language (XML) A general methodology for structuring data; defined (but not trademarked) by *W3C*.

Hyperlink or link Within a *hypertext* document, an *element* associated with a pointer to another (part of the) document; activating the element—typically by clicking on its *rendering* as a button, image, or highlighted text—takes the user to the new document (part); the most essential component of a hypertext document.

Hypertext Technique for interactively traversing an electronic document (or collection of documents) by following *hyperlinks* as an alternative to sequential traversal.

HyperText Markup Language (HTML) The original language for marking up content on the World Wide Web; invented by Tim Berners-Lee; further developed (upon the suggestions of many) and codified (but not trademarked) by *W3C*.

Hypertext Transfer Protocol (HTTP) A generic, stateless, object-oriented, application-level (level 7 in the ISO OSI model) protocol (typically associated with Port 80) that allows for negotiation of data representation; usable for a variety of tasks, e.g., distributed object-management systems, but most famous as the means of sending HTML code through the Internet.

Inline element *Element* that can contain only data (text, images, etc.) and other inline elements; "smaller" than a *block element* in the sense that it must be embedded within a block element; rendering generally does not automatically add whitespace as for block elements.

Markup language Language for defining the structure and formatting of a document by means of ordinary characters (rather than, for example, escape characters) within the document.

Metalanguage Language used to describe another language (for example, *XML* is the metalanguage for *XHTML*.)

Parse To analyze a document to determine its structure.

Render For a *user agent,* to present a document on a screen, printed page, etc.

Standardized General Markup Language (SGML) A *metalanguage* standardized by the International Standards Organization (ISO 8859); used to define the syntax of textual markup languages; the precursor of HTML.

Tag The opening or closing marker of an *element;* in *SGML*-derived languages, delimited by angle brackets (< >). (Empty elements are represented by a single tag.)

Universal Resource Identifier (URI) A string identifying a resource on the Web by means of a name or address. There are a multitude of URI addressing schemes (see http://www.w3.org/Addressing). Two particular subtypes of URI are

1. **Universal Resource Locator (URL)** The original term for resource addresses associated with common URI schemes such as http and ftp; no longer used in technical specifications, but still appears within values in `meta` *elements*.

2. **Universal Resource Name** (**URN**) A persistent name that identifies a resource regardless of the particular address to which the name currently resolves; address resolution requires an institutional commitment to persistence.

User agent Application, such as a Web browser or an e-mail program, that can *render* an *HTML* or *XML* document.

Valid (relative to a particular *DTD* or *XML Schema*) Conforming to the DTD or XML Schema—all *elements* and their *attributes* are declared in the DTD or XML Schema, each element has appropriate attributes (perhaps as required), and each attribute's value is in the appropriate range.

XML namespace (for a document or specified portion of a document) A (structured) collection of names that can be used as *element* types and *attribute* names, identified by the URI of the file defining the collection.

XML Schema An alternative to *DTD*s for specifying a language (module); written in *XML* syntax, so it can be manipulated like any other XML document.

Well-formed Having a well-defined structure—*elements* are correctly nested and contain any required child element/elements.

World Wide Web Consortium (**W3C**) An international organization that makes recommendations regarding Web-related technologies; a registered trademark. (Note: While one commonly sees "the W3C" in print, W3C itself generally does not apply the definite article to its abbreviation; compare this with the broadcasting corporations "ABC" and "the BBC.")

CROSS REFERENCES

See *Cascading Style Sheets (CSS); Extensible Markup Language (XML).*

REFERENCES

Boumphrey, F., Greer, C., Raggett, D., Raggett, J., Schnitzenbaumer, S., & Wugofski, T. (2000). *Beginning XHTML.* Birmingham, AL: Wrox Press.

Chami.com (2002). *HTML-Kit*. Retrieved October 15, 2002, from http://www.chami.com/html-kit

Deitel, H. M., Deitel, P. J., Nieto, T. R., Lin, T. M., & Sadhu, P. (2001). *XML how to program*. Upper Saddle River, NJ: Prentice-Hall.

Design Science (2003). *Authoring for MathPlayer*. Retrieved April 23, 2003, from http://www.dessci.com/en/products/mathplayer/authoring.htm

Graham, I. S. (2000). *XHTML 1.0 language and design sourcebook* New York: Wiley.

International Organization for Standardization/International Electrotechnical Commission (2000). ISO/IEC15445:2000/COR1:2002(E). Retrieved January 3, 2003, from http://www.cs.tcd.ie/15445/15445.html

Reizel, C. (2003) *Charlie's Tidy Add-ons.* Retrieved April 19, 2003, from http://users.rcn.com/creitzel/tidy.html

Sauers, M., & Wyke, R. A. (2001). *XHTML essentials*. New York: Wiley.

Source Forge (2002). *HTML Tidy project page*. Retrieved October 15, 2002, from http://tidy.sourceforge.net

St. Laurent, S., & DeLong, B. K. (2000). *XHTML: Moving toward XML*. New York: Wiley.

World Wide Web Consortium (1999). *Namespaces in XML*. Retrieved October 15, 2002, from http://www.w3.org/TR/REC-xml-names

World Wide Web Consortium (2000a). *Extensible Markup Language (XML) 1.0* (2nd ed.). Retrieved April 19, 2003, from http://www.w3.org/TR/REC-xml

World Wide Web Consortium (2000b). *XHTML Basic*. Retrieved October 15, 2002, from http://www.w3.org/TR/xhtml-basic

World Wide Web Consortium (2001a). *Mathematical Markup Language (MathML) version 2.0*. Retrieved October 15, 2002, from http://www.w3.org/TR/MathML2

World Wide Web Consortium (2001b). *Modularization of XHTML*. Retrieved October 15, 2002, from http://www.w3.org/TR/xhtml-modularization

World Wide Web Consortium (2001c). *XHTML 1.1-module-based XHTML*. Retrieved October 15, 2002, from http://www.w3.org/TR/xhtml11

World Wide Web Consortium (2001d). *Ruby annotation. Retrieved October 15*, 2002, from http://www.w3.org/TR/ruby

World Wide Web Consortium (2002a). *XHTML 1.0: The Extensible HyperText Markup Language* (2nd ed.). Retrieved October 15, 2002, from http://www.w3.org/TR/xhtml1

World Wide Web Consortium (2002b). *XHTML 1.0 in XMLSchema*. Retrieved October 15, 2002, from http://www.w3.org/TR/2002/NOTE-xhtml1-schema-20020902

World Wide Web Consortium (2002c). *W3C MarkUp Validation Service*. Retrieved April 19, 2003, from http://validator.w3.org

World Wide Web Consortium (2003a). *XHTML 2.0* (W3C Working Draft, 31 January 2003). Retrieved April 19, 2003, from http://www.w3.org/TR/xhtml2/

World Wide Web Consortium (2003b). *Hypertext Markup Language (HTML) home page*. Retrieved April 19, 2003, from http://www.w3.org/MarkUp

World Wide Web Consortium (2003c). *Amaya home page*. Retrieved April 19, 2003, from http://www.w3.org/Amaya

Human Factors and Ergonomics

Robert W. Proctor, *Purdue University*
Kim-Phuong L. Vu, *Purdue University*

INTRODUCTION

Human factors and ergonomics (HFE) is an interdisciplinary field devoted to designing and evaluating products and systems for effective and safe use by humans (Proctor & Van Zandt, 1994). The fundamental idea underlying HFE is that systems must be designed with the users in mind if the systems are to accomplish their goals effectively. Transactions initiated through the Internet typically involve extensive human interactions with the system because the purpose of the Internet is to convey information from user to user, business to business, and business to consumer. Because the focus of HFE is on human–machine interactions in general, with a specific emphasis on human–computer interactions, the field has much to offer to the design of user interfaces and applications for the Internet (Proctor et al., 2002).

HFE emerged from issues involving use of military systems in World War II (Green, Self, & Ellifritt, 1995). In subsequent decades, however, the field expanded into many areas, including computer and information systems. Traditionally, the main contributors to HFE have been industrial engineers and psychologists, but more recently, individuals with a range of backgrounds and interests from both academia and industry have become involved. Current technical groups for the Human Factors and Ergonomics Society include, for example, cognitive engineering and decision making, communication, computer systems, consumer products, Internet, performance and perception, system development, test and evaluation, and virtual environments (see http://www.hfes.org).

The concept of human–machine systems implies that all products and systems that involve humans can be divided into subsystems composed of the human aspects and the machine aspects. The focus is on designing interfaces that promote effective human–machine interaction and are consistent with the physical and mental capabilities of the human. HFE relies on basic research in psychology, biomechanics, physiology, and other areas for knowledge regarding the capabilities and limitations of humans (Proctor & Vu, 2003). Examples include the following: (a) central vision is specialized for perceiving detail and peripheral vision for detecting stimulus onsets and movement, (b) the capacity to attend to multiple stimuli is limited, (c) irrelevant features of stimulation often affect performance, (d) items in immediate memory are subject to rapid forgetting, (e) retrieval from long-term memory is dependent on the cues provided, and (f) human reasoning and decision making often rely on use of heuristics that simplify the task but may lead to errors.

The applied research that is at the heart of HFE has the goal of analyzing human performance in ecologically valid environments of the type in which people must operate on a daily basis. For the Internet, the environmental context of interest is the user interacting with Internet applications such as a Web browser or e-mail application. Principles, guidelines, and methods have been developed for assisting designers in designing for human use (Salvendy, 1997), and many techniques have been formulated for evaluating the usability of systems and products at all stages of the design process (Proctor et al., 2002). These principles and techniques are directly pertinent to designing for human interactions with the Internet.

HFE focuses on both physical and cognitive aspects of human use. Physical ergonomics is concerned with accommodating the physical properties of the human body in system design. Examples include design of chairs to be comfortable and support good work postures and design of control panels that allows the switches and controls to be reached and operated easily. Cognitive ergonomics is concerned with accommodating psychological properties of the human in system design, broadly defined to include everything from perception to execution of an action. Examples include use of color coding to distinguish alternative categories of events (e.g., traffic light) and design of decision support systems to aid human decision making.

HFE initially focused primarily on individual operators and users. Over the past decade, however, macroergonomics, which involves organizational and social factors, has received increasing research interest (Hendrick, 1997). Systems exist within an organizational context and often involve interactions among individuals within a group, and successful prediction of human performance requires consideration of these group factors. Among the macroergonomic topics of interest are team cognition and performance. Teams of people are involved in performance of many tasks. Examples include the flight crew of a commercial aircraft and the engineers of a company at different sites working together on a design project. The focus of concern in team cognition is how to enhance communication among team members to optimize performance of the team as a whole.

Human–computer interaction (HCI) has been a topic of burgeoning interest over the past 20 years (Jacko & Sears, 2003). During this period, several organizations devoted to HCI have been established (e.g., the Special Interest Group on Computer–Human Interaction (SIGCHI) of the Association for Computing Machinery; see http://www.acm.org/sigchi/), along with associated journals and conferences. For any task involving computer use, a range of physical and cognitive issues arise. One widely known physical injury associated with computer use is carpal tunnel syndrome. People who operate a keyboard for many hours on a daily basis are at risk for developing this syndrome, which involves neural damage in the area of the wrist. Many factors reduce the probability of developing carpal tunnel syndrome, including use of split, curved keyboards that allow the wrists to be kept straight rather than bent. An example of a cognitive consideration is the development of icons that convey specific actions of the interface to minimize memory demands of where specific commands are located in a menu. Icons can increase the speed in which functions are carried out, but if the icons are not recognizable, then more time is lost by wrong executions. Because of the rapid development of the Internet, an increasing portion of HCI research has been devoted specifically to issues of usability associated with the Internet (Nielsen, 2000).

HUMAN FACTORS AND ERGONOMICS IN DEVELOPMENT OF INTERNET INTERACTIONS

Initially, the Internet was used to enhance communication and transaction between businesses and organizations, and the population of users was restricted primarily to highly educated persons with some background in computers. In recent years, use of the Internet has expanded to the mass population, for which the user characteristics are highly varied. For the novice user, effective use of the World Wide Web can be a daunting challenge. Whereas designing for human use was important initially, it is essential now to allow easy access and use for the whole range of users.

Not only has the population of users expanded, but the variety of uses to which the Internet is put has increased substantially as well. Individuals can use the Internet to send and receive e-mail messages; to buy merchandise; to manage financial accounts; to download files, applications, and programs; to search for specific information or browse information relating to general topics; and so on. Each of these broad categories of activities comprises many specific tasks, which in turn comprise many subtasks. Given the enormous variety of user-initiated interactions with the Internet, it should be clear that a system designed without consideration of the user will not be successful. Evidence consistent with this view has been demonstrated repeatedly in the HCI literature (see, e.g., Najjar, 2001).

Considerable effort has been devoted to applying HFE to different aspects of Web design and evaluation (see Figure 1). Research has been conducted to identify problems associated with browser compatibility, Web site navigation, specific functions and features of a site, and search effectiveness. For example, studies have demonstrated that if a Web page takes a long time to download, users will become frustrated, have increased feelings of being lost, and judge the material to be less interesting and of lower quality than they would otherwise (e.g., Jacko, Sears, & Borella, 2000). To allow faster download times for Web pages, use of graphics and multimedia effects should be kept to minimum (Nielsen, 2000). Although efforts to improve usability with respect to the aforementioned features continue today, issues such as accessibility for different populations of users and promoting security for Web interactions have also become topics of concern. The increased emphasis on security and accessibility for all users is in part a consequence of the emergence of e-commerce in the global economy and use of the Web for educational purposes.

Because e-commerce yields large income for companies, Web sites that are not user-friendly and not perceived as secure will result in loss of income because users abandon the sites. It has been shown that increasing the usability of a Web site improves sales. IBM found a 400% increase in sales by improving the usability of their Web site (Tedeschi, 1999). A Web site that is user-compatible can also improve customers' satisfaction with the site, and Creative Good (2000) showed that by improving the users' experience, the number of customers and order size increases. When users go to a company's Web site in search of a specific product or category of products, they must be able to locate the product easily. Consequently, a major concern of most e-commerce sites is to develop or adopt a search engine that allows users to find their products easily. If the users cannot find what they are looking for, they cannot buy the item and will leave the site. Once the user is able to locate the product, the remaining interactions with the site also need to be user-friendly, or the user will not complete the checkout process and will abandon the items in the "shopping cart." Factors that influence a user's willingness to complete the checkout process include the length of the process, perceptions of security for credit card information, the ease with which information about shipping costs can be located, and privacy issues associated with registration (Najjar, 2001). Poor usability will reduce the number of consumers who purchase products from a business, as well as the likelihood of whether those consumers will return to the site for future purchases.

The Internet has also allowed exchange of data between researchers and collaboration among people at different locations in the world. For example, a researcher can e-mail a manuscript or post it to a Web page so that a colleague can access the file. When this capability first became available, the main benefit was to allow the individuals to exchange files. With such a procedure, however, it is difficult for the receiver to detect the changes made by the sender. Consequently, tools have been

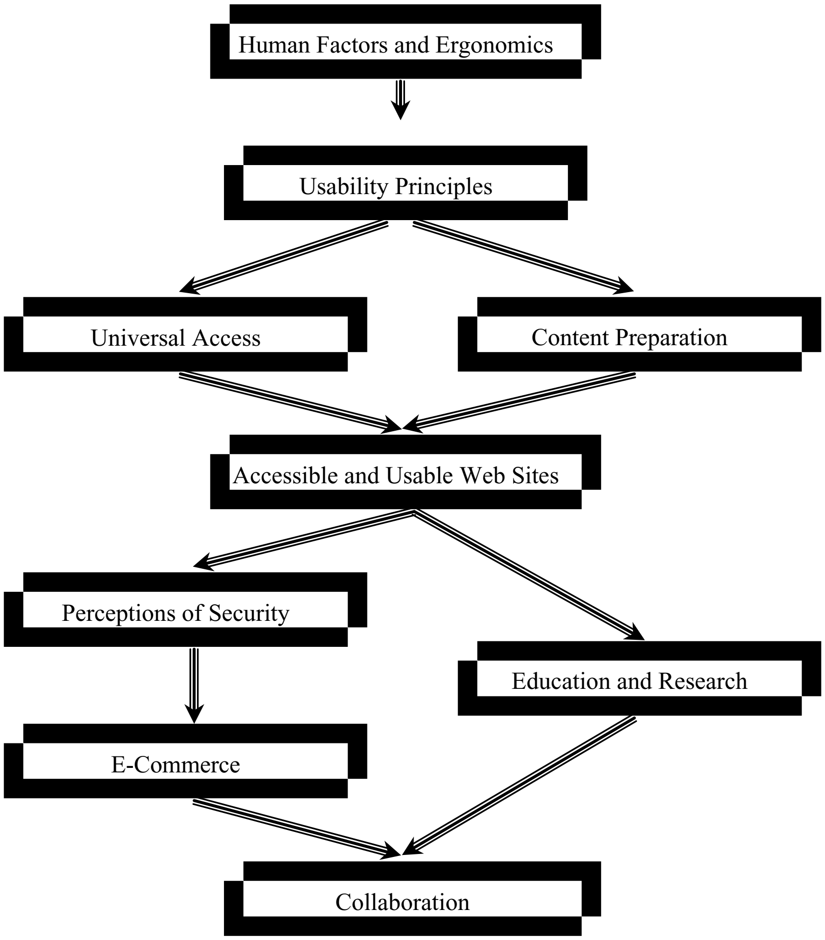


Figure 1: Illustration of how human factors and ergonomics is applied to different aspects of Web design and evaluation.

developed that enable the changes to be marked so that the receiver can locate them easily and either accept or reject them. How to facilitate collaboration on research and design projects is a topic of current interest in research on HCI (e.g., Neuwirth, Chandhok, Morris, Wenger, & Regli, 1998). For example, researchers are developing methods for automated assessment of team cognition using communication data. These methods allow online assessment of team cognition and can be used for designing and evaluating system interfaces and training programs (Kiekel, Cooke, Foltz, & Shope, 2001).

Web sites have also been developed for educational purposes that contain information relevant to material covered in a specific course as well as demonstrations of specific experiments or phenomena relevant to the subject matter (e.g., Birnbaum, 2000). These demonstrations have the potential of helping students understand the material better than traditional text coverage in textbooks, but not if the Web site and demonstrations are hard to use or the format difficult to understand. Sites intended to promote learning must be designed to match human information processing characteristics for them to be effective. For example, because reading from the computer screen is much slower than reading hard copies of the same material, text on the Web should be concise (Nielsen, 2000). The Internet has led to the possibility of better quality for distance education courses in that it can include real-time lectures and interactions between the instructor and the students. Also, the efficient transmission of documents reduces the time lag for students to receive feedback on assignments. Although the technology is available, if it is not usable by the targeted population, it cannot achieve the goals of promoting learning. Because the use and effectiveness of Web-based courses and materials is in its infancy, the continued application and incorporation of human factors principles and usability evaluations are needed to ensure its success in the future.

RETRIEVAL OF INFORMATION: IMPROVING WEB SEARCH

The Internet contains a vast amount of information that is growing and changing daily. Consequently, it is a challenge for users to locate the information they seek and that is most relevant to their concerns. Because of the vast information, a powerful search engine is essential. General and specific search engines for particular sites help users narrow down the information for which they are looking. If the search is successful, the user will find the desired information, but if it is not, the user must repeat the process until the information is found or he or she gives up. For an e-commerce site, with its own search engine, the number of searches in which users must engage before locating the targeted information affects the successfulness of the Web site. According to Nielsen, Snyder, Molich, and Farrell (2000), 33% of e-commerce users give up after three search queries. This statistic has important implications for business and stresses the importance of a good search engine for any large Web site.

Although search engines are important to the success of a Web site, current search engines are for the most part inadequate because they return too many irrelevant hits or provide search options that are difficult to use. Furthermore, search engines tend to be inconsistent with each other. For example, some search engines do not require the use of Boolean connectors such as "and," whereas others do. Furthermore, even if the same terms are entered into the search engines, different hits are often returned. The major problem leading to the poor design of search engines is that they were designed without incorporating knowledge regarding how users search for information (see, e.g., Proctor et al., 2002). Although some research has been devoted to understanding users' interactions with search engines, the results still do not provide an adequate description of what strategies users use to search for information and whether these strategies are specific to the task or generalizable to all Web searches.

Because humans' working memory and attentional capabilities are characterized by a limited capacity to process information, it is important for search engines to minimize the memory and attention loads. Tools should be provided to support users' decision making in high load situations. In addition to using the Web to search for specific information, it can be used to browse topics of general interest. Users approach their tasks with an analytic, goal-directed strategy when searching for specific information but with a more intuitive strategy when browsing. Chen, Wang, Proctor, and Salvendy (1997) found that the mean path length (the number of steps users take through continually using hyperlinks from a start URL) was twice as long for a nonspecific browsing task as for a specific search task. In other words, when browsing, users have a tendency to follow links from successive pages much more than they do when searching for specific information. Moreover, the specific browser functions that are used vary as a function of the type of task. The basic point is that Web sites and search engines need to be designed with both search and browsing in mind.

CONTENT PREPARATION FOR THE WEB

No matter how good a search engine for a particular Web site is, if the information contained in the site is poorly organized and displayed, users will not be able to locate and use the information. Thus, effective content design is required to promote successful interactions of the users with the Web site. Proctor et al. (2002) identified three areas of focus for content preparation in addition to search and retrieval of information: (a) knowledge elicitation, (b) structure and organization of information, and (c) presentation of information.

A Web site should convey information relevant to the user's task in a manner that is easy to understand. To determine what information needs to be conveyed, content designers must identify the information from the domain of interest that is needed to perform various tasks. Knowledge can be elicited from domain experts or from a group of potential users. Experts can specify what information should be conveyed, the way in which it should be structured, and the typical strategies used to retrieve the information. Users can also provide information regarding what information seems to be relevant to them and how this information should be presented. The difference between knowledge elicitation of experts and of potential users is that for the former the emphasis is placed on capturing what the expert knows, whereas for the latter, the emphasis is placed on understanding the knowledge that users bring for interacting with the system to predict their actions.

Because it is often difficult to identify which individuals will be able to provide the knowledge appropriate for different tasks, the process of knowledge elicitation is not simple. For example, the question arises of who the expert is. An expert in one area is not necessarily an expert in another, even if the two areas are related. Furthermore, different experts view issues from different points of view, making it difficult to generalize. Thus, it is best to include several experts in a domain and a range of users in the knowledge elicitation process.

Once information is elicited, it needs to be organized and structured in a manner to promote easy navigation through the structure and easy retrieval of information in the structure. Because each organization has different goals and characteristics that need to be reflected their Web site, the information that needs to be organized may differ in terms of the types of format and structure and need to be integrated into the Web site in a simple and intuitive manner. Users must be able to navigate through the Web site easily. To help communicate the site's structure to the users, the major sections and subsections of the site must be obviously displayed and navigation aides such as breadcrumbs and site maps should be made available (Najjar, 2001; Nielsen, 2000). Information also needs to be structured and organized in a manner that allows the site administrator to update the information easily. More important, the organization and structure of the information must satisfy the objectives and preferences of users or they will not use the site. As a result, the mode of behavior in which the user will be engaged should be determined to help organize the information in a manner that will be advantageous to the user. Although end users are typically not involved in the design of the information architecture of a Web site, their input in early design phases is crucial because the success of the site is often measured by the ability of the end users to interact with it to achieve the intended goals.

The last stage of content design involves presentation of information. The way in which the information is presented can affect the overall success of the Web site because even if the information can be elicited, organized, and retrieved, it needs to be presented in an effective and efficient manner to enhance usability. Preferences in the way that information should be presented differ across individuals. As a result, information can be presented in a way that is intuitive for some users but not others. For example, a Web designer may present information in a way that makes sense to him or her, but this presentation format may be difficult for most end users to understand, in turn making it difficult for them to interact with the site. To remedy this problem, user feedback from the targeted population can be obtained to improve the presentation format used in Web sites. There have been many demonstrations of the effectiveness of incorporating user input into the design of Web sites (see, e.g., Nielsen et al., 2000), but to date, most Web sites are designed without significant user input.

UNIVERSAL ACCESS

Because the Internet allows access of information from anywhere in the world and at any time, the issue of universal access has become a topic of concern (Stephanidis, 2001). Universal access is based on the idea that if Web sites were designed to comply with standard accessibility guidelines (e.g., the W3C-WAI guidelines produced by the World Wide Web Consortium), then the sites could be used by people with various characteristics. An important change in design conceptualization is that Web sites are no longer designed to be useable for the average user, but for all potential users. Thus, universal access requires a user-centered approach to Web design that would satisfy the broadest range of users.

Designing universally accessible Web sites would promote international and multidisciplinary exchange of information and help the expansion of e-commerce to the global market. One way to make Web sites accessible to an international audience is to translate them into different languages. Although this may seem like a trivial task because translation software is available, the translated information must be accurate and match the language structure of the population. It should not be raw translations that do not make much sense or require the user to engage in additional translations. More important, the way in which the information is conveyed when translated should match the mental models of the targeted population. For example, it has been shown that Chinese users prefer concrete knowledge representations, whereas American users prefer more abstract representations (Choong & Salvendy, 1999). Also, the symbology used in Web site should conform to the cultural norms of the population.

Another issue in promoting universal access is to make the Internet accessible to elderly and disabled users. The perceptual and cognitive capabilities of the various user groups must be taken into account. For example, because the ability to perceive detail decreases with age, larger font sizes should be available for elderly users. For users with disabilities, information presented in the Web site must be redundant across modalities (e.g., auditory and visual) so that the user is able to extract the information. Thus, there is a need to evaluate translated and modified versions of the Web site with respect to the users' preferences, abilities, and mental models.

One way to facilitate universal access is to have the interface adapt to the user. Adaptive user interfaces are those that automatically adapt to the behavior, needs, and requirements of users and to different contexts in which the interface is used. Because of the range of users, adaptive systems promote a good fit between the users and the Web interface. Weber and Stephanidis (2001) distinguish between two characteristics of adaptive systems: adaptability and adaptivity. Adaptability refers to the ability of a system to be preprogrammed to recognize known user characteristics. Adaptivity refers to the ability of a system to update its knowledge of the user while he or she interacts with the system and modify the system's behavior accordingly. Although adaptive systems can allow a larger range of users to access a site, adaptivity takes control away from users. If the level of adaptivity is more than what users are willing to tolerate, then they may stop using the site. The development of adaptive systems as means for establishing universal access is still in its initial stages. Thus, it is important that designers take human factors considerations into account when developing these systems to ensure successful interactions of the system with the human.

INFORMATION SECURITY

Because unauthorized users have potential access to any computer connected to the Internet, information security has come to be recognized as an increasingly important aspect of the Internet. Computer security breaches including theft of confidential corporate information, compromise of intellectual property, unauthorized modification of systems and data, and denial of service may occur if systems and the information stored in them are not kept secure. Most of the sophisticated methods that have been developed to provide increased computer security rely on individuals to implement and use them, but many of them were designed without the user in mind. Improper use by end users or system administrators often occurs because the procedures required of them are usually complicated and annoying. Furthermore, users may not be aware of the best methods for increasing security (e.g., how to generate passwords that are difficult to crack) or may just resist them. Consequently, the level of security that is actually provided is much less than the level that the methods could potentially provide. Not too surprisingly, then, the user is often considered to be the main limiting factor in security. Although systematic HFE evaluations to the area of information security are in their infancy, it is clear that that HFE has much to offer for improving the degree of security (Schultz, Proctor, Lien, & Salvendy, 2001).

A properly executed information security program requires appropriate program management and administration. Security administrators must (a) configure security methods appropriately, (b) monitor their status and detect intrusions, and (c) enable features and run add-on applications that increase the level of security. Methods of identification and authentication are concerned with determining who the user is and whether the user should be allowed access to information and systems. The most common method is that of entering a username and password. There are many issues regarding passwords, and they center around the fact that passwords impose a large memory load on the users. Passwords that are easy to remember are also easy to crack (Proctor, Lien, Vu, Schultz, & Salvendy, 2002). Alternatives to the username–password method of authentication include biometric measurements such as fingerprint recognition that are dependent on individual differences in some physical characteristic, and use of smart cards and token devices. All of these methods require additional steps for the user and create the potential for error and user resistance if usability is not considered.

Another concern of information security is ensuring data integrity, confidentiality, and availability. Corruption of stored information, release of information to unauthorized individuals, and lack of availability of information when it is needed can create legal and financial problems. For example, alteration of Web page content can lead to liabilities associated with unauthorized changes advertised prices, and other terms and conditions that can create public relations damage. The most common security method for guarding against unauthorized changes in data is to restrict access by setting permissions or access control lists. Yet the steps for setting permissions and access lists are usually long and complicated, making it difficult for users and administrators to follow them. Another method used to combat the data integrity problem involves computation and storage of cryptographic codes for all of the data on the server. Effective operation of this method requires numerous steps of user interaction, both in implementing programs that check data integrity and checking any output of these programs to determine whether unauthorized data changes have been made. Again, the complexity and workload imposed on the administrator must be kept within reasonable bounds for this method to be effective.

Remote intrusions and unauthorized accesses of databases likely will go unnoticed if there is not an intrusion detection program in place. Such a program can be based on system audit logging and intrusion detection systems, both of which pose high cognitive demands on system administrators. Regular inspection of system audit logs, which provide information about system usage characteristics, can lead system administrators to detect unauthorized usage patterns. The success of system audit logging for intrusion detection depends on the ability of administrators to recognize patterns that indicate intrusions. System logs usually are poorly formatted, displaying lines of monochrome text in uniform size and font, arranged in columns. With this format, it is difficult for system administrators to detect break-ins. It is important to display information in such a way as to make the patterns easier to detect, for example, by using perceptual cues to make critical information “pop out” and to train humans to recognize specific patterns “automatically.”

Information security is of particular concern to e-commerce and its numerous transactions (Garfinkel & Spafford, 1997). Corporations have invested billions of dollars toward providing secure transactions and fraud protection. The focus of providing a secure site for e-commerce is crucial for users to have confidence in purchasing the products offered by the site. Nonetheless, no matter how much security a system can provide, if a system is unusable and not perceived as being secure, users will choose alternative Web sites that are perceived as more usable and secure. In fact, as many as 20% of users reported that they discontinued an online purchase because they felt the site was not secure (Hill, 2001). It has been suggested that sites provide prominent links and symbols to promote information security and privacy (see, e.g., Najjar, 2001).

DETERMINING USER CHARACTERISTICS THROUGH TASK ANALYSIS, USER MODELING, AND USABILITY TESTING

The first step in designing, modifying, or improving a Web interface should be a task analysis to identify the activities that users must perform when interacting with the interface to achieve their tasks (Pearson & Weeks, 2001). The task analysis provides designers with the users’ background knowledge, expectancies, and actions while interacting with the system, allowing the designer to identify the weaknesses of the interface and insights on how to modify the interface to eliminate these weaknesses. A task analysis can begin with observation of users in the task environment to understand the setting in which they interact with the system. Afterward, information regarding specific tasks and activities can be obtained through an interview. The interview can consist of a series of questions and answers or inquiries that are answered while the user performs a specific task. Based on the observations and task descriptions, the evaluator can determine the knowledge needed, the procedures involved, and the users’ preferences for how to perform the task, as well as the assumptions and background knowledge that users bring to the task. Workflow charts can also be developed to capture the dynamic processes involved in performing a task. Information regarding difficulty in performing the task, design flaws, and aspects of the procedure that are not preferred by users can also be obtained.

A task analysis can be used for more formal user modeling to predict users’ performance in the specific context in which the task will be performed. User models can relate system features to users’ needs and be used to compare alternative designs. The most widely implemented user modeling technique in HCI is that of GOMS (goals, operators, methods, selection rules; Card, Moran, & Newell, 1983) analysis and its variants. Using this technique, the task is decomposed into goals and subgoals; the perceptual, cognitive, and motor operators used in performing the task; alternative methods for achieving the goals; and selection rules for determining which method to use. Estimates of the times for the cognitive operators are plugged in to the user models to generate predicted time to perform the tasks. GOMS analysis has been used successfully to predict the relative time required to use alternative interfaces (Pirolli, 1999).

Usability evaluations can improve a Web site by making the functions and goals of the Web site intuitive, effective, and consistent with users' preferences. Usability testing is the most well-known method used to evaluate the ease with which humans can interact with a system (Nielsen, 1997). Usability evaluations can be used to reveal much information about users interacting with the site such as features of the site that are not compatible with users' expectancy, the effectiveness of the site at facilitating users achieve their goals with ease, and changes to the site that would enhance its performance.

Usability testing is important when designing for the Internet because the sample of users is large, making it difficult for any site to maintain a "help desk" for questions. Furthermore, the number of similar and competing Web sites is numerous, making it easy for users to abandon one site if it is difficult to use in favor of a user-friendly site. Despite the benefits of incorporating usability evaluations in the design process, many sites are not including usability evaluations in their design process. There are many reasons usability evaluations are not incorporated throughout the design process, including, but not limited to (a) emphasis on conducting usability evaluations only in late design phases, when changes are difficult and expensive to make; (b) lack of support from management despite demonstrations of the effectiveness of usability evaluations; and (c) the use of guidelines as a substitute for usability testing.

There are many methods used for evaluating the usability of Web sites, which are briefly described next. Some methods are used in the initial stage of the design process, which allows the designer to gather information about the population that the Web site targets. Others methods involve usability experts systematically evaluating the individual Web pages and site to ensure that functions are consistent and follow basic standards and guidelines. Finally, some methods allow the designer to determine how users use and interact with the Web site either directly or indirectly.

Naturalistic and field observations allow designers to see how users interact with a Web site unobtrusively in a natural setting. For example, designers can observe users interacting with a particular Web site on a display computer. Contextual inquiries can be used to supplement observation data that helps the usability engineer understand the background of the user and how he or she will use the Web site. Focus groups usually consist of a group of users who are brought together to discuss different functions and features of a Web site. The group is usually moderated by a usability engineer leads the discussion. Focus groups can reveal information regarding what functions of the Web site are problematic or undesirable and allow brainstorming on how these problems can be remedied.

Questionnaires are often used to gather information about user preferences for the features and presentation of the Web site. With the use of e-mail and Web-based surveys, questionnaires allow quick gathering of preference data from a small or wide range of users. The response rate for questionnaires is typically low, however, and users may not always prefer the most optimal designs. User journals, logs, and diaries involve users writing down the specific activities in which they engage when interacting with the site. The users are asked not only to log their activities, but also to write down their thoughts regarding any interactions with the site. It is important to emphasize that the users be conscientiousness in keeping complete logs to ensure that the data are accurate and that the entries can be understood at a later date.

The usability specialist conducts heuristic analyses and walkthroughs to find problems with a Web site. For heuristic analysis, several usability evaluators interact and examine the interface of the Web site to rate the degree to which the site complies with established guidelines. The evaluators should work independently to increase the likelihood that most problematic aspects will be identified. Heuristic analyses are conducted in early design phases because they provide quick methods to find major usability problems. Whereas heuristic analyses are conducted from a usability specialist perspective, walkthroughs are conducted from user's point of view—that is, the usability specialist pretends to be the user interacting with the site with a set of scenarios. This cognitive walkthrough allows the engineer to note problems that prevent users from completing their tasks. With a pluralistic walkthrough, the usability specialist works with the user when he or she is interacting with the Web site. The users interact with the Web site as they would normally, and the usability specialist records any problems or issues with the product.

Usability lab testing is a technique for assessing the usability of a completed Web site or an incomplete version of a Web site. In a usability lab test, users are given a representative set of scenarios that they would typically perform with the Web site. The session is usually videotaped and mediated by a usability specialist. Performance measures include the time needed to complete a task, whether the user is able to complete the task, and the severity of errors. While the user is interacting with the interface, the user may be asked to think aloud or verbalize his or her thought processes, feelings, preferences, and problems experienced with the site or task. The user also answers questions relating to the usability of the Web site as a whole.

GENERAL GUIDELINES FOR INCORPORATING HUMAN FACTORS AND ERGONOMICS INTO THE DESIGN OF INTERNET SITES AND INTERNET-BASED APPLICATIONS

The material reviewed in this chapter suggests the following general guidelines for incorporating human factors and ergonomics into Web site design and evaluation:

- Determine the targeted population of users.
- Make the Web site compatible with user characteristics and preferences.
- Use established guidelines to aid in the design of Web interfaces.
- Incorporate usability evaluations throughout the design cycle.
- Present information in a concise and accurate manner.

- Structure and organize information in a way that enables easy navigation and access of information within the site.
- Design the site so that the information can be easily updated and maintained.
- Make search options available and ensure that the search features are efficient and effective.
- Ensure that the Web site or application is perceived as secure and that the information stored or transmitted is secure.
- Allow and promote universal access.

More detailed guidelines regarding specific aspects of usability issues and the Internet can be found in a variety of sources, which are cited in papers by Najjar (2001) and Proctor et al. (2002) and summarized in books such as that by Nielsen (2000). Some of the specific methods and guidelines are more effective for some purposes than for others. Thus, readers should make sure that the guidelines and methods they are consulting are appropriate for the intended purpose.

GLOSSARY

Cognitive ergonomics The area of ergonomics that focuses on consideration of psychological properties of users in system design.

Content preparation Preparing a Web site or application to be user-friendly, including determining the information to include, the way to structure this information, how to facilitate its retrieval, and how to present the information.

E-commerce Business transactions conducted via the Web, including business-to-business and business-to-consumer transactions.

Human factors and ergonomics An interdisciplinary field devoted to designing and evaluating products and systems for effective and safe use by humans.

Human–computer interaction An interdisciplinary field devoted to facilitating user interactions with computers.

Information security Methods for protecting information posted, stored, and transmitted electronically that are intended to ensure privacy, appropriate access to systems and information, and accuracy of information.

Macroergonomics The area of ergonomics that focuses on social and group interactions in the system context.

Physical ergonomics The area of ergonomics that focuses on consideration of physical and physiological properties of users in system design.

Task analysis An analysis of user tasks into goals, subgoals, strategies for achieving the goals, and processes involved in completing the task.

Universal access The availability of the Internet for use by anyone from anywhere at anytime and through any Web device.

Usability evaluations Methods that are intended to determine whether Web interfaces and other products are easy to use, easy to learn, and accepted by users.

Usability guidelines Guidelines established to promote usability and to standardize Web interfaces.

Web search A process by which users retrieve desired information from the Web.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Universally Accessible Web Resources: Designing for People with Disabilities; Usability Testing: An Evaluation Process for Internet Communications; Web Search Fundamentals.*

REFERENCES

Birnbaum, M. H. (Ed.). (2000). *Psychological experiments on the Internet*. San Diego, CA: Academic Press.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human–computer interaction*. Hillsdale, NJ: Erlbaum.

Chen, B., Wang, H., Proctor, R. W., & Salvendy, G. (1997). A human-centered approach for designing World Wide Web browsers. *Behavior Research Methods, Instruments & Computers, 29*, 172–179.

Choong, Y.-Y., & Salvendy, G. (1999). Implications for design of computer interfaces for Chinese users in Mainland China. *International Journal of Human–Computer Interaction, 11*, 29–46.

Creative Good. (2000). The dotcom survival guide. *Creative Good*. Retrieved June 12, 2000, from http://www.creativegood.com/survival

Garfinkel, S., & Spafford, G. (1997). *Web security & commerce*. Cambridge, MA: O'Reilly.

Green, R. J., Self, H. C., & Ellifritt, T. S. (1995). *50 years of human engineering*. Wright-Patterson Air Force Base, OH: Armstrong Laboratories.

Hendrick, H. W. (1997). Macroergonomics: A proposed approach for use with anthropotechnology and ergonomic work analysis in effecting technology transfer. *Travail Humain, 60*, 255–272.

Hill, A. (2001). *Top 5 reasons your customers abandon their shopping carts (and what you can do about it)*. ZDNet. Retrieved February 12, 2001, from http://www.zdnet.com/ecommerce/stories/main/0, 10475,2677306, 00.html

Jacko, J. A., & Sears, A. (Eds.). (2003). *The human–computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*. Mahwah, NJ: Erlbaum.

Jacko, J. A., Sears, A., & Borella, M. (2000). The effect of network delay and media on user perceptions of Web resources. *Behaviour and Information Technology, 19*, 427–439.

Kiekel, P.A., Cooke, N. J., Foltz, P. W., & Shope, S. M. (2001). Automating measurement of team cognition through analysis of communication data. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality* (Vol. 1, pp. 1382–1386). Mahwah, NJ: Erlbaum.

Najjar, L. (2001). E-commerce user interface design for the Web. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability evaluation and interface* *design: Cognitive engineering, intelligent agents, and virtual reality* (Vol. 1, pp. 843–847). Mahwah, NJ: Erlbaum.

Neuwirth, C. M., Chandhok, R., Morris, J. H., Wenger, G. C., & Regli, S. H. (1998). Envisioning communication: Task-tailorable representations of communication in asynchronous work (pp. 265–274). *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. New York: Association of Computing Machinery.

Nielsen, J. (1997). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed., pp. 1,543–1,568). New York: Wiley.

Nielsen, J. (2000). *Designing Web usability: The practice of simplicity*. Indianapolis, IN: New Riders.

Nielsen, J., Snyder, C., Molich, R., & Farrell, S. (2000). *E-commerce user experience*. Fremont, CA: Nielsen Norman Group.

Pearson, G., & Weeks, P. (2001). Task analysis: The best first step in user interface design. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality* (Vol. 1, pp. 968–972). Mahwah, NJ: Erlbaum.

Pirolli, P. (1999). Cognitive engineering models and cognitive architectures in human-computer interaction. In F. T. Durso (Ed.), *Handbook of applied cognition*. (pp. 443–477). Chichester, England: Wiley.

Proctor, R. W., Lien, M.-C., Vu, K.-P. L., Schultz, E. E., & Salvendy, G. (2002). Influence of restrictions on password generation and recall. *Behavior Research Methods, Instrumentation, and Computers, 34*, 163–169.

Proctor, R. W., & Van Zandt, T. (1994). *Human factors in simple and complex systems*. Boston, MA: Allyn & Bacon.

Proctor, R. W., & Vu, K.-P. L. (2003). Human information processing: An overview for human–computer interaction. In J. A. Jacko & A. Sears (Eds.), *The Human–computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (pp. 35–51). Mahwah, NJ: Erlbaum.

Proctor, R. W., Vu, K.-P. L., Salvendy, G., & 19 other authors (2002). Content preparation and management for web design: Eliciting, structuring, searching, and displaying information. *International Journal of Human-Computer Interaction, 14*, 25–92.

Salvendy, G. (Ed.) (1997). *Handbook of human factors and ergonomics* (2nd ed.). New York: Wiley.

Schultz, E. E., Proctor, R. W., Lien, M.-C., & Salvendy, G. (2001). Usability and security: An appraisal of usability issues in information security methods. *Computers & Security, 20,* 620–634.

Stephanidis, C. (Ed.). (2001). *Universal access in HCI: Towards an information society for all.* Mahwah, NJ: Erlbaum.

Tedeschi, B. (1999). Good Web site design can lead to healthy sales. *New York Times e-commerce report*. Retrieved February 12, 2003, from http://www.nytimes.com/library/tech/99/08/cyber/commerce/30commerce.html

Weber, H., & Stephanidis, C. (2001). Enabling universal access: Minimum requirements for content preparation. In M. J. Smith, G. Salvendy, D. Harris, & R. J. Koubek (Eds.), *Usability evaluation and interface design: Cognitive engineering, intelligent agents, and virtual reality* (Vol. 1, pp. 1390–1394). Mahwah, NJ: Erlbaum.

Human Resources Management

Dianna L. Stone, *University of Central Florida*
Eduardo Salas, *University of Central Florida*
Linda C. Isenhour, *University of Central Florida*

INTRODUCTION

During the last decade the Internet has had a dramatic effect on our social and economic lives, changing the way we communicate, interact with others, and purchase products. It has also had a profound effect on the way we manage organizational processes including human resources (HR) management systems. For instance, results of recent surveys revealed that 70% of firms now use Internet-based employee self-service (ESS) systems to maintain employee records, 80% conduct online recruiting, and 40% use Web-based portals to communicate HR policies (Cedar, 2001; Towers Perrin, 2001). Although there are numerous reasons to use the Internet to facilitate HR processes, one of the primary reasons is that it helps organizations attract and retain employees, as well as improve employee productivity. Furthermore, researchers contend that the Internet helps organizations (a) streamline HR processes, (b) reduce cycle time, (c) facilitate organization communication, (d) increase employee access to information, and (e) improve the organization's overall ability to adapt to change (cf. Cardy & Miller, 2003; Gueutal, 2003; Stone, Stone-Romero, & Lukaszewski, 2003).

Although the Internet is thought to have a number of unique advantages, some researchers (Stone & Stone-Romero, 1998; Stone et al., 2003) have also argued that the Internet may pose potential threats to personal privacy. For example, the use of Web-based platforms may enable individuals and organizations to gain access to highly sensitive information about employees. In addition, the storage of personal information online may create permanent marks that may stigmatize individuals and preclude them from receiving valued, job-related outcomes (e.g., promotions, pay raises, training opportunities).

Despite the rise in the use of the Internet in the field of human resources management, few studies (Gueutal, 2003; Salas, Kosarzycki, Burke, Fiore, & Stone, 2002; Stone et al., 2003) have assessed the impact of the Internet on HR processes. Furthermore, little research (Eddy, Stone, & Stone-Romero, 1999; Lukaszewski & Stone, 2001; Stone & Stone-Romero, 1998) has directly examined the extent to which Web-based HR systems have the potential to violate employee privacy. Therefore, the primary purposes of this chapter are to (a) describe the role of the Internet in achieving HR system goals, (b) consider the impact of the Internet on the efficiency and effectiveness of HR systems, (c) identify the potential dysfunctional consequences of using the Internet in HR, with emphasis on employee privacy, and (d) offer guidance for future research and practice on the impact of the Internet on HR systems. As a result, in the sections that follow we consider the impact of the Internet on several specific HR processes, including recruitment, selection, performance appraisal, training, compensation and HR planning.

RECRUITMENT

Recruitment has been defined as the process of "searching for and obtaining potential job candidates in sufficient numbers and quality" to fulfill organizational goals (Dowling & Schuler, 1990, p. 47). As noted previously, the Internet is rapidly transforming the recruitment process in three key ways (a) facilitating the attraction of both internal and external job applicants, (b) providing a virtual realistic preview of organizations and jobs, and (c) evaluating the effectiveness of the recruitment process.

Attracting Job Applicants

Although traditional recruitment sources such as newspaper advertisements, private search firms, job fairs, and campus visits are still used to recruit job applicants, the Internet has changed the way in which many organizations announce job openings. For example, in 1997 approximately 58% of large firms posted job vacancies on the Internet; today almost 100% of firms announce job openings via their own Web sites or commercial sites (cf. Cedar, 2001; Gueutal, 2003).

The practice of using the Internet as a recruiting source, often called "e-cruiting," (Galanaki, 2002), started as a means of attracting individuals for technical positions by technology-driven firms such as Cisco Systems and Microsoft but has now expanded to include all types of job openings. Currently, the use of Internet recruitment strategies is thought to be beneficial to internal applicants, external applicants, and the organization as a whole. For example, external job candidates can easily search company Web sites for job openings and compare their knowledge, skill, and ability levels to job requirements. In addition, the use of the Internet can extend traditional searches worldwide, and potential job applicants can easily access job vacancy information anytime of the day or night. Furthermore, the shift to Web-based systems has resulted in a proliferation of national and international job data banks (e.g., Hotjobs.com) that enable applicants to compare and evaluate a wide array of positions (Higgs, Papper, & Carr, 2000). Moreover, some Internet recruiting software facilitates the electronic submission of resumes, bypassing the days or weeks associated with traditional resume handling.

Finally, firms can respond quickly to promising applicants to establish interviews. They can also use the Internet to notify applicants rapidly that their skill or education levels do not meet job requirements. Thus, the use of Internet recruiting may reduce the cycle time associated with traditional recruitment methods and streamline many of the labor-intensive administrative processes (e.g., reviewing resumes, sending follow-up letters to applicants) associated with recruiting. Given that applicants often judge organizations by the time lines and efficiency of administrative processes, a streamlined recruitment system may also have a positive influence on job applicants' image of the organization (Rynes, 1991).

Similarly, Internet recruiting and job posting may facilitate internal recruiting processes. Prior to the use of Internet recruiting, current employees may not have known about job openings within a firm either because they failed to inquire about openings or because that information was not easily accessible. With the advent of the Internet, however, firms are placing their vacancy notices on Web-based portals that may provide more timely job vacancy information to current employees and encourage them to apply for job openings. Given that internal applicants offer some advantages over external applicants, in terms of familiarity with the firm, prior training, and demonstrated job performance, organizations are likely to find that Internet recruiting provides a valuable source of unidentified, qualified internal job candidates. It may also allow organizations to use effectively the skills acquired by employees through training or educational programs (e.g., master of business administration or management of information systems degrees).

Interestingly, some Web-based systems used by large organizations (e.g., SAP, PeopleSoft) have the capacity to search employee records automatically to determine which employees are qualified for job openings. These same systems can then identify a match between employee knowledge, skill, and ability levels and job-based profiles. Finally, they are used to notify qualified applicants about internal job opportunities and inquire about their interest in the job vacancies. As a result, the use of Internet recruiting may increase job opportunities for current employees, enhance their satisfaction levels, and decrease turnover rates.

Furthermore, use of the Internet gives current employees the ability to take advantage of the rapidly growing ESS systems in which the Internet or firm Intranet are used to provide them with a wide array of HR information. For example, ESS systems allow employees to ensure that their records are updated to reflect current training and educational levels. Likewise, ESS systems provide current employees with the opportunity to review training requirements associated with jobs, which should help them plan their careers and enhance their promotion opportunities in organizations (cf. Gueutal, 2003).

Virtual Realistic Previews

As noted previously, another benefit of Internet recruiting is that it may allow organizations to provide a virtual realistic preview to job applicants. Realistic job previews (RJP) are typically used to ensure that the naive expectations of job candidates are brought more in line with organizational realities. Most research on RJPs has shown that their use increases employee satisfaction levels and reduces turnover rates (Rynes, 1991). Thus, organizations have long used organizational brochures or videotapes to provide applicants with a realistic preview of the job or organizational environment. The Internet can easily and inexpensively be used to provide job candidates with a "virtual" preview of a job or show them what it is like to work for the organization. For example, Ford Motor Company presents applicants with a series of typical situations they might encounter on the job to ensure that candidates know what to expect in the organization (Higgs et al., 2000). Such "high-tech–high-touch" approaches to RJPs should provide an important alternative to using brochures that are traditionally less flexible and often more expensive than online systems. In addition, virtual job previews may offer more detailed, current information than traditional brochures and allow candidates to personalize the recruitment process by indicating their preferences for working arrangements (e.g., telework, on-site training). Thus, the use of virtual job previews may help organizations improve the effectiveness of their overall recruitment and retention processes.

Evaluating Recruiting Effectiveness

In addition to helping organizations attract both internal and external candidates for job vacancies and providing virtual previews to job applicants, using the Internet can aid organizations in assessing the effectiveness of recruiting processes and strategies. For example, firms can monitor the Web-based portals of competitors to determine the types of hiring strategies being employed. Such tracking and monitoring capabilities provide insights that would have been difficult to secure in the past and may help organizations remain competitive in the labor market.

Similarly, organizations can use Internet tracking capabilities to build an internal database of candidates for future vacancies to reduce overall costs of future recruiting (Kavanagh, Gueutal, & Tannenbaum, 1990). In addition, these online systems can help HR managers evaluate the cost, timeliness, and effectiveness of the recruiting process. For example, the use of an Internet recruiting systems may provide HR managers or supervisors with reports that will enhance their ability to optimize recruitment strategies. Finally, the use of an Internet recruiting system can help control recruiting costs by improving the ratio of qualified candidates attracted to successful placements made. In summary, the Internet may improve the recruitment process for both job applicants and the organization by attracting a wider array of job applicants, improving the quality and amount of information available to applicants, increasing the speed and timeliness of the recruitment process, and enhancing the firm's ability to evaluate the recruitment process.

Caveats and Implications for Practitioners

Although we have noted a number of advantages of Internet recruiting, we urge HR professionals to understand the challenges that Internet recruiting may pose for organizations. For example, the use of Internet recruiting may result in a deluge of job applications, which can be extremely burdensome for HR departments to process manually. Furthermore, some ethnic minorities (e.g., Hispanic Americans, African Americans, Native Americans) and individuals from low socioeconomic environments may not have access to the Internet, resulting in a "digital divide." Therefore, the use of Internet recruiting may have a potential adverse impact on members of these groups. Thus, Stone et al. (2003) have argued that organizations may want to use Internet recruiting in conjunction with more traditional recruiting methods to ensure that all individuals have equal opportunities for employment with an organization.

In addition, not all Internet recruiting software offers the same capabilities. For example, those systems that will not accept online resumes directly but instead require applicants to "cut, paste, or rekey information" to forward their resumes may actually deter job candidates from applying for jobs. As a result, the degree to which their Web sites are not user-friendly can be a potential disadvantage for some organizations. Similarly, unless HR professionals require automatic search features on their software packages, an important source of internal job candidates may be lost. Likewise, the use of these same systems may serve as an obstacle to enhancing satisfaction and retention of valuable employees when organization policies are not aligned internally.

HR professionals must review policies to identify potential conflicts before introducing Internet recruiting techniques. For example, those organizations with policies that preclude disclosing salary ranges or use broadbanding salary ranges for multiple job levels may find that internal candidates applying for Internet-posted job vacancies are actually applying for jobs that would be a "down-grade." As a result, HR professionals should consider listing open job classification systems with clearly identified salary ranges to ensure that candidates are not discouraged or do not waste time applying for lower level jobs. In addition, managers who view subordinates as belonging solely to their department may perceive automatic vacancy matching as a "raid," threatening their operations. Thus, HR policies must be designed to emphasize cooperation among departments to ensure that all employees are recognized as valued assets to be developed through movement into jobs for which they are qualified.

SELECTION

The selection process has been the focus of considerable research aimed at helping organizations optimize the match between employee knowledge, skills and abilities, and job requirements (Guion, 1991). Thus, selection systems are often used to ensure that employees have the talents and skills needed to perform jobs, which should enhance the overall performance of the organization. Interestingly, Internet selection systems capabilities promise to improve the selection process in three ways: (a) by facilitating selection decisions and reducing biases in the selection process, (b) by reducing the time needed to fill job vacancies, and (c) by increasing the organization's ability to evaluate the validity of inferences made in the selection process.

Facilitating Selection Decisions

As noted previously, the use of the Internet or online systems may facilitate selection decision making in a number of ways. First, online systems may facilitate selection decisions through the use of sophisticated resume screening software. For example, upon receipt of an online resume, Web-based software can be used to scan the resume using key-word and other search techniques to determine if applicants meet basic skill and education requirements associated with job vacancies. When large numbers of applicants are involved, some firms also use interactive voice response technology (e.g., Wonderlic systems) to screen job applicants (Stone et al., 2003). Thus, the use of Internet-based selection systems may decrease the time it takes to fill job vacancies and improve the process of screening large numbers of resumes or job applications.

Second, although most selection systems have historically involved face-to-face interactions between job candidates and decision makers, the Internet has made it possible to gather a wide array of information about candidates in a rapid manner. For example, the Internet has made it easier to conduct background checks that identify criminal convictions or credit-related problems that may be important to success in many jobs (e.g., bank teller, police officer). Likewise it is now simple to verify degrees, licenses, and conduct reference checks that indicate prior employment dates, salary, job duties, and previous training. In addition, the same systems may also allow organizations to conduct online ability testing or conduct assessment centers that measure job applicant performance in a virtual work simulation or virtual team. Some organizations (e.g., Texas Instruments) also establish Web pages that allow job seekers to complete self-assessments to ensure that they are qualified for posted jobs (Higgs et al., 2000).

Furthermore, Ulrich (2001) argued that decision support systems (DSS) can be used to improve the selection process by providing interviewers with advanced information about job candidate strengths and weaknesses to help them structure interviews. Given this information, organizations can conduct initial screening interviews online, which saves interviewer time and recruiting costs. As noted previously, some organizations also use scoring keys to review resumes or application blanks for applicant qualifications. Thus, the use of Internet selection systems may allow organizations to improve the selection process by increasing the amount and timeliness of information available to decision makers.

Apart from decreasing the cycle time associated with the selection process and increasing the amount of information available to decision makers, another important benefit of using Internet selection is that the same systems may increase the quality of selection decisions by decreasing biases inherent in traditional face-to-face systems. As should be evident from our description of the resume screening process, Internet selection systems may be more likely to focus on applicants' knowledge, skills, and abilities rather than on factors unrelated to job performance (e.g., age, race, ethnicity, gender, or disability). Thus, the Internet may ensure that decision makers focus on objective information about applicants rather than information that may be potentially stigmatizing. As a result, the Internet can be a great equalizer and may decrease the biases inherent in face-to-face interactions via the interview or other selection systems. Consequently, the use of Internet selection systems may allow organizations to improve their selection decisions and enhance their overall performance (cf. Kiesler & Sproull, 1992; Spears, Postomes, Lea, & Wolbert, 2002).

Assessing the Validity of Selection Techniques

Selection processes must meet the highest possible standards to comply with legal, moral, and ethical hiring concerns. As a result, organizations may use the monitoring and tracking capabilities of Internet systems to ensure that tests (e.g., interviews, cognitive ability tests, work samples) are valid predictors of success on the job. Interestingly, although an automated screening system may save time and increase efficiency, some researchers have argued that organizations rarely ensure that the screening criteria are effective (Higgs et al., 2000). Thus, HR professionals should ensure that scoring keys are valid or correlated with job success before using them to screen large numbers of applicants.

Furthermore, the Internet may be used to assess the utility or cost-effectiveness of various selection techniques. For example, although assessment centers may be a valid means of selecting employees, their high cost may prohibit organizations from using them for all types of jobs. Use of the Internet to conduct virtual work samples may make such assessment centers more affordable, however, and increase the extent to which organizations use them to screen applicants. Moreover, ESS systems may also make it possible to extend the hiring process to its logical conclusion by assisting the new hire with his or her records and registration for company benefits or training. Ultimately, Internet tracking systems can be used to streamline and improve the effectiveness of the selection process (Gueutal, 2003).

Caveats and Implications for Practitioners

Although we have noted a number of benefits of Internet selection, HR professionals must be aware that applicants may react more negatively to online selection than to traditional selection systems because Internet selection depersonalizes the selection process. As a result, applicants may perceive that they cannot communicate all of their abilities (e.g., interpersonal or communication skills) via the Internet and may resent not being able to control the types of information disclosed in the hiring process. Some research by Martin and Nagao (cited in Higgs et al., 2000) supports this argument. Likewise, Spears et al. (2002) argued that Internet selection may be deindividuating because it does not provide a complete picture of the applicants' skills and abilities. As a result, decision makers may fill in gaps caused by missing or incomplete data and make negative inferences about job candidates' abilities to perform the job (Stone & Stone, 1987). Thus, applicants may be more likely to perceive that Internet selection systems are unfair, when compared with traditional selection systems. Furthermore, HR professionals may want to be aggressive in ensuring that Internet selection systems do not lead to applicant resentment which may preclude them from accepting jobs with the organization. Given that the Internet may not provide an accurate portrayal of an applicants' knowledge, skills, and abilities, we suggest that Internet systems be used in conjunction with more traditional procedures (e.g., situational interviews, work samples) to finalize applicant selections.

Likewise, HR professionals should be cautious when using the Internet to conduct ability testing or personality assessment. First, there is currently no way to guarantee that the applicant being considered is the one actually completing the test or inventory or that the applicant is completing the material without assistance from others. Thus, the integrity of ability test scores, personality inventories, and interview responses may be compromised when conducted via the Internet.

PERFORMANCE MANAGEMENT

Performance management is the umbrella term used to describe the HR function associated with "improving the performance of individuals, teams and the organization as a whole" (Cascio, 1998, p. 154). Performance management functions often include setting performance standards, appraising individual and team performance, giving feedback, and designing compensation and reward systems to motivate performance. The Internet contributes to performance management and performance appraisal in at least three ways: (a) by facilitating evaluation and monitoring of employee performance, (b) by assisting managers with administration and provision of feedback to employees, and (c) by assessing the effectiveness and accuracy of performance appraisals, including the identification of rating errors.

Evaluating Employee Performance and Providing Feedback

Internet systems may help managers monitor and appraise employee performance through the use of management self-service (MSS) systems, which track, analyze, and organize performance data about subordinates (Gueutal, 2003). These same systems may be beneficial because they provide managers with timely data about performance, including the consistency of performance over time, and may encourage managers to provide more frequent feedback to subordinates. Furthermore, the use of MSS systems may enable managers to compare performance across units, allowing them to evaluate more accurately the overall performance of their unit.

Internet performance appraisal systems may also facilitate the use of 360-degree performance feedback, which provides multisource information to individuals from supervisors, peers, customers, and subordinates. Although London and Smither (1995) note that 360-degree feedback has been used primarily for developmental purposes, Internet systems may ensure that feedback flows directly from internal and external customers to employees in a timely fashion. As a result, employees may be able to improve their performance and better meet the needs of their customers. Thus, Internet systems may help organizations reduce the cycle time associated with the performance evaluation process and enable managers to provide more direct and timely feedback to subordinates.

Evaluating the Effectiveness of Performance Appraisal

Apart from the impact on performance evaluation, Internet systems offer organizations the capability of tracking managers' ratings to identify rating errors (e.g., halo, central tendency, or leniency errors). Researchers have long argued that there are biases in the rating process that negatively affect the accuracy of performance ratings and negatively impact the extent to which employees accept feedback (Borman, White, Pulakos, & Oppler, 1991). Thus, the use of Internet systems designed to track and analyze patterns of ratings by individual managers may enable organizations to identify problems associated with ratings (e.g., friendship biases, central tendency, leniency errors) and help improve the accuracy and effectiveness of performance ratings. As a result, Internet and online systems may be used to evaluate the effectiveness of appraisals and provide heightened support for performance management and feedback systems in organizations.

Furthermore, these same online tracking capabilities may be used to design processes that monitor nonexempt employee tardiness, attendance, turnover, and disciplinary problems in organizations as part of a comprehensive performance management program. Ultimately, such attendance systems may also enable managers to improve forecasting of staffing needs and identify potential behavior problems early enough to allow preventive supervisory interventions.

Caveats and Implications for Practitioners

Although we have identified a number of unique advantages of Internet performance appraisal systems, HR professionals should recognize that Internet systems may also pose problems in organizations because such systems may depersonalize the rating process. First, the use of online performance appraisal systems may provide an incomplete picture of an employee's overall performance and neglect important factors such as good citizenship behaviors. Second, employees may react negatively to online performance evaluations because they create a greater psychological distance between raters and ratees, engendering less trust between supervisors and subordinates. For example, when performance is monitored and evaluated online, employees may perceive they have less control over the situation and become suspicious of supervisors' intentions. As a result, they may perceive that electronic monitoring is less fair than traditional appraisal systems and argue that such systems violate their individual rights. Consistent with this argument, research by Ambrose, Alder, and Noel (1998) suggests that individuals typically perceive that electronic performance monitoring is a violation of their rights unless there is justification for such systems and that justification is clearly conveyed to employees. Conversely, organizations that have geographically dispersed employees and teams may find that Internet-based systems aid the remote supervisor in staying abreast of activities and outcomes, thus positively facilitating effective performance management. Indeed, such Internet systems may be the best means of performance management for virtual organizations (Johnson & Isenhour, 2003).

In addition, HR professionals should ensure that use of Internet performance management systems do not inadvertently undermine commitment to legal and ethical organization goals. For example, use of these systems to monitor tardiness or actual hours of work on a daily basis for individuals classified as exempt under the Fair Labor Standards Act guidelines can result in forcing a change in classification from exempt to nonexempt solely because of that management practice. The consequence of this would be immediate company liability for 2 years of back pay for any overtime worked. Moreover, inappropriate use of monitoring tools could leave supervisors and the organization open to charges of invasion of privacy from customers and employees (Stone & Stone-Romero, 1998). Thoughtful consideration of how best to use online performance management, coupled with careful training of supervisors and periodic reviews by managers, can ensure that these systems fulfill their promise of focusing individuals and teams on goal achievement.

Finally, HR professionals need to prepare supervisors and managers adequately for the coproduction requirements associated with many managerial self-service systems (MSS) designed to help view subordinate performance records, access policies and procedures and complete performance appraisal. Although such systems provide a heretofore unimagined access to useful performance management tools, supervisors may be burdened with additional work previously performed by the HR department and perceive the system negatively. Thus, it is essential that managers be trained and become skilled in using the new MSS tools.

COMPENSATION AND BENEFITS

Compensation has been defined as all direct or indirect payments to individuals designed to induce them to join, remain, and contribute effectively to organization goals (March & Simon, 1958). It may include benefits such as health care, paid time off, education assistance, as well as salaries, wages, bonuses, or stock options (Gerhart & Milkovich, 1991). Managers are often concerned with developing compensation systems that simultaneously allow them to reduce labor costs while maintaining a competitive position in attracting and retaining desirable job candidates. Payroll and compensation support systems were among the first information systems developed and remain a crucial part of an organization's information systems today (Gueutal, 2003). Internet and online systems, however, can now provide additional value for managing compensation systems in three ways (a) by modeling the costs and benefits of various salary decisions, (b) by facilitating communication of compensation and benefits information to employees in a timely manner and (c) by giving employees the ability to enroll or change benefits through ESS.

Modeling Compensation Decisions

As noted previously, Internet systems are especially useful in designing compensation systems in organizations because such systems provide managers with the ability to model the costs and benefits of various compensation strategies, as well as evaluate various incentive systems. Furthermore, Internet systems may enable HR professionals and managers to gather salary and benefits data (compensation surveys) within and across industries and track trends in compensation systems (e.g., gainsharing, merit pay, skill-based pay) to remain competitive in the labor market. In addition, such systems can assist in tracking and analyzing the degree to which the compensation systems are effective in achieving organizational objectives, permitting modifications that improve the process on an ongoing basis.

Communicating Information to Employees

Historically, organizations have relied on employee handbooks and other printed materials to communicate compensation and benefit information to employees. Effective communication is essential because employees who fail to understand the rewards and benefits provided will not be motivated by the systems. The use of printed materials is often problematic, however, because the information is quickly outdated, requiring frequent updates each time benefits change. Such changes are not only expensive but often result in numerous inquiries from employees that are time-consuming and expensive to handle. In addition, employees do not always read the printed compensation materials, leaving them unaware of the compensation or benefits they receive in an organization. The use of Internet compensation systems and Web-based portals may greatly alleviate these problems and improve the extent to which employees receive timely communication about their compensation and benefits. Such communication is likely to enhance employee perceptions of the value of their benefits and provide a greater breadth of information through the use of such tools as FAQ (frequently asked questions) or online links to organization call centers. Furthermore, the ease with which Web sites can be updated makes the benefits communication process more flexible and may reduce overall organization costs. Although more than half of all organizations in North America use Web-based portals as a means of providing HR information, including benefits, to employees (Cedar, 2001), older technologies, such as interactive voice response and toll-free telephone systems, are still prevalent.

Use of Employee Self-Service Systems

One of the fastest growing trends in human resources management is the use of Web portals or ESS (Gueutal, 2003). The growing use of ESS systems parallels the rise in flexible (e.g., cafeteria) benefit plans aimed at meeting the needs of an increasingly diverse workforce. In particular, flexible benefit systems allow employees to choose among several benefit plans that may meet their individual needs. For example, a married employee may choose increased health care benefits, whereas a single employee may choose to allocate funds to supplemental retirement programs (e.g., 401K). The increased flexibility of cafeteria benefit plans has also increased their complexity, however, often including numerous providers and frequent changes in plan terms and conditions. The new Web-based ESS systems have enabled organizations to manage benefit systems more effectively. In particular, ESS capabilities permit employees to research and compare alternative benefit coverage, establish initial enrollment or change plans at a time or place convenient for them, with full round-the-clock access via the Internet. Access to other benefits, including 401K investment account information and retirement planning calculators, is frequently provided as well, including online links from the HR portal to remote service providers. In fact, such approaches to benefits management allow employees increased control over their benefits and may reduce the time it takes to make changes or update benefit plans. In general, increased Web-based system capabilities promise to expand the ways in which employees provide and receive information about all aspects of their work lives not just benefits (e.g., training, career planning and advancement opportunities).

Caveats and Implications for Practitioners

HR professionals have long used compensation systems as a means to motivate and retain employees. The use of Internet HR systems has the potential to shift much of the traditional HR compensation work to line managers. Explaining the advantages of moving this work and providing training to line managers will be essential in making such a transition. For example, one especially useful feature of Internet systems for the supervisor is the fully electronic processing of periodic salary increases or changes. The new systems will allow managers who are knowledgeable about a subordinate's performance to ensure that salary increases are equitable and consistent with organization policies. Likewise, Internet systems may be particularly useful for managers with subordinates in remote geographic locations or those who travel extensively and must perform their work via Internet links. Similarly, the move to global, geographically dispersed virtual organizations makes the communication of compensation practices and benefits particularly difficult. Offering ESS capabilities via the Internet can help bring such dispersed groups together, emphasizing the common link that all share with the organization.

TRAINING

Although the Internet has certainly influenced recruitment, selection, appraisal, and compensation processes, it may have had its most profound effect on the design and delivery of training in organizations. The primary reasons for this are that (a) organizations have transitioned from an industrial to a global, knowledge-based economy that relies heavily on the skills and abilities of employees and (b) training is a critical means of ensuring that employees have the skills needed to help organizations adapt to changes in the new environment (Salas & Cannon-Bowers, 2001). Furthermore, the rapid proliferation of technology has facilitated the provision of training via alternative electronic methods. Taken together, these trends have prompted organizations to place increased emphasis on enhancing employee knowledge, skills, and abilities and inspired them to use distance learning as a means of delivering training in a timely and cost-effective manner (PriceWaterhouseCoopers, 2002). Distance learning is often used as an umbrella term that includes e-learning (e.g., Web-based learning, virtual classrooms, computer-based learning, and digital collaboration) as well as delivery of training content via a variety of media (e.g., videotape, interactive television, Internet/intranet, CD-ROM, satellite broadcasts, etc.; Kosarzycki, Salas, DeRouin, & Fiore, 2003). Distance learning does not preclude the use of traditional classroom instruction, however, and frequently encompasses the integration of a variety of electronic media with face-to-face instruction.

As should be evident from these comments, the use of the Internet to facilitate distance learning is thought to have a number of benefits in organizations including (a) reduced training cycle times allowing the delivery of just-in-time and on-demand training, (b) increased flexibility allowing employees to learn at their own pace and control the sequence of instruction, (c) decreased training costs, and (f) improved online enrollment and tracking systems to facilitate training evaluation. These benefits, as well as specific issues associated with system capabilities and training content, are discussed in the following paragraphs.

Reduced Training Cycle Time

One of the potential advantages of using the Internet for distance learning is that the time needed to conduct training can be greatly reduced (Driscoll, 1999). For example, when a training need is identified, distance learning can be conducted immediately (e.g., just in time) in a consistent manner for employees throughout the world. As a result, the competency levels of all employees can be enhanced in a timely, efficient manner. Furthermore, distance learning may be especially useful when training is deconstructed into small content modules that can be administered over time, without loss of learning effectiveness. Small procedural changes requiring training, which may have been deferred because of the expense of modifying training materials and transporting employees to an off-site training center, can be administered more expeditiously via distance learning than traditional classroom methods (Peterson, Marostica, & Callahan, 1999). Moreover, desirable training capabilities, such as trainer–trainee or trainee–trainee interactions can also be facilitated via Internet e-mail or listserv systems. As a result, trainees may actually be able to communicate more frequently with trainers using distance learning than traditional training techniques.

Increased Training Flexibility

Another advantage of distance learning is that it may increase the flexibility of training delivery and allow for self-paced instruction (Driscoll, 1999). For example, the use of online enrollment and tracking systems allows employees to choose the timing and amount of training to be conducted at any point in time. Employees can complete all or any part of the training during a given session. Furthermore, employees may receive training via the Internet while traveling away from their primary work locations or elect to take the training in segments appropriate to their stage of development in a job. This unconstrained time-and-place approach to training may be particularly attractive for those employees who want to improve their skills but also need to balance work and family demands. Likewise, distance learning may enhance individual performance because employees are not absent from their jobs for long periods of time or can be trained during off-peak hours. Given that distance learning typically provides online skill assessment and feedback, employees may also be able to use such information to target or enhance their career opportunities (Anderson, 2002).

Decreased Training Costs

Another important benefit of distance learning is that it is thought to decrease the overall costs of training (Driscoll, 1999). For example, distance learning programs often result in economies of scale (e.g., standardized training programs) that are likely to reduce the overall expense of training in organizations. In addition, distance learning may also reduce travel costs, labor costs, and the cost of equipment or facilities (e.g., classroom space). Distance learning may also allow organizations to save on the costs of printing training materials and reduce costs by using the existing technical infrastructure (e.g., computers, networks) to facilitate training. As a result, Driscoll (1999) estimates that distance learning often results in an average cost savings of $1,500 per employee. In reported examples, Eli Lilly saved $800,000 in travel and salary costs, Hewlett-Packard saved $5.5 million, and Aetna saved $3 million, while training approximately 3,000 workers (Horton, 2000).

Evaluation of Training

A final benefit of using the Internet for distance learning is that such systems can facilitate the evaluation of training. Although distance learning may be less expensive, organizations are justifiably concerned with determining the return on investment associated with its use. For example, IBM noted a 20-fold return on its management training program conducted via distance learning (Vaas, 2001), whereas GE estimated annual savings of $1 million by using distance learning for its geographically dispersed auditing organization (Lohr, 2002).

Although return on investment is an important criterion, organizations are also concerned with the degree to which distance learning facilitates employee learning and skill development. Therefore, organizations may want to evaluate distance learning in terms of Kirkpatrick's (1983) four evaluation criteria, which include reactions, learning, behavior, and outcomes. First, reactions criteria typically refer to trainee reactions to the overall training and might address the question of whether employees react more positively (or negatively) to distance learning than traditional classroom training. Second, Kirkpatrick argues that training should be evaluated in terms of the overall level of learning. Thus, distance learning could be evaluated in terms of the level of declarative or procedural knowledge acquired through such a program. Alternatively, learning acquired through distance learning could be assessed relative to other training methods (e.g., traditional classroom instruction, CD-ROM). In addition, Kirkpatrick contended that training should be evaluated in terms of behavioral criteria, referring to the degree to which training transfers back to the work setting (Hall & LeCavalier, 2000). As a result, organizations may want to assess the degree to which knowledge or skills acquired through distance learning affect on-the-job performance. Finally, Kirkpatrick suggests that training should be evaluated in terms of outcomes, referring to the degree to which training has a positive impact on organization goals. Therefore, organizations may want to evaluate the degree to which distance learning affects important organizational outcomes such as product quality, customer satisfaction, or organization adaptability. Given that Kirkpatrick and others (cf. Kraiger, Ford, & Salas, 1993) encourage the use of multiple criteria to evaluate distance learning, we suggest that organizations move beyond traditional return on investment measures alone and incorporate multiple evaluation criteria to assess the overall costs and benefits of using distance learning programs.

Apart from the advantages associated with using distance learning noted here, researchers (Kosarzycki et al., 2003; Salas et al., 2002) have argued that specific design and implementation issues noted in the following section should be considered by organizations proposing to use distance learning.

Caveats and Implications for Practitioners

Two primary issues facing organizations interested in implementing distance learning are (a) the fragmentation of distance learning systems and (b) the types of training content that can be delivered most effectively via distance learning. First, although distance learning is widely embraced across organizations, distance learning system capabilities vary widely. As a result, training professionals are often forced to focus on identifying and securing those functions that are required for the specific entity. High-quality products are available at lower prices for those willing to expend effort matching products to training need, however (e.g., collaborative capability, student tracking, interoperability with existing hardware and software; Adkins, 2002). In addition, purchasers of distance learning systems must be wary of the disappearance of a preferred vendor and the resulting support problems that might ensue. Furthermore, organizations should be aware that the distance learning industry is currently moving toward consolidation and is striving to establish standards on which to base future growth (Dobbs, 2000).

In light of the industry fragmentation and instability, some training professionals are now looking to a new segment of the distance learning market known as learning management systems (LMS). Designed to coordinate courses across multiple providers, LMS vendors offer management reports, user registration, and tracking capabilities and a framework within which content can be adapted or developed via learning object format to meet the learning needs of a variety of functional groups within an organization. At the same time, learning portals (i.e., Web sites that act as a single contact point for multiple sources of content, courses, and learning information; Mantyla, 2000) are becoming more prevalent. LMS, in concert with the growing number of learning portals, promise increased opportunity and ease of use for distance learning adopters (Kaplan-Leiserson, 2002).

A second issue associated with the design of distance learning systems is the type of training content that can be delivered most effectively via distance learning programs. Although some experts argue that distance learning is suitable for all types of training content (Moe & Blodget, 2000), others (Schreiber & Berge, 1998) contend that training content should be aligned with desired type of learning (e.g., declarative knowledge, procedural knowledge, or strategic knowledge) and the skill level of the trainee. For example, employees with low computer skill levels may find distance learning more difficult and intimidating than those with high computer skill levels.

Acquiring declarative knowledge, with its emphasis on structured domains of facts, with clear "right or wrong" answers, can easily be facilitated by e-learning. Topics include such wide-ranging content as foreign languages, software coding, and mathematics. Similarly, procedural knowledge can also be taught via distance learning (e.g., safety compliance training, business writing; Conners, 2001). As the knowledge being acquired becomes less structured (e.g., strategic planning), however, some researchers (Anderson, 2002) maintain that a combination of methods (e.g., classroom instruction, chat rooms) should be used in conjunction with distance learning to help trainees learn, analyze, and synthesize the material.

Given that e-learning may not be appropriate for all types of training content, some organizations are pursuing a hybrid of instructor-led and distance learning known as "blended learning" (Kaplan-Leiserson, 2002). The hybrid approach seems particularly appropriate for learning interpersonal skills, which are primarily procedural and conceptual rather than declarative. Such a blended approach would include having the trainees review and complete tests on the theory and underlying concepts to be learned (e.g., communication techniques) to prepare themselves for instructor-led classroom experiences (Prentice, 2001). Other researchers have suggested that distance learning can be used to supplement classroom instruction in order to change trainee attitudes and increase motivation (Anderson, 2002). For example, online discussions could be used in conjunction with instructor-led training to advance trainee skills associated with evaluating alternative values and understanding different perspectives (Duckworth, 2001). However, Duckworth (2001) also maintains that training of psychomotor skills will generally require live practice opportunities (e.g., donning safety gear properly, administering cardiopulmonary resuscitation (CPR), flying an airplane), even though some declarative knowledge (e.g., steps in the CPR process) may be obtained via distance learning. Overall, distance learning promises to revolutionize how organizations upgrade competencies in the battle for competitive advantage.

HUMAN RESOURCE PLANNING

Human resource planning is recognized as the link between organizational strategic planning and detailed HR program decisions (Kavanagh et al., 1990). Two primary areas of HR planning affected by the use of Internet and online systems include (a) forecasting and (b) legal compliance tracking.

Forecasting

There are a number of advantages of using the Internet for HR planning in organizations. First, the Internet helps ensure that organizations use employee skills through the use of workforce skills inventories that help HR professionals identify the number and skill levels of current employees (Kavanagh et al., 1990). These analyses assist organizations with filling anticipated vacancies and allow organizations to meet new staffing requirements rapidly. They also provide an important means for increasing the flexibility of organizations and facilitating an organization's ability to adapt to strategic change.

Second, the Internet helps organizations forecast labor supply and demand to project future staffing needs. As a result, Internet systems help ensure that human resources management functions are aligned with the strategic goals of the organization. These systems are also used to assist managers in succession planning, which typically involves identifying and targeting individuals as replacements for key managerial vacancies caused by retirements, promotions, or voluntary turnover. Through the use of succession planning, career plans can be developed for "high potential" individuals to ensure that they are properly trained or prepared to fill important vacancies (Kavanagh et al., 1990).

Legal Compliance Tracking

Another benefit of Internet and online systems is that they are a valuable means of monitoring, tracking, and reporting compliance with government laws and regulations (e.g., EEO, Affirmative Action, OSHA). For example, OSHA compliance reports are required periodically, but information must be available at any time for unannounced inspections or in conjunction with accident investigations. Failure to have up-to-date, accurate data can result in fines or other serious consequences.

In addition, Internet systems make HR data available to managers and help identify deficiencies or problems, allowing managers to formulate corrective action plans. For example, managers can track hiring rates by department and assess the possibility of adverse impact on protected groups. Likewise, managers can analyze turnover rates and uncover potential supervisory or job design problems. Without such systems support, managers may only learn about problems after lawsuits are filed or the organization is subjected to a governmental audit. Thus, Internet systems help managers become proactive in solving problems and avoiding legal sanctions. In general, HR planning can be greatly enhanced through the judicious use of Internet and online systems.

PRIVACY CONCERNS

As noted previously one of the potentially dysfunctional consequences of using the Internet for HR purposes is that employees may perceive such systems as an invasion of privacy. In particular, Stone et al. (2003) have argued that employees are concerned about the extent to which Internet HR systems permit access or disclosure of personal information or contain inaccurate data that may potentially stigmatize individuals or lead to other negative consequences.

Access or Disclosure of Personal Information

Researchers (Stone & Stone-Romero, 1998) have long argued that employees are concerned about the storage of data in HR systems because these systems contain a great deal of personal information about them, including data about their creditworthiness, benefits, performance, medical history, and family background. Thus, employees often fear that HR Web-based systems are insecure, allowing users (both inside and outside the organization) to gain access to their personal information. As a consequence, employees may perceive that if some individuals access their personal data they will be stigmatized or experience other negative outcomes (e.g., embarrassment, loss of job opportunities). Furthermore, employees are often concerned that organizations may release personal information about them to third parties (e.g., credit grantors, courts, potential employers).

Interestingly, there is considerable evidence that employee concerns about organizational access and disclosure of personal information are warranted. For example, surveys suggest that organizations collect and store a great deal of personal information about employees and do release the data to third parties (cf. Stone et al., 2003). In particular, a survey by the Society for Human Resources Management (2000) revealed that 34% of firms collect and store medical and prescription drug information about employees, often releasing such information to insurance companies and potential employers. Similarly, surveys show (Electronic Privacy Information Center, 2000) that 70% of employers disclose information to creditors, 47% release information to landlords and 19% disclose data to charitable organizations. Furthermore, 60% of employers do not inform employees that they disclose such personal data (Society for Human Resources Management, 2000).

Apart from the survey results reported here, findings of studies (Eddy et al., 1999) on employee reactions to human resources information systems (HRIS) revealed that individuals were more likely to perceive that HRIS were unfair and an invasion of privacy when they were unable to control the release of information, when data were released outside the organization, and when employees were unable to authorize the release of information. Similarly, results of a study by Stone, Lukaszewski, and Stone-Romero (2001) indicated that an HRIS was considered most invasive of privacy when (a) employment data were disclosed to supervisors rather than HR representatives, (b) the same data were used for HR planning purposes rather than emergency notification purposes, and (c) employees had no ability to check the accuracy of the data in the HRIS. In addition, results of a study by Lukaszewski and Stone (2001) suggested that employees were more likely to perceive their privacy had been invaded when medical data were collected and stored in an HRIS and when they had no choice about whether the data would be stored in a Web-based system or a traditional paper file system.

Data Accuracy

Another potential concern with the use of Web-based HR systems is that such systems often contain inaccurate data about employees that may stigmatize them or result in negative outcomes (Stone et al., 2003). Consistent with this concern, results of survey research suggest that some data in HR systems may be inaccurate and employees are often unable to access and correct their personal data (e.g., Electronic Privacy Information Center, 2000). For example, a study by Pillar (1993) indicated that 48% of credit data are inaccurate, including the data used for employment purposes. Likewise, a survey by Linowes (Electronic Privacy Information Center, 2000) showed that 72% of firms do not allow employees to access and correct records. Thus, employees may be justifiably concerned that HR systems used to store employment data may contain inaccuracies that can potentially affect their outcomes in organizations.

Caveats and Implications for Practitioners

Given the growing use of Internet-based HR systems, HR professionals and Web site developers must be cognizant that employees are often concerned about the privacy of their personal data. Although recent court cases have affirmed the organization's right to monitor employee e-mails and Internet usage, privacy concerns are likely to receive increasing scrutiny (Stone & Stone-Romero, 1998). Protection of individual information was easier when most records were on paper in the HR department or in legacy systems for which there was no easy external access. With the increased use of Internet HR systems and Web-based portals, access to personal or sensitive information about employees is more of a potential problem than ever before. Thus, organizations must ensure that particular care is taken to safeguard personal information about employees. Although there is no federal privacy law that applies to the private sector, several states have passed legislation giving employees the right to access, review and amend their employment records (cf. Stone & Stone-Romero, 1998). Also, the newly passed European Data Act places restrictions on the transmission of employee data across national boundaries. As a result, HR professionals should be proactive in developing policies to protect employee privacy and ensure that employee data are recorded accurately and handled fairly.

IMPLICATIONS FOR RESEARCH

Although we have described a number of advantages of using Internet HR systems, we indicated previously that little theoretical or empirical research has focused on the effectiveness of such systems. Thus, we believe that future research should be conducted to examine the effectiveness of many of these systems (e.g., Internet recruiting, distance learning, ESS, and MSS). Furthermore, research is needed to assess the degree to which Web-based systems and portals are a more effective means of communicating HR policies and practices than traditional communication systems. Likewise, given that Internet-based HR systems often transfer traditional HR work to employees and managers, research should focus on the reactions of internal customers (e.g., employees, managers, HR staff) to these systems. In particular, employee acceptance may be critical to the success of ESS systems designed to allow employees to change or review benefits. Likewise, managerial reactions to the increased workload (i.e., coproduction) and data provided by MSS systems may also be a key to the successful implementation of the new HR systems. Furthermore, given the widespread use of distance learning in industry and academia, research is needed to examine (a) the effectiveness of these systems in terms of key learning criteria (e.g., trainee reactions, behavior, outcomes), (b) the types of training content that can be effectively delivered through distance learning, and (c) the extent to which trainee individual differences affect the success of distance learning. Finally, research is needed to examine strategies for reducing employees' fears and concerns about the privacy of Web-based HR systems.

CONCLUSION

We have reviewed the impact of the Internet on several key HR processes, including recruitment, selection, performance management, compensation, training, and HR planning. Overall, our review revealed that the use of the Internet promises to (a) reduce the cycle time associated with many HR processes, (b) increase the flexibility and effectiveness of these systems, and (c) provide better service to employees and managers. Despite the many benefits of using Web-based HR systems, we also noted that the Internet may pose a number of new problems for organizations, including negative reactions from employees due to violations of privacy and the increased depersonalization of HR systems.

Although we have attempted to describe the impact of the Internet on HR systems, we must stress that we have only begun to analyze the effects of the Internet on organizational processes. As the Internet grows and develops, we believe it will offer numerous benefits and pose a host of unknown challenges for its users. We also believe, however, that HR professionals should be cognizant of the following issues when implementing and using Web-based HR systems. First, HR professionals should stress the importance of aligning Internet-based HR systems with organization goals. For example, they must consider whether the system selected is appropriate for the type and size of the organization and determine whether the system will help support its mission. Second, HR professionals should ascertain whether the organizational infrastructure will support the new technology. For example, HR professionals will need to ensure that they have financial resources needed to implement Web-based HR systems, including hardware, software, training, and ongoing technical support. Likewise, given that technology evolves rapidly, they may want to determine whether the system is adaptable and can be expanded to meet future needs. Finally, HR professionals will want to ensure that this change is managed effectively, which means that key stakeholders (e.g., employees, managers, HR staff) should be involved in every step of the transition from traditional to Internet-based HR systems.

In conclusion, we believe that our chapter has shed new light on the ways that the Internet will affect the practice of human resources management and hope that it has sparked an interest in scientific research that will help evaluate the effectiveness and acceptability of these new systems.

GLOSSARY

Blended learning A hybrid approach to training that combines instructor-led and distance learning techniques to facilitate types of learning (e.g., interpersonal skills) which are primarily procedural and conceptual rather than declarative).

Digital divide The term used to describe the lack of access to computer and Internet technologies that affects some ethnic minorities and individuals from low socioeconomic environments.

Distance learning The umbrella term used to describe a variety of electronic learning (e-learning) techniques and tools intended to facilitate specific skill and knowledge enhancement efforts.

Electronic learning (e-learning) The term used to describe a variety of Internet- and computer-based learning techniques, including Web-based learning, virtual classrooms, computer-based learning, CD/DVD learning, to facilitate individual and group acquisition of specific skills and knowledge.

Electronic monitoring The process of collecting and using computer-based performance measures as part of individual performance management efforts.

Electronic recruiting (e-cruiting) The term used to identify Internet recruiting as a formal sourcing of online job information.

Employee self-service systems (ESS) Internet-linked, coproduction computer systems that provide firm employees with a variety of human resources program and policy information and access to multiple sources of relevant personal information maintained by the employing firm.

Human resources information systems (HRIS) "Systems used to acquire, store, manipulate, analyze, retrieve, and distribute pertinent information regarding an organization's human resources . . . including people, forms, policies and procedures, and data for the purpose of providing service in the form of accurate, timely information to clients for supporting strategic, tactical, operational decision making (selection, recruitment), evaluating programs, policies or practices (compensation plan comparison, sick leave costs, diversity hiring), and enhancing efficient fulfillment of daily functions (time reports, attendance, performance management)" (Kavanagh, et al. 1990, pp. 29–30).

Learning management systems (LMS) Internet-based computer systems designed to coordinate training course offerings and management across multiple providers.

Management self-service systems (MSS) Internet-linked, coproduction computer systems that provide support to firm managers in fulfilling desired human resources goals and policies, including tracking, analyzing, and organizing data about subordinates.

Privacy the expectation by individual employees that personal information maintained by employers for a variety of legitimate purposes will be closely protected by human resources professionals and will be divulged only under limited circumstances to internal sources.

Realistic job previews (RJP) are one of several ways firms may attempt to manage expectations of job candidates regarding actual job responsibilities and firm working environments prior to job offers as a means of increasing job acceptance, enhancing initial new-hire socialization tactics, and reducing early-stage voluntary turnover.

Stigmatized the term used to describe individual physical, psychological, or sociological marks that may preclude full inclusion and unbiased treatment in organizations.

Three-hundred-sixty (360) degree feedback The term used to describe multiple sources of performance information from supervisors, peers, customers, and subordinates provided to individuals as part of performance management efforts.

Web-based portals Internet sites maintained by firms for the specific purpose of encouraging contact with both internal and external stakeholders.

CROSS REFERENCES

See *Digital Divide; Distance Learning (Virtual Learning); Internet Literacy; Internet Navigation (Basics, Services, and Portals); Legal, Social and Ethical Issues; Privacy Law.*

REFERENCES

Adkins, S. (2002). Market analysis of the 2002 U.S. e-learning industry: Convergence, consolidation and commoditization. *Brandon-Hall 2002 Market Analysis Series*. Retrieved January 30, 2002, from http://www.brandon-hall.com/public/execsum_simmarket.pdf

Ambrose, M., Alder, G., & Noel, T. (1998). Electronic performance monitoring: A consideration of rights. In M. Schminke (Ed.), *Managerial ethics: Moral management of people and processes* (pp. 61–80). Mahwah, NJ: Erlbaum.

Anderson, T. (2002, January). Is e-learning right for your organization? *Learning circuits*. Retrieved February 4, 2002, from http://www.learningcircuits.org/2002/jan2002/anderson.html

Borman, W., White, L., Pulakos, E., & Oppler, S. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76,* 863–72.

Cardy, R., & Miller, J. (2003). Technology: Implications for HRM. In D. Stone (Ed.), *Advances in human performance and cognitive engineering research Vol 3* (pp. 99–118). Amsterdam: JAI.

Cascio, W. (1998). *Applied psychology in human resource management*. Upper Saddle River, NJ: Prentice-Hall.

Cedar. (2001). *Cedar 2001 human resources self-service/portal survey: Fourth annual survey.* Baltimore, MD: Author.

Connors, K. (2001). Online learning can be a cost-effective alternative to traditional training. *Managed Healthcare Executive, 11*(7), 45–46.

Dobbs, K. (2000). What the online world needs now: Quality. *Training, 37*(9), 84–94.

Dowling, P., & Schuler, R. (1990). *International dimensions of human resource management*. Boston: PWS-Kent.

Driscoll, M. (1999). Web-based training in the work place. *Adult Learning, 10*(4), 21–25.

Duckworth, C. L. (2001, April). ISD for live e-learning. *Learning Circuits*. Retrieved December 12, 2001, from http://www.learningcircuits.org/2001/apr2001/duckworth.html

Eddy, E., Stone, D., & Stone-Romero, E. (1999). The effects of information management policies on reactions to human resource information systems: An integration of privacy and procedural justice perspectives. *Personnel Psychology, 52,* 335–358.

Galanaki, E. (2002). The decision to recruit online: a descriptive study. *Career Development International, 7,* 243–251.

Gerhart, B., & Milkovich, G. (1991). Employee compensation: Research and practice. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 481–569). Palo Alto, CA: Consulting Psychology Press.

Gueutal, H. (2003). The brave new world of eHR. In D. Stone (Ed.), *Advances in human performance and cognitive engineering research,* Vol 3 (pp. 13–36). Amsterdam: JAI.

Guion, R. (1991). Personnel assessment, selection, and placement. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 327–97). Palo Alto, CA: Consulting Psychologists Press.

Hall, B., & LeCavalier, J. (2000). The case for level 3. *Learning circuits*. Retrieved December 12, 2001, from http://www.learningcircuits.org/nov2000/hall.html

Higgs, A., Papper, E., & Carr, L. (2000). Integrating selection with other organizational systems. In J. Kehoe (Ed.), *Managing selection in changing organizations* (pp. 73–122). San Francisco: Jossey-Bass.

Horton, W. (2000). *Designing Web-based training: How to teach anyone anything anywhere anytime.* New York: Wiley.

Johnson, R., & Isenhour, L. (2003). Changing the rules? Human resources in the 21st century virtual organization. In D. Stone (Ed.), *Advances in human performance and cognitive engineering research,* Vol 3 (pp. 119–152). Amsterdam: JAI.

Kaplan-Leiserson, E. (2002). E-learning glossary. *Learning Circuits.* Retrieved February 21, 2001, from http://www.learningcircuits.org/glossary.html#P

Kavanagh, M., Gueutal, H., & Tannenbaum, S. (1990). *Human resource information systems: Development and application.* Boston: Kent.

Kiesler, S., & Sproull, I. (1992). Group decision making and communication technology. *Organizational Behavior and Human Decision Processes, 52,* 96–123.

Kirkpatrick, D. (1983). Four steps to measuring training effectiveness. *Personnel Administrator, 28*(1), 19–25.

Kosarzycki, M., Salas, E., DeRouin, R., and Fiore, S. (2003). Distance learning in organizations: A review and future needs. In D. Stone (Ed.), *Advances in human performance and cognitive engineering research*, Vol 3 (pp. 69–98). Amsterdam: JAI.

Kraiger, K., Ford, J., & Salas, E. (1993). Applications of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78,* 311–28.

Electronic Privacy Information Center. (1996, April 22). Many companies fail to protect confidential employee data. Retrieved April 5, 2001, from http://www.epic.org/privacy/workplace/linowesPR.html

Lohr, G. (2002). Learning the value of online training. *Washington Business Journal, 20*(38), 37.

London, M., & Smither, J. (1995). Can multisource feedback change perceptions of goal accomplishment, self-evaluations and performance-related outcomes?

Theory based applications and directions for research. *Personnel Psychology, 48,* 803–39.

Lukaszewski, K., & Stone, D. (2001). *The effects of the type of data and choice of using a Web-based or traditional HR system on reactions to human resources information systems.* Unpublished manuscript, State University of New York at New Paltz.

Mantyla, K. (2000). Learning portals: Online e-learning options. In K. Mantyla (Ed.), *The 2000/2001 ASTD distance learning yearbook* (pp. 15–18). New York: McGraw-Hill.

March, J., & Simon, H. (1958). *Organizations.* New York: Wiley.

Moe, M., & Blodget, H. (2000). Merrill Lynch e-learning: The knowledge web, part 4. Corporate e-learning–Feeding hungry minds. Retrieved December 12, 2001, from http://www.internettime.com/itimegroup/MOE4.PDF

Peterson, R., Marostica, M., & Callahan, L. (1999). *E-learning: Helping investors climb the e-learning curve.* US Bancorp Piper Jaffray Equity Research. Retrieved on May 7, 2002, from http://www.eoskl.it/paper/piper-elearning.pdf

Pillar, C. (1993). Privacy in peril. *MacWorld*, *10,* 124–30.

Prentice, E., III. (2001). Answer Geek: Can soft skills be taught with Web-based training, or should soft skills training take place in an instructor-led, classroom environment? *Learning Circuits.* Retrieved December 12, 2001, from http://www.learningcircuits.org / 2001/sep2001/geek2.html

PricewaterhouseCoopers (2002). *Trendsetter barometer.* Retrieved February 15, 2002, http://www.barometer-surveys.com/pr/tb020122.html

Rynes, S. (1991). Recruitment, job choice and post-hire consequences. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 399–444). Palo Alto, CA: Consulting Psychology Press.

Salas, E., & Cannon-Bowers, J. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52,* 471–501.

Salas, E., Kosarzycki, M., Burke, C., Fiore, S., & Stone, D. (2002). Emerging themes in distance learning research and practice: Some food for thought. *International Journal of Management Reviews, 4*(2), 1–19.

Schreiber, D., & Berge, Z. (1998). *Distance training: How innovative organizations are using technology to maximize learning and meet business objectives.* San Francisco: Jossey-Bass.

Society for Human Resources Management (2000). *2000 Workplace privacy survey.* Washington, DC: Author.

Spears, R., Postumes, T., Lea, M., & Wolbert, A. (2002). When are net effects gross products? The power of influence and the influence of power in computer-mediated communication. *Journal of Social Issues, 58,* 91–108.

Stone, D., Lukaszewski, K., & Stone-Romero, E. (2001, April). *Privacy and HRIS.* Paper presented at the meeting of the Society of Industrial and Organizational Psychology, San Diego, CA.

Stone, D., & Stone, E. (1987). The effects of missing application blank information on personnel selection decisions: Do privacy protection strategies bias the outcome? *Journal of Applied Psychology, 72,* 452–456.

Stone, D., & Stone-Romero, E. (1998). A multiple stakeholder model of privacy in organizations. In M. Schminke (Ed.), *Managerial ethics: Moral management of people and processes* (pp. 35–59). Mahwah, NJ: Erlbaum.

Stone, D., Stone-Romero, E., & Lukaszewski, K. (2003). The functional and dysfunctional consequences of using technology to achieve human resource system goals. In D. Stone (Ed.), *Research in human performance and cognitive engineering technology,* Vol. 3 (pp. 37–68). Amsterdam: JAI.

Towers Perrin. (2001). *HR survey.* Retrieved November 2001, from http://www.towers.com/towers/services-products/TowersPerrin/hrsurvey2001.htm

Ulrich, D. (2001). From e-business to e-HR. *International Human Resources Information Management Journal, 5,* 90–97.

Vaas, L. (2001). The e-training of America. *PC magazine, 20*(22), 1–5. Retrieved January 31, 2002, from http://ehostvgwl.epnet.com/fulltext.asp?resultSedId = R00000001&HitNum = 114&booleanT

I

Information Quality in Internet and E-business Environments

Larry P. English, *Information Impact International, Inc.*

INFORMATION QUALITY APPLIED TO THE INTERNET

As organizations move into the e-business arena, failure to address information quality issues can produce garbage-in–garbage-out at the speed of the Internet. Information quality is one of the most important ingredients to a successful Web site whether the site is informational or e-business transactional. Consider the following:

Twenty-nine million U.S. adults stopped using the Internet in 1999, double the number who dropped out the year before (Iwata, 2000).

A UK retail company received and confirmed orders for thousands of television sets for an online mispriced £3 (US$5.00) instead of the actual price of £300 (US $500.00) before discovering the error.

A European telecommunications company failed to charge their customers of Internet services an amount equivalent to 8 million Euros ($7.8 million) for six months before they discovered the "oversight," creating a significant public relations and recovery problem.

A U.S. airline offered trans-Atlantic flights for $25.00 fares over the Web, honoring them at a significant loss of revenue.

A U.S. computer maker offered personal computers online for $1.00, resulting in thousands of sales and legal problems.

One industrial supplier, whose typical orders were $1,000–$10,000, received an Internet order from a fictitious customer, named "Dr. Evil," for $10 million worth of materials. Although this was an obvious prank, intentional falsification is not uncommon on the Web.

A major airline sued one of its Internet business partners for "misleading" information about its flights.

This discussion applies to all aspects of information quality (IQ) on the Internet and e-business, including IQ in Web-provided information and business-to-consumer (B2C) and business-to-business or government (B2B, B2G, G2G) supply chain management.

Information Quality Defined

To define information quality in the context of the Internet, we must first define "quality":

> Quality is a customer determination, not an engineer's determination, not a marketing determination or a general management determination. It is based upon the customer's actual experience with the product or service, measured against his or her *requirements*—stated or unstated, conscious or merely sensed, technically operational or entirely subjective—and always representing a moving target in a competitive market. (Feigenbaum, 1991, p. 7)

Quality, then, is "the total set of product and service characteristics . . . through which the product and service in use will meet the expectations of the customer" (Feigenbaum, 1991).

Information quality is "consistently meeting all knowledge worker and end-customer expectations" (English, 1996, p. 1.4) in all characteristics of the information products and services they deem important. In e-business Web visitors, e-consumers, and e-business partners will assess information as "quality" or not, based on their needs and expectations as they conduct business on the Internet.

Components of Information Quality

Information consumers have quality expectations in three areas of information: *definition, content,* and *presentation* (English, 1999, p. 27–30). Internet and e-business information require the following:

Data Definition Quality

With data definition quality e-customers understand *the meaning* of the information and any underlying policies, such as terms and conditions of business transactions. They know exactly the meaning of information a Web transaction asks them to provide so they can provide the correct data values.

Data Content Quality

With data content quality e-customers *trust* the information they get to be *accurate, complete,* and *current,* and they *trust* that the organization will protect their information (security and privacy). They trust the organization will handle their information properly and will fulfill their transactions accurately, properly, and on a timely basis.

Information Presentation Quality

With information presentation quality e-customers can find the information they need quickly *when* they need it, and the information is *presented* in an *objective,* intuitive way, enabling them to take the right action or make the right decision.

Categories of Internet Information Requiring Quality

There are three categories of information in the Internet environment and the processes that produce and maintain them to which quality principles must be applied:

Web-Based Documents and Web Content

Misleading, inaccurate, or noncurrent information provided to Web visitors can cause harm to the enterprise or drive customers away without even being noticed.

Data "Shared" by Internal Processes and Internet Processes

Disparity in designs between Internet accessible databases and internal databases drives operating costs up and increases points where errors can be introduced.

Information Collected or Created over the Internet

Data created by customers or business partners in e-business are outside enterprise control and introduce new sources of nonquality that must be overcome.

E-business Supply-Chain Management

E-business is not just a new channel to be "bolted on" to existing processes. It demands rethinking and reengineering *how* the organization conducts work. The entire value chain of work from the Web activities to internal business activities, and from originating supply to ultimate distribution to the customer, must be managed and controlled for quality and customer satisfaction.

Quality Processes for Internet and E-business Information

To provide and collect quality information, effective enterprises will implement quality management processes and apply them to the processes that "produce" information in the same way that world-class manufacturing and service organizations applied quality management principles. These processes include the following:

Assess Information Quality

An organization must measure the extent of nonquality information and the costs of process failure and "information scrap and rework" to understand whether its information processes are in control and consistently meet customers' expectations.

Improve and Control Information Processes

Process improvement and control enable organizations to operate at peak performance to meet their customer's needs.

Implement an Information Quality Improvement Environment

Information quality does not just happen. To have sustainable information quality, organizations must embed quality management beliefs, principles, and processes into the fabric of their cultures.

QUALITY PROBLEMS UNIQUE TO THE INTERNET AND E-BUSINESS INFORMATION

While IQ is important in all of business, it is even more critical in e-business. In cyberspace, the information *is* the business, making IQ a *strategic* requirement for successful e-business. Here, organizations expose both their business processes and information to their customers and business partners. Nonquality information and defective processes creates high costs of process failure and information scrap and rework and indirect costs of missed opportunity and lost customer lifetime value.

All customer or supplier interactions and transactions take place electronically without people-to-people interaction. Quality problems occur in *gemba,* the Japanese word that means "the real place." In quality management, *gemba* is where value work happens. In manufacturing, *gemba* is where products are made or the place services are performed (Imai, 1997, p. 13). In the Information Age, Gemba is anywhere information is discovered, created or "produced," updated, moved, or exchanged. Information quality principles and techniques must be employed at each point. On the Internet, *gemba* is a virtual place, unlike the shop floor or at a sales counter where people can see when problems occur. Internet IQ problems include the following:

General Internet Information Quality Problems

Unscrupulous and unreliable Web sites have driven potential e-customers off the Internet. Anyone can create a Web site and place any kind of information on it with little concern to its accuracy. The "quality" Web site must establish its credibility for its e-customers to trust both the information and the organization that provides it.

The ease of posting information to the Web coupled with the emphasis on accelerated cycle times to publish on the Web has had the impact of decreasing quality control of the content (Calvert, 1999).

Nonquality information in cyberspace can go unnoticed—except by the visitors—for much longer periods of time than in a store, where staff can recognize and adjust for obvious IQ problems.

Internet information sources can be unreliable with pages being modified without documentation, or moved or removed without notice. According to DeletedDomains.com, there were 29,335,210 currently registered domains and 19,052,261 deleted domains as of May 2, 2003.

Misinformation, complaints, and rumors can be spread globally quickly and easily over the Internet, increasing risks and damage caused by nonquality Web information.

Business-to-Consumer (B2C) Information Quality Problems

On the Internet, customers who have a "complaint" are just a few clicks away from competitors. If they cannot find something easily or perceive noncurrent, missing, inaccurate, or biased information on the Web, they only have to click to the competition.

Staff no longer control information captured. E-customers are the new *information producers* in e-business. Organizations cannot "train" e-customers in the same way they train point-of-sales staff. Complex transactions produce higher error rates.

The anonymity of the Web increases the potential for misinformation, maliciously provided information, and fraud.

Privacy laws impose additional minimum "quality" requirements on data.

Mistrust of e-businesses has created a new IQ problem—deliberate falsification of data by e-customers to protect their privacy. An increasing number of Web visitors provide false information when registering with Web sites. According to a Georgia Tech Research Corporation study, 46.2% of 4563 respondents said they falsified information at least once when they registered on Web sites, with 12.3% saying they falsified information more than 50% of the time (GVU, 1998; the most recent study on this topic known to the author at time of publication). See Figure 1.

Business-to-Business (B2B) Information Quality Problems

The exchange of information among businesses or organizations (business, government, etc.) exposes risks of errors, omissions, and duplication. Overloaded data elements (multiple fact types) may be handled by the source organization's immediate operational application.

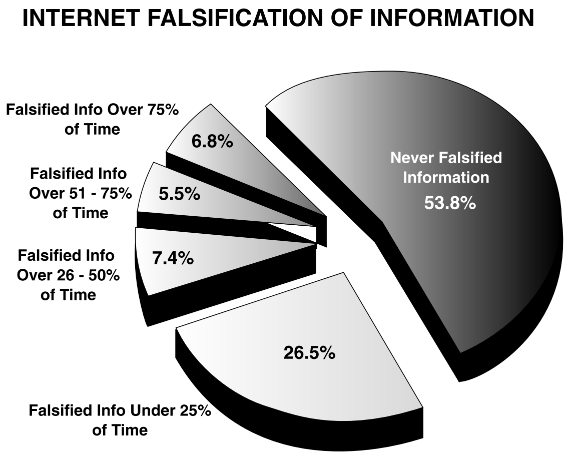


Figure 1: Web site registrants frequently falsify information.

However, that same data can cause business partner processes to fail.

Nonquality information passed can expose embarrassing information or breach regulation or confidentialities.

Subtle differences in the data definition, domain value sets, and business rules across organizations can cause interorganizational processes to fail. Effective communication requires partners to understand the precise definition of the data and the business rules constraining it.

XML by itself cannot solve data definition problems. XML, as a language for information exchange, addresses only part of the problem by passing a label (not a definition) of the values passed. However, XML by itself will no more solve communication problems across organizations than DBMS by itself solved the data-sharing problem within an organization. Cross-organization and even cross-industry data standards and common information models are required to eliminate communication (data definition) problems in B2B communications.

QUALITY PRINCIPLES APPLIED TO INTERNET AND E-BUSINESS INFORMATION

To solve the IQ problems inherent to the Internet, an organization must understand e-customers' quality expectations, the fact that information itself is a "product," and how to implement quality principles and processes into the organization culture.

The Internet "Customer"

Quality management has taught us that business success depends on how well an organization meets or exceeds customer expectations. E-business is the same. "Visitors" to a Web site are not unempowered "users." They are, first, information "customers" and knowledge workers who gather information for their purposes and, second, they are product/service "prospects" who hopefully become satisfied "customers."

Successful e-businesses will understand the quality expectations of its market and will measure "customer satisfaction" as part of its balanced scorecard. Increased customer satisfaction results in "customer lifetime value" of return visits and repeat business.

An effective e-business leverages its information content, especially information about its existing customers and their relationships. It analyzes its information to understand its customer behavior and needs, and uses it to plan, design, and implement its Web site and e-business processes.

The Internet "Product"

On the Internet, the information *is* the "business." Whether an organization is providing information or selling products or services, the only commodity that exists is the information. There are no tangible goods for the customer to feel, try on, or try out. Because e-customers are not able to *physically examine* products, they must *trust* the information.

Only after e-customer expectations are understood can the organization design (or enhance) a Web site to provide "quality" information that "consistently meets information customer expectations." Only then can it design the e-business transactions that enable the customers to conduct business in ways that satisfy their requirements.

The E-business "Processes"

Three categories of processes and information require quality management:

Providing information, including Web site design, Web content, and information produced by internal business processes;

Collecting information from e-customers or business partners about themselves; and

Conducting business over the Internet, including B2C and B2B transactions.

Quality principles applied to each of these processes are described in subsequent sections.

THE CUSTOMER-CENTRIC E-BUSINESS VALUE CHAIN

A customer-centric e-business value chain, illustrated in Figure 2, identifies the cycle of activities of an e-customer's request for information or services and their satisfaction from benefits received. Successful e-businesses must understand this cycle and manage and control the activities across the value chain, not just the individual functions to meet e-customer expectations.

Customer Focus

World-class e-businesses, like world-class brick-and-mortar businesses, focus on customer satisfaction for success. The first of Deming's 14 Points of Quality confirms that an organization must "create constancy of purpose

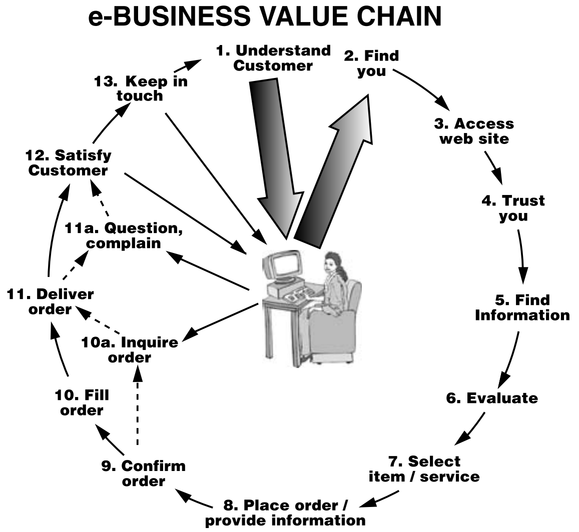


Figure 2: The e-business value chain. From English (2000). © 2000. Reprinted with permission.

Table 1 Customer Satisfaction: Why E-customers Return

Quality Requirement	E-customer Expectation
Valuable content	Meets my need to know and my purposes.
Accurate	I do not waste time or make costly mistakes as a result.
Complete	I have all the data I need for my purposes.
Current	Updated frequently; new things of interest to me.
Objective	I can trust the information because it is not misleading; I easily draw the right conclusion.
Concise	I do not have to waste time in "information overload."
Fast	Fast downloads and few clicks minimizes my time to use.
Easy to use	Minimizes my time to learn; I don't get lost.
Timely	Conduct business on *my* time table.
Reliable	I get goods and services that meet my expectations.
Meets expectations	Makes my life easier or more successful.

Note. From English (2000). © 2000. Reprinted with permission.

for improvement of product and service," an obligation to the customer that never ceases, for "the consumer is the most important part of the production line" (Deming, 1986, p. 26). A Web site, whether informational only or B2B or B2C, will succeed or fail based on the e-customers' experiences. Table 1 illustrates why e-customers return to a Web site.

E-business Value Chain Steps

Success in e-business requires an organization to understand the complete value chain of e-business from the request of a prospective e-customer to a satisfied e-customer. The 13 discrete value chain steps are illustrated in Figure 2.

This section introduces the quality requirements required for each value chain step. Specific quality techniques are described in the subsequent sections:

Step 1: Understand Customer Expectations

The first step of the e-business value chain is central to quality management: Understand and define customer expectations in order to design quality into information, products, and services that meet or exceed them.

Steps 2–6: Provide Quality Information

A Web site must first be found *easily* by prospective e-customers. Both general information and product/service information must be found *easily* and enable e-customers to evaluate the right products and information for them. The site *must contain* information that e-customers find valuable and engender *trust* for them to conduct e-business.

Steps 7–9: Provide Quality Products and Services

As e-customers conduct business over the Internet, they become *information producers,* directly providing personal information and transactional data. Error-proofing techniques must be applied to *information collected from and provided by* the e-customer to prevent inadvertent errors, such as transpositions, unreasonable values, or deliberate falsification.

Steps 10–12: Meet/Exceed Customer Expectations

These steps fulfill the e-business requests for information, or orders for products or services, or requests to conduct business transactions, whether online, such as software upgrade download or bank account transfer, or physical product delivery, such as office supplies or custom-made products. The e-business must meet delivery and quality expectations, communicating to e-customers any delivery problems, handling questions, and resolving any complaints. E-customers judge e-business' quality with the simple criteria: "I got what I expected in the way I expected it with value for my time and money."

Steps 13 and 1: Continually Improve

Customer-focused organizations will understand their customers better by analyzing all customer contacts. Through an *appropriate* level of contact with e-customers and analysis of customer experiences, customer loyalty and customer lifetime value can be increased.

E-BUSINESS VALUE SYSTEM AND CODE OF ETHICS

With myriads of unscrupulous and unreliable Web sites along with growing incidents of identity theft, reputable e-businesses must establish credibility and trustworthiness in their Web sites and e-business operations (Alexander & Tate, 1999; Ciolek & Goltz, 1996; English, 2000).

The quality organization will differentiate its Web site as a reputable place to find reliable information and successfully transact business. An organization with a customer-centered value system can translate that into a customer-centered code of ethics that instills e-customer confidence. A code of ethics contains statements of commitment to the following:

Integrity, Credibility, and Legality

This establishes professional credibility or authority (Alexander & Tate, 1999, p. 11) and a commitment to compliance to an industry or professional group code of ethics that the e-customer can verify. For example, Rite Aid and drugstore.com display their VIPPS certification provided by the National Association of Boards of Pharmacy. To be VIPPS certified, a pharmacy must comply with NABP criteria, including quality assurance, patient privacy rights, state licensing and inspection, security of prescription orders, and provision of meaningful consultation between patients and pharmacists (NABP, 1999).

Security

The highest level of security must be provided relative to the nature of the information exchanged to assure no accidental disclosure of or unauthorized access to personal data or confidential or sensitive business data. Data must be classified as to its security requirements, including regulatory and privacy, and implemented with the proper level of security control, such as firewalls, secure servers, effective log-in ID and password techniques, and encryption.

Privacy and Confidentiality

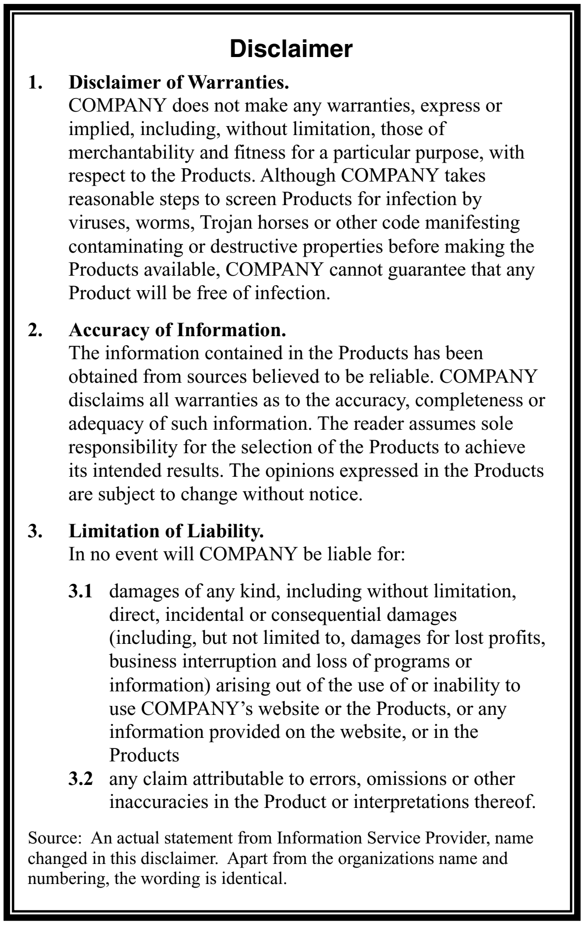
Consumers are increasingly frustrated with what they perceive as "intrusive" invasion of their privacy. The UK and EU Data Protection Acts and the U.S. Gramm–Leach–Bliley Act require organizations to take a more customer-focused stance on privacy. These laws confirm customer "rights" as to what data may be collected about them and how it may be used.

Conforming to the letter of the law is only the *entry point* for privacy. A quality organization understands its individual customer's privacy expectations and honors them. Reputable organizations develop customer-centric privacy policies and have third-party certification of their privacy statements through organizations such as TRUSTe (1997) or BBBOnline (1996).

Information Quality

Credible Web sites will *not* just post a *disclaimer* disavowing responsibility for the completeness, accuracy, or currency of the information they post. Rather, they will *claim* what they *are* doing to assure quality information. See the upcoming section Quality Principles for Web-Based Documents, Web Content, and Information Presentation. Only after they communicate what they are doing to provide quality information will they describe any disclaimer for legal purposes.

Figure 3 illustrates an actual disclaimer on an information provider's Web site. There is no statement as to what, if any, quality management controls are applied.



Disclaimer

1. **Disclaimer of Warranties.**
   COMPANY does not make any warranties, express or implied, including, without limitation, those of merchantability and fitness for a particular purpose, with respect to the Products. Although COMPANY takes reasonable steps to screen Products for infection by viruses, worms, Trojan horses or other code manifesting contaminating or destructive properties before making the Products available, COMPANY cannot guarantee that any Product will be free of infection.
2. **Accuracy of Information.**
   The information contained in the Products has been obtained from sources believed to be reliable. COMPANY disclaims all warranties as to the accuracy, completeness or adequacy of such information. The reader assumes sole responsibility for the selection of the Products to achieve its intended results. The opinions expressed in the Products are subject to change without notice.
3. **Limitation of Liability.**
   In no event will COMPANY be liable for:

   3.1 damages of any kind, including without limitation, direct, incidental or consequential damages (including, but not limited to, damages for lost profits, business interruption and loss of programs or information) arising out of the use of or inability to use COMPANY's website or the Products, or any information provided on the website, or in the Products

   3.2 any claim attributable to errors, omissions or other inaccuracies in the Product or interpretations thereof.

Source: An actual statement from Information Service Provider, name changed in this disclaimer. Apart from the organizations name and numbering, the wording is identical.

Figure 3: Example of an Information Quality Disclaimer with a nonquality focus.

Describing what the organization is doing to *assure quality* information increases customers' trust in their information. U.S. federal law requires federal agencies to ensure and maximize the quality, objectivity, utility, and integrity of the information (including statistical information) they disseminate. "Agencies must issue their own implementing guidelines that include 'administrative mechanisms allowing affected persons to seek and obtain correction of information maintained and disseminated by the agency' that does not comply with the OMB guidelines" (OMB, 2002). The National Science Foundation Web site illustrates its "warranty" and procedures for how to communicate an information complaint and how the complaint will be handled (NSF, 2002). Information consumers can identify suspected nonquality information and describe how the information consumer uses the information, how they are "hurt" by the information in error, and reasons for correcting the suspected error.

This law may be a forerunner of laws requiring *warranties for information quality* and customer recourse in the private sector if consumer well being becomes at risk.

Customer Satisfaction

Credible e-businesses post customer satisfaction policies and product/service warranties. It is even more important when conducting business without prior physical evaluation of goods and services. See Table 1 above.

QUALITY MANAGEMENT PROCESSES FOR E-BUSINESS

Like manufacturing quality, information quality must be planned with assessment and improvement processes defined. Quality must be designed into the business processes for quality information products. Because information is a product, the same principles that apply to manufacturing quality apply to information quality (English, 1999; Huang, Lee, & Wang, 1999; Redman, 2001).

Information Quality Assessment

This process measures the quality of data definition, data content, and information presentation to assure information meets customers' expectations. Assessments measure the level of quality in important characteristics, such as accuracy, completeness, and currency of the information. Measuring the costs of nonquality information establishes the economic impact of nonquality information and prioritization of information quality improvement and error-proofing initiatives (English, 1999, pp. 199–236).

Information Process Improvement and Control

Without process definition, improvement, and control, processes will produce a(n)—often high—degree of error. A process improvement technique, such as the Shewhart Cycle of PDCA or Plan–Do–Check–Act (Deming, 1986; English [PDCA applied to information processes], 1999; Imai, 1997) or the Six Sigma DMAIC or Define–Measure–Analyze–Improve–Control (Brue, 2002), is used to apply various quality techniques to improve processes. The goal is to error-proof processes and prevent *errors* or *omissions* in information created by the enterprise for its Web site content or information collected from the e-customers. Processes must be in place to maintain *currency* of information subject to information quality decay.

Error-proofing techniques must also be applied to B2B and B2C processes that collect information from e-customers, whether consumer or business-to-business, to assure its *completeness, validity* to allowed values *and* correctly defined business rules, *accuracy,* and *nonduplication* (multiple records about the same customer). Accuracy cannot be "guaranteed" with edit and validation routines electronically for most data elements, such as birth date, name (spelling), address-to-person matching, and product price. Accuracy requires additional error-proofing techniques to prevent digit or character transpositions, deliberate falsification, and reasonable but inaccurate data entry.

Planning and Implementing an Information Quality Improvement Environment

Producing quality products or services is not a one-time project. It requires a culture of process excellence and customer focus in tangible goods and information products. This requires management accountability for information quality, training in information quality principles and methods, and implementation of key performance measures of customer satisfaction of information products.

Planning for quality requires management to understand the costs of nonquality information. Costs include *direct* costs of process failure and "information scrap and rework" caused by nonquality information as well as *opportunity* costs of lost and missed revenue caused when missing or inaccurate data cause the organization to miss an opportunity or drive customers away. Especially important in e-business, e-customers who cannot find easily what they are looking for, find suspect data, or cannot trust the organization can, and will, go elsewhere. E-customer complaints can go unnoticed and, therefore, unresolved, while the e-customer clicks off to the competition.

Deming's 14 Points of Quality describe the 14 components of management transformation to create a quality culture. The 14 Points must be implemented to sustain a habit of continuous improvement to "do jobs right the first time" and eliminate the high costs of process failure and scrap and rework, whether in manufactured products or information (Deming, 1986, pp. 24–96; English, 1999, pp. 337–399; English, 2000, pp. 6.10–25):

1. Create constancy of purpose for *improvement* of *information* product and service. The obligation to the customer never ceases. Planning, designing, and operating the e-business for e-customer success leads to e-business success.
2. Adopt the new philosophy of quality information, for "reliable [quality] shared information reduces costs." Improving processes to error-proof them is less costly than the direct costs (alone) of process failure and information scrap and rework due to defective data.
3. Cease reliance on mass data inspection as a means of quality: design quality into the processes that create, update, distribute, and deliver information to knowledge workers. This applies to internal processes creating Web content and to processes that capture data from Web visitors.
4. End the practice of "on-time" and "within-budget" alone as measures of information systems' delivery of success. The collapse of the Internet revolution taught us that it is not who is first into the Internet space who will win, but those with *quality* information, products, and services that satisfy e-customers. Of the three project objectives "quality, cost, and speed," sacrificing quality always drives up costs of maintenance and scrap and rework, and prevents optimizing time-to-market in the long term.
5. Improve constantly and forever the system of information production and service. Information quality is not a one-time data "cleanup" initiative. World-class organizations make a habit of continuous improvement.
6. Institute training. Content and information producers within the organization require training to know how to provide quality information and how to improve their processes. They must know how their customers use information and the costs and consequences of nonquality. "Training" of e-customers requires intuitive site and business transaction design, providing online training and help mechanisms for ease of use, including intuitive clear labels, glossaries and definitions, search engines, easy site navigation, and help where necessary.
7. Institute leadership for information quality. IQ does not just happen. Someone must lead the enterprise into implementing the principles, policies, and

processes of IQ management. Management must accept accountability for the information produced by its staff. Managers of the Internet processes must be held accountable for the quality of information captured. Managers must implement processes that prevent defective data.

8. Drive out fear. E-customers will NOT conduct business with an e-business if they "fear" for their privacy, or they will be motivated to falsify information.
9. Break down barriers between staff areas, including between information systems areas and business areas, between application and e-business development staff and information resource management, and between business areas and business areas. Enterprise failure occurs when business areas operate to their own autonomous goals. When they do so, they create problems for their downstream information "customers," forcing them to spend time in scrap and rework instead of value work. World-class organizations manage *across* the e-business and other value chains.
10. Eliminate *slogans* and exhortations for quality, and replace them with *actions* for IQ improvement. An organization cannot just *tell* people to do a quality job, it must provide people with the training, resources, and tools, such as a process improvement method, to *do* a quality job.
11. Eliminate quotas of "productivity" and replace them with metrics of customer satisfaction that keeps customers coming back, increasing customer lifetime value and profit.
12. Remove barriers of pride of workmanship. Empower and involve information producers to improve their processes. They know the barriers, and given an opportunity will improve processes to eliminate waste, for people want to be able to say, "we have provided valuable work."
13. Institute a vigorous program of education and self-improvement. The organization that learns tomorrow's skills will be the organization that leads tomorrow's markets. In the Information Age, information is the new currency. We must learn that information management is a core competency, manage information and improve the processes that produce it with quality, and exploit it for the benefit of its customers and its internal and external stakeholders.
14. Management must organize itself to make the other 13 Points happen. To do so, senior management must understand—and feel the pain—of the costs of information scrap and rework. Senior management must communicate to a critical mass of people why change is necessary. Management must put *everybody* to work to make information quality happen as a proactive reality, and not as a reactive, scrap-and-rework consequence of not producing information systems, databases, and data right the first time.

A methodology for Total Information Quality Management (TIQM) is represented in Figure 4, and documented in English (1999). Other quality methods describe information quality assessment (Huang et al., 1999) and data quality tracking and control (Redman, 2001).

TOTAL INFORMATION QUALITY MANAGEMENT (TIQM®)

TIQM® is *not* a program; it is a *value system, mind set,* and *habit* of continuous improvement of:

1. Application and data development processes

2. Business processes

By integrating *quality* management *beliefs, principles* and *methods* into the culture

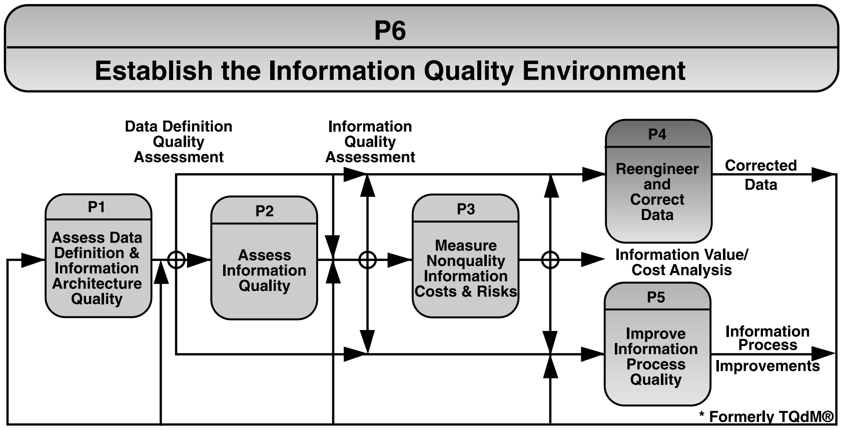

P6
Establish the Information Quality Environment
Data Definition Quality Assessment
Information Quality Assessment
P1
Assess Data Definition & Information Architecture Quality
P2
Assess Information Quality
P3
Measure Nonquality Information Costs & Risks
P4
Reengineer and Correct Data
Corrected Data
Information Value/ Cost Analysis
P5
Improve Information Process Quality
Information Process Improvements
* Formerly TQdM®

Figure 4: Methodology for total information quality management. From English (1999). Reprinted with permission.

QUALITY PRINCIPLES FOR WEB-BASED DOCUMENTS, WEB CONTENT, AND INFORMATION PRESENTATION

Quality management principles must be applied to information, both Web pages and database data, made available to e-customers on the Web site and disseminated electronically. These principles address e-business value chain steps 1–6 (Figure 2).

Quality Principles for Web Design and Information Presentation Quality

This section seeks not to describe Web site design, but to describe the *quality principles* for "designing quality into" Web site and document information presentation. Quality principles include the following:

Quality Design Principles

Quality Function Deployment (QFD) techniques are used to involve e-customers in the *design* (*redesign*) of the Web site and presentation style as a "product." QFD techniques link the needs of the customer with design, development, engineering, manufacturing, and service functions. QFD helps organizations discover the spoken needs ("Voice of the Customer") and unspoken needs, translate them into actions and designs, and focus the various business areas developing products and services that exceed normal customer expectations (QFD Institute, 1995).

Web sites should be designed for simplicity and ease of navigation to prevent casual e-customers from getting lost (Alexander & Tate, 1999; Nielsen, 2000; Useit, 1996). As many as two out of three customer complaints are caused by confusing marketing and inappropriate customer use (nonquality information presentation) (TARP, 2000).

Site search capability allows e-customers to quickly find specific information.

Document design techniques like Information Mapping help chunk information and documents into structured formats that make it easy for readers to scan and find pertinent information quickly (Infomap, 1995).

Pictures or graphic representations should clearly and accurately represent products. Allow pictures with enlargement options for increased clarity to minimize returns.

A "minimalist" approach to text information, using standards for meaningful titles and descriptive abstracts, enables e-customers to evaluate content efficiently and reduces downloading of inappropriate documents.

Provide e-customer-friendly error messages. Intercept cryptic system error messages and restate them to be meaningful.

Design and test for the most common denominator for the browsers and intended access platforms, such as laptops, PDAs, and e-mobile devices.

Assure all hyperlinks correctly link each time a page is updated. Periodically check hyperlinks to assure currency. If linked pages are moved, links should be updated. If 6,000 e-customers link to a moved page with a 10-s timer redirecting it, 10 hours of e-customer time is wasted.

Quality Feedback Principles

E-customers need an easy way to provide feedback or "complain" about nonquality site information, products, or services. Feedback is critical to detect broken processes early, as well as to prevent losing customers.

Clickstream quality analysis and customer satisfaction surveys (TARP, 2000) help understand e-customers' experiences in the *gemba* of The Internet, especially to identify prospective customers who drop out of a site after starting a business transaction. The pages people drop out from may identify causes of e-customer dissatisfaction.

Quality Principles for Web Data Definition

Because e-customers have different backgrounds and range from casual to experienced, labels and terms used must be intuitive and unambiguously named and defined with help mechanisms as necessary for ease of understanding:

Use universally accepted names and terms.

Provide glossary with hyperlinks to technical terms, or terms that might be misunderstood. If multiple definitions may apply to a term, define all pertinent meanings and the context in which each definition applies (English, 1999, p. 97).

Develop guidelines for descriptive key words and metatags to facilitate indexing and search engine management.

Quality Principles for Electronic and Web Document Management

Information management includes all forms and formats of information, not just electronic databases or files. This is even more important when documents, such as product literature, may be produced in multiple physical formats, such as glossy brochures, plain paper copies, and HTML or Web documents and presentation media format.

A document management process, such as illustrated in Figure 5, provides consistent control of document and distribution (ISO 9001, Standard 4.1).

Assure updates are made to a single record-of-reference document with revisions or revised document published or propagated to secondary copies.

Maintain document "expiration dates" or "review-by dates" as well as revision numbers and dates, requiring review and approval. This enables proactive management of content subject to "information quality decay" to maintain currency. These documents can be automatically deleted if not updated, or marked to advise readers they are no longer current.

Quality Principles for Web Content Creation and Maintenance

Because on the Internet, *the information is the business,* information created for the Web must be subject to rigorous quality management processes. Standards for Web content include criteria of authority (credibility), accuracy, objectivity, currency, and adequate coverage to meet

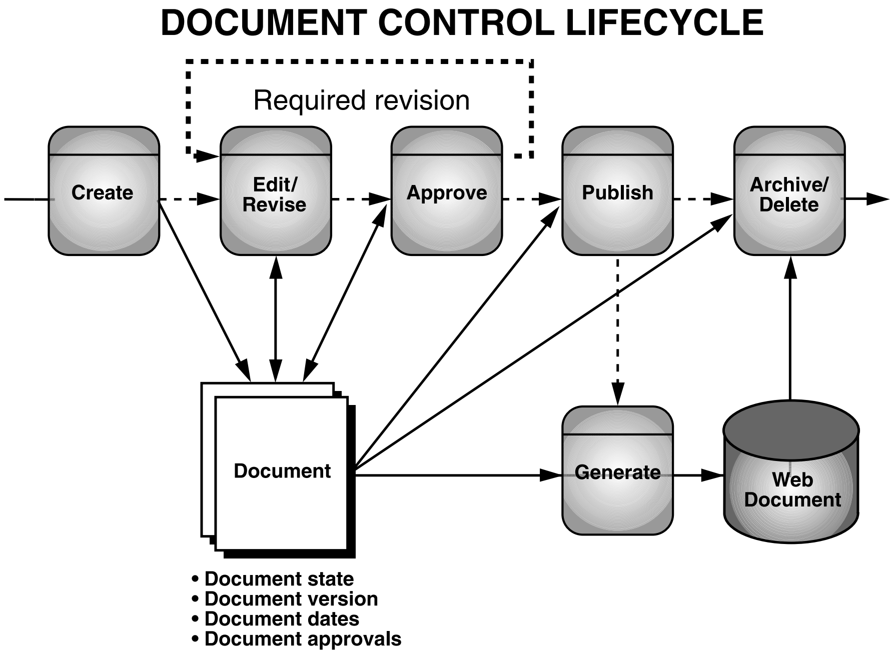


Figure 5: Document control lifecycle. From English (2000). © 2000. Reprinted with permission.

the needs of the intended audience (Alexander & Tate, 1996, 1999; Beck, 1997; Ciolek & Goltz, 1996; English, 2000; Virtual Chase, 1998).

Implement accountability and training:

- Implement management accountability for information content. Documents and information created within a business area are the responsibility of that business area manager. In the same way managers are accountable for their budgets and personnel, they *are* also accountable for the information produced in their area, as managerial information stewards (English, 1999, p. 406).
- Provide training to managers and information and Web content producers in quality principles and defect prevention techniques.
- Develop contracts for quality standards with third-party information sources.

Implement standards for site and page design that establishes the authority and credibility of the information (Alexander & Tate, 1999):

- Clear identification of organization responsible for the content, parent or sponsoring organization, and editorial or oversight board, if any.
- Name of copyright holder.
- Method to contact organization.
- Content author name and qualifications.
- Source(s) of information not produced by the host organization and how obtained.

Implement standardized processes which help content producers to provide quality information:

- Confirm accuracy of details with reliable sources.
- Corroborate information provided by external sources to assure accuracy.
- Verify summary or calculated data.
- Assure consistency of redundant data, such as to source data, between data in body and abstract or introduction, or in tables or figures.
- Conduct periodic IQ assessments of information content from both internal and external sources and process improvement initiatives as necessary.

Implement standardized procedures for all content producers to keep data current:

- Establish review or expiration dates for pages containing time-sensitive data.
- Provide dates when data were first provided to the site and date last updated.
- Use international date format (dd month yyyy) to avoid ambiguity; for example, 11/12/02 could be 11 December 2002 or 12 November 2002 or either date in 1902.
- List frequency of update for volatile information, such as prices.

Implement reasonability tests in data created for or migrated to the Web. This would have prevented the magnitude of error in the UK retail company's £3 television sets cited earlier.

- Assure information coverage is complete for the needs of the intended audience.
- Use the "who, what, where, how, when, why" checklist to assure all relevant data are included
- Include links to supplemental or support data.
- Provide definitions or links to glossaries of technical terms for understanding.
- If the content also exists in print format, identify any inconsistencies, such as if parts are missing, or if some content is updated from the print version.
- Provide legends, footnotes, or explanations for anomalies. If e-customers doubt data that appear "wrong" but are actually correct, they may lose trust in the entire Web site and not come back.

Present information objectively:

- Avoid bias that may lead the reader to a wrong decision. Although marketing information is inherently biased, there is a fine line between acceptable marketing bravado and deception. When e-customers feel misled, they can become vociferous complainers through chat rooms, consumer action Web sites, and e-mail.
- Differentiate informational, advertising, and editorial pages.
- Present both sides of issues fairly, even if the recommendation is for one point.
- If providing a survey or sampling to create summary information, assure the use of sound statistical sampling techniques and communicate the survey or sampling techniques used and the confidence level of the results. Present the questions as asked, especially if presenting a short form of the question in graphics.
- Publish IQ assessment results with assessment dates, so knowledge workers know the degree of error and omission, if any. Knowledge workers may not need zero-defect data, but they need to know the degree of inaccurate and missing data, to factor into their decisions or use.

QUALITY PRINCIPLES FOR DATA SHARED BY INTERNAL AND INTERNET PROCESSES

Data integration of Internet and internal data is best achieved by accessing data from a commonly defined database. If, for security or real performance reasons, the Internet databases must be physically separate from the internal databases, they must have *controlled* processes that keep the data, such as product and product price or customer name and address, equivalent.

Information Architecture and Internet Applications

Databases that house an enterprise's data must be designed, just as a building must have a sound blueprint. For shared business data maintained in an enterprise database such as product and product price or customer name and address, the quality principle is to design and define the data so they can be shared directly from a single database or be controlled replicated copies or partitions. Data designed for direct sharing eliminates the need for disparately defined and redundant databases for different functional uses. In this way only a single copy, or controlled replicated copies, needs to be maintained, reducing the costs of maintenance and data transformation.

Database architecture has quality design, resulting in databases that are

Stable—new applications, such as Web applications, are added with only the addition of new date elements or files and no destructive change is required,

Flexible—processes and applications can be reengineered with minimal modification to the database design, and

Reusable—minimal redundant databases are required, and where they are, they have consistent definition and domain value sets for common data elements, reducing development and maintenance costs and points of potential error introduction.

Develop standards and guidelines for the use of XML consistent with enterprise information architecture. XML as the basis for information exchange among businesses is often misunderstood and misused. It is a tool that *can be* used to help IQ if it is used effectively. Although XML allows passing data along with a label of the data, it is generally insufficient to completely define the meaning of that data. XML cannot be used as a substitute for complete definition of data elements and entity types, nor of rigorous data modeling. Standard XML labels should be consistent with the standard data names for data elements, and they must be backed up with complete, correct, and clear definition. Clear and precise business rule documentation should be available to business partners.

Managing Consistency of Redundant Data

When redundant databases must be developed, strict quality controls on data creation, updates, and data movement must be implemented. The quality principles for controlling redundant, disparately defined data in application software packages or in non-architected legacy systems include the following:

Define data with a standard name, single meaning, and valid data value set enterprise-wide. In this way data values need only to be replicated and not transformed. See below the principle for controlling data transformation when it is required.

Develop single-application programs or callable modules that maintain information about a specific entity type, such as product. Where there are subtypes of a given entity type that have unique attributes (data elements) not common to the supertype, then develop a unique callable module to capture unique attributes for each subtype. For example, all insurance products, regardless of whether they are life insurance or homeowners insurance, will have common attributes, such as product id, product name, product description, product first available date, product retired date, and product type (life or homeowners). Life products will have some unique attributes not applicable to homeowners and vice versa. Each subtype of product should have an application module dedicated to capture its unique attributes. This is the principle of single source supplier (Deming, 1986, pp. 35–38; English, 1999, p. 356). Having only one application module to capture common product information, rather than one application for each business area (life, homeowners, etc.), simplifies common data capture and reduces the costs of data capture and maintenance.

Define a single authoritative record-of-origin for each information type, or set of occurrences within one information type (partitioned database) to which data of that type will be created or updated.

Propagate data in one direction, from the record-of-origin to the record-of-reference database from which the data can be published or replicated where distribution is required.

When application software packages are *required* and/or when non-architected legacy systems must be interfaced with enterprise standard databases, the enterprise data model should control the data transformation and interfacing as a hub for information exchange. Controlling data transformation in data movement requires defining and implementing appropriate audit and control routines to assure proper extract, handling, and loading. Data transformation requiring control includes attributes with different data definitions (Web visitor to customer), different formats (person name [free form] to person first name, middle name, first surname, last surname), and/or different data value sets (M [male] and F [female] = 1 and 2, respectively). Most extract, correct, transform, and load (ECTL) software handles record counts satisfactorily, but few handle control total balancing and value transformation control balancing well. Content control audits generally must be defined and built as separate controls embedded in the ECTL job streams. Data control, tagging, or tracking techniques (Redman, 2001; Segev, 1996) assure data are moved and transformed properly across redundant databases.

Periodically assess information quality characteristics of timeliness (of data propagation) and consistency of data that should be equivalent in the multiple databases. Assess information quality by extracting sample records in the record-of-origin database. Assess them for completeness, accuracy, and nonduplication. Then extract the counterpart records in the downstream databases and measure for "equivalent" values, (gender code value of "M" in one database is a value of "1" in the downstream database) to assure that no corruption has taken place in the data movement (English, 1999; Redman, 2001).

QUALITY PRINCIPLES FOR INTERNET-COLLECTED INFORMATION

As Web sites matured from static read-only sites to registering e-customers and to conducting business, new information quality problems emerged. Capturing customer or business partner data and their business transaction data shifted from internally trained staff to the e-customers and e-business partners themselves. External customers and suppliers became direct information producers. Information quality became subjected to new factors:

E-customer lack of "training";

E-customer typing ability;

E-application design intuitiveness and ease of use;

Lack of strong e-application edit and validation capabilities;

Lack of e-application help capabilities; and

E-customer trust and deliberate falsification.

This set of IQ principles centers on the e-business value chain steps 7–9. Processes must have quality designed in to error-proof them such as check digits for numeric identifiers, double entry of critical data, and real-time look up of address and other verifiable data, to prevent inadvertent keying transpositions, to prevent deliberate falsification, and to verify the data using IQ software or internal edit and validation.

Understanding E-customer Expectations Regarding Their Personal Information

E-customers will provide quality information if they trust the organization and understand how their information will be used or why it is important.

Customers are increasingly concerned about their privacy and have fears about how their information might be used.

A customer-focused privacy policy helps customers trust an organization, minimizing deliberate falsification.

Cookies should be used from a customer-care focus to customize screens, offerings, and prevent re-entry of data already provided or known.

Defer to an e-customer's preferred form of (valid) address should that be different from the official postal address.

Allow e-customers to omit providing optional information not required for the processes. Create a "customer does not wish to provide" attribute along with their privacy requests.

Designing Quality into Electronic Forms

To minimize errors, forms must be intuitive, easy to understand and use, and error-proof.

Form labels and data names should be intuitive and clear from the reader's perspective with an accessible glossary of terms, to prevent misinterpretation by the readers and to prevent readers from providing the wrong information.

Information collection forms and procedures should be simple and clear. If procedures are required, use a well-defined set of procedure guidelines with prompts. Procedures should be written in a style readable by two grade levels below the target audience education level. Indices such as the Gunning Fog Index (1999) can assist in readability analysis. The goal is to eliminate or minimize errors due to misinterpretation or complex procedures.

E-customers should not have to re-enter data they have already provided. Propagate e-customer previously provided data into the forms, with opportunity for them to confirm or update inaccurate or changed information.

Pull-down menus and selection boxes minimize e-customer keystrokes in providing data and standardize data values.

Pull-down menus and selection boxes should have all possible options, with an "other" category to capture rare or unexpected, but real, values.

Designing Quality into Internet Processes That Collect Information

Design quality by defining the process, defining the data to be produced, and implementing error-proofing techniques. Techniques include the following (English, 1996, 8.35–36; English, 2000, 5.8; Redman, 2001):

Implement IQ software with "defect prevention capability" or edits in the source transactions of e-customer data creation. Several IQ software products provide defect prevention capability for e-business applications (Infoimpact, 1996, particularly the IQ Products page: http://www.infoimpact.com under the Tools and Resources Section, and IQ Products. Understand the strengths and limitations of IQ software tool functionalities (English, 1999, chap. 10).

Implement appropriate edit and business rule validation routines in data capture applications. Use intelligent edits. For example, avoid simply testing for an "@" to test for a "valid" e-mail address.

Defect prevention software and edit and validation tests should be performed in real time. If not possible, execute tests "near real time," before the data are acted on by subsequent processes to minimize process failure from defective data.

Do not overvalidate data. Correct values may fall outside expected limits. Allow e-customers to provide correct data and provide a reason. Then re-evaluate the validation rule tests.

Maintain assessment attributes to allow anomalistic data that are correct to be maintained without unintentional "correction" to a wrong value. For example, both George D. and Sam M. are women whose gender codes tend to get changed to "male" by "gender-assignment" software. Assessment attributes include the assessment type ("AV" = accuracy verified, or "CA" = corrected to authoritative source) and assessment date.

Periodically assess information quality characteristics of completeness, validity, accuracy, and nonduplication (English, 1999; Huang et al., 1999; Redman, 2001).

Provide feedback of acceptance of transactions to prevent the resubmissions that create redundant transactions.

Notify e-customers of data "corrections" for their confirmation. Even reputable defect prevention software can have errors in their databases or in their business rules.

Provide data "corrections" or suspect data back to the e-customer, such as address, product characteristics (such as an illegal color or size), or quantity for confirmation or correction.

IQ problems should be prevented by conducting an information quality improvement initiative, using the Shewhart PDCA Cycle (Deming, 1986; English, 1999; Imai, 1997) or the Six Sigma DMAIC technique (Brue, 2002). In PDCA, the "plan" phase seeks to discover the root cause(s) and define an improvement that eliminates recurrence of the defects. The "do" phase implements the improved process in a controlled way, enabling a "check" to confirm its effectiveness. If the improvement has accomplished the quality goal with no negative side effects, "act" standardizes and institutionalizes that improvement, putting the process into control.

QUALITY PRINCIPLES IN E-BUSINESS SUPPLY-CHAIN MANAGEMENT

Once e-customers have visited a Web site, liked what they have seen (e-business value chain steps 2–6), and placed their order or conducted their business (steps 7 and 8), the e-business must deliver the goods and services or complete the business transactions to meet expectations (e-business value chain steps 9–12). See Figure 2.

Early e-business failures often resulted from a rush to get into the Internet revolution with "bolt-on" processes to conduct business on the Web. Lack of thought-out process design and control can result in nonquality service delivery. E-business requires the enterprise to reengineer how it performs business.

Reengineer Business Processes

The Internet is not just a new channel for organizations to conduct business. It represents an opportunity to "rethink" *how* it can perform its business. Quality principles include the following:

Reengineer business processes; do not just "bolt-on" e-business processes.

Minimize developing applications that create redundant and disparately defined databases that must be integrated by "interfaceation." Define and develop integrated, commonly defined, and shared operational databases.

Eliminate activities that add only costs and focus on processes that add value. The more interface programs an organization has, the more opportunity for error in data movement and the more the costs of "scrap and rework" of maintaining the redundant databases and applications—and the lower the profit margin.

Develop contracts with service/product providers that specify quality requirements to satisfy your e-customer expectations.

Deliver as Promised

Whether delivering products from one's own or third-party warehouses to fulfill order distribution, e-customers will hold the e-business accountable for any failure. The entire order fulfillment value chain must be managed, not just the online portion.

Confirm the e-customer order. The communication loop is not completed until the sender is assured the receiver has understood accurately their request. Information quality means providing feedback to confirm correct understanding of their order, business transaction, or request. Feedback should include specific dates or times e-customer should receive their order or have the business transaction completed.

Enable e-customers to have personal contact with help staff when they need it.

Reply to e-customer inquiries and complaints in a timely fashion to minimize lost customers.

If there are hitches in delivery, proactively and quickly communicate the delay and rescheduled delivery, confirming whether this is acceptable to the e-customer.

Maintain a customer-friendly returns policy. Analyze return reasons to identify quality problems, such as the item "did not look like the picture."

UNDERSTANDING E-CUSTOMERS' EVOLVING QUALITY REQUIREMENTS

Because there is no *physical* interaction with e-customers in e-business, new ways to understand e-customer behavior and their needs must be found.

Keep in Touch with E-customers Appropriately

While keeping in touch with customers is easier than ever, it is also easier to alienate them. Avoid the appearance of spam and exploitation. Quality principles include the following:

Use appropriate forms of contact with e-customers. This includes new ways of segmenting customers based upon their preferences for forms and nature of contacts.

E-mail periodic reminders to e-customers to verify their personal data if you have not had recent contact. (Always provide value when sending such reminders.)

Customize content options based on learned preferences of e-customers.

Provide varied ways for e-customers to communicate, such as newsletters (always with value-adding content), surveys with results going back to respondents in ways that add value to them, discount opportunities, and ways to "complain" and communicate how your e-business can improve.

Use appropriate e-customer satisfaction surveys to gather likes and dislikes (TARP, 2000). Make them easy to complete. Analyze and *act* on the feedback.

Clickstream Quality Analysis

One way to "see" e-customer behavior is through their clickstream "trail" of the pages they have visited. Quality principles include the following:

Where do e-customers come from (third-party site links; banners, search engines, etc.)?

What terms are used to access a site via search engines? This indicates subjects of interest or customer terminology to enhance finding the site.

What terms do e-customers look up on the site search engine. This identifies areas to expand content and offerings.

What glossary terms do the e-customers tend to look up? This identifies possible ambiguous terms and labels to improve or define.

What inner pages are "bookmarked" and linked to? This indicates content of interest for continuation or expansion.

What pages do e-customers leave from? This may indicate potential complaints, especially if they were in midst of a transaction, such as going through the checkout process with one or more items in the shopping cart, or from a registration page or the privacy statement page. Analyze and understand the following:

- Did they have to go through too many pages?
- Does the privacy policy not engender trust?
- Did they get lost in the site navigation?
- Were the pages ambiguous or too complex?

How long do people stay on various pages? This may identify where they find value. However, excessive length generally indicates interruption.

Limitations of clickstream analysis must be understood to maximize planning and analysis:

Spiders access all pages and should be filtered out.

Masked IP addresses and IP proxies provide anonymity.

Caching of Web pages prevents accurate page trails. The use of dynamic pages increases better page trail data by causing pages to be regenerated each time.

Statistics of average time on a page can be skewed by visitor interruptions.

Time of visits represents the site time of access and does not adequately reflect global e-customer local time of access. Mapping local transaction time, when known, to site access time compensates for this to understand e-customer behavior. Mapping organizational IP addresses to local time zones can provide some adjustment, but cannot compensate for traveling e-customers.

Individual e-customers may access a site from different IP addresses (office, home, and laptop) and single IP addresses may have many individuals accessing a site (library or shared machines).

Focus on Making the E-customer Successful

Because customers measure quality of products and services in terms of how well those products and services help them to be successful, world-class organizations concentrate on the customer's success. The Internet provides new ways to help customers to be more successful:

Use e-customer knowledge to help people avoid making mistakes.

Customize instructions for customers given their unique environment, such as customized installation instructions for a car CD player based on the specific make and model of the automobile.

Use customized e-mail early warnings. Some airlines customize e-mail and e-mobile communications for canceled or delayed flights to their frequent flyers.

Use renewal or reminder notices. Pharmacies send e-mail remainder notices for regular prescriptions before the patient's supply runs out, especially helpful for patients with failing memories.

Contacts with the e-customer should always provide value. E-customers can be alienated if they perceive that they are being contacted only for "selling" purposes.

CONCLUSION

Creating and collecting quality information on the Internet and in e-business transactions has additional complexities and challenges beyond internally controlled information production. This increases the importance of implementing quality management principles, processes, and systemic changes into the culture of the enterprise.

Key components of quality e-business include implementing a customer-centric code of ethics that focuses on understanding e-customer expectations and implementing improvement and quality control in processes that create Web content to engender trust in the e-business and its information. Additional techniques for error-proofing processes in which the e-customers—the new information producers—provide information with quality, both to minimize inadvertent error and to discourage deliberate falsification, are required.

Organizations that do not apply quality management principles to their Internet information and e-business will suboptimize their investment at best or fail at worst. Those organizations who apply quality management principles will reduce the costs of information scrap and rework, reduce the instances of e-customer dissatisfaction and lost business, and will increase opportunity gains and customer lifetime value.

GLOSSARY

Accuracy A characteristic of information quality measuring the degree to which a data value (or set of data values) correctly represents the attributes of the real-world object or event.

Corrective maintenance The process of correcting or reworking defective products to an acceptable level of quality, i.e., correcting defective data, such as missing, inaccurate, or duplicate data; also called scrap and rework, data cleansing, or data scrubbing.

Clickstream quality analysis The activity of using data-mining techniques to analyze patterns of e-customer behavior to identify possible points of dissatisfaction, such as leaving a retail Web site with items in their shopping cart, or frequent links to a site map from a specific page (from which they might be confused).

E-customer Persons or organizations who access information or conduct business over the Internet, including both prospects and purchasing customers.

Gemba Japanese for "the real place"; in information quality, where information is exchanged, gathered, or created in the course of conducting business (adapted from Imai, 1997).

Information producer The role of individuals in which they originate, capture, create, or maintain data or knowledge as a part of their job function or as part of the process they perform; they create the actual information content and are accountable for its accuracy and completeness to meet all information stakeholders' needs. In e-business, the external consumers and business partners are the new information producers.

Information quality (IQ) Consistently meeting all knowledge worker and end-customer expectations in all the characteristics of the information products and services they deem important; the degree to which information consistently meets the requirements and expectations of the knowledge workers in performing their jobs.

Information quality characteristic An aspect of information that an information customer deems important in order to be considered "quality information"; includes completeness, accuracy, nonduplication, timeliness, conciseness, understandability, objectivity, and presentation clarity, among others.

Information quality decay The characteristic of data such that formerly accurate data will become *inaccurate* over time because the characteristic about the real-world object will change without a corresponding update to the data applied; for example, a John Doe's marital status value of "single" in a database is subject to information quality and will become inaccurate the moment he becomes married.

Kaizen Japanese for "improvement," including continuous improvement in all aspects of life, including personal, social, professional, and work. In work, kaizen means continuous improvement involving everyone in the organization, both managers and workers. Trademark of the Kaizen Institute.

Knowledge worker (1) The role of individuals in which they analyze or apply information in any form as part of their job function or in the course of performing a process or making a decision; also referred to as an *information consumer* or *customer*. Accountable for work results created as a result of the use of information and for adhering to any policies governing the security, privacy, and confidentiality of the information used. (2) Coined by Peter Drucker no later than 1973 and refers to business persons or workers in the Information Age as opposed to the term "user."

Overloaded data element A data element that contains more than one type of fact, usually the result of the need to know more types of facts growing faster than the ability to make additions to the data structures, which causes process failure when downstream processes find unexpected data values.

PDCA (Plan–Do–Check–Act) A closed-loop process for planning to solve a problem, implementing suggested improvements, analyzing the results, and standardizing the improvements. Also called a Shewhart cycle after its developer, W. A. Shewhart.

Preventive maintenance The activity of process improvement to error-proof a process and eliminate the causes and occurrences of defects.

Quality Function Deployment (QFD) The methods and techniques to actively involve customers in the design of products and services to better understand and capture customer requirements, and design and develop products and services that better meet their needs on initial product or service delivery.

Record of origin The first source or artifact, such as a paper document, audio or video recording, or electronic recording, in which data are captured or recorded.

Record of reference The single, authoritative document or database for a collection of data elements for a set occurrences of an entity type, representing the most reliable source of operational data for these data elements. In a fragmented data environment, a single occurrence may have different collections of data elements whose record of reference is in different files. Synonyms include database of record and system of record.

TIQM (Total Information Quality Management) A complete methodology for assessing IQ, measuring the costs of nonquality information, correcting data and controlling data transportation and movement, improving processes, and implementing IQ management principles into the culture of the enterprise. TIQM® is a registered trademark of Information Impact International, Inc.

User A term used by many to refer to the relationship of a person to information technology, computer systems, or data, implying dependence on something, or one who has no choice, or one who is not actively involved in the use of something. The term is inappropriate to describe the role of information producers and knowledge workers who perform the work of the enterprise by employing information technology, applications, and information in the process. Thus, the role of business personnel to information technology, applications, and data is one of information producer and knowledge worker, and their relationship to information systems personnel is not as users, but as partners. If Industrial-Age personnel were (machine) "operators" or "workers," then Information-Age personnel are "knowledge workers," a term coined by Peter Drucker in or before 1960 and used by him since then to refer to business persons.

Value chain An end-to-end set of activities that begins with a request from a customer and ends with specific benefits for a customer, either internal or external; also called a business process or value stream.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Consumer-Oriented Electronic Commerce; Data Warehousing and Data Marts; Privacy Law; Value Chain Analysis.*

REFERENCES

Alexander, J. E., & Tate, M. A. (1996). *Evaluating Web resources*. Retrieved January 2, 2003, from http://muse.widener.edu/Wolfgram-Memorial-Library/webevaluation/webeval.htm

Alexander, J. E., & Tate, M. A. (1999). *Web wisdom: How to evaluate and create information quality on the Web*. Mahwah, NJ: Lawrence Erlbaum.

BBBOnline (1996). Retrieved January 2, 2003, from http://www.bbbonline.org/

Beck, S. E. (1997). *The good, the bad & the ugly: Or, why it's a good idea to evaluate Web sources*. Retrieved January 2, 2003, from New Mexico State University, Institute for Technology-Assisted Learning Web site at http://lib.nmsu.edu/instruction/eval.html

Brue, G. (2002). *Six sigma for managers*. New York: McGraw-Hill.

Calvert, P. J. (1999). Web-based misinformation in the context of higher education. *Asian Libraries, 8*(3), 83–91. Retrieved January 2, 2003, from http://www.emeraldinsight.com/pdfs/17308cb2.pdf

Ciolek, T. M., & Goltz, I. M. (1996). *Information quality WWW virtual library*. Retrieved January 2, 2003, from http://www.ciolek.com/WWWVL-InfoQuality.html

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: MIT Center for Advanced Engineering Study.

English, L. P. (1996). *Information quality improvement: Principles, methods, and management.* (Seminar.) Brentwood, TN: Information Impact International.

English, L. P. (1999). *Improving data warehouse and business information quality*. New York: Wiley.

English, L. P. (2000). *e-Quality: Quality in Internet and e-business information.* (Seminar.) Brentwood, TN: Information Impact International.

Feigenbaum, A. V. (1991). *Total quality control*. New York: McGraw-Hill.

Gunning Fog Index (1999). Retrieved January 2, 2003, from http://isu.indstate.edu/nelsons/asbe336/PowerPoint/fog-index.htm

GVU's 10th WWW user survey (GVU). (1998, October). *Falsification of information*. Retrieved January 2, 2003, from http://www.gvu.gatech.edu/user_surveys/survey-1998-10/graphs/general/q65.htm

Huang, K.-T., Lee, Y., & Wang, R. (1999). *Quality information and knowledge*. Upper Saddle River, NJ: Prentice-Hall.

Imai, M. (1997). *Gemba kaizen*. New York: McGraw-Hill.

Information Impact International, Inc. (Infoimpact). (1996). Retrieved January 2, 2003, from http://www.infoimpact.com

Information Mapping, Inc. (Infomap). (1995). Retrieved January 2, 2003, from http://www.infomap.com

Iwata, E. (2000, August 21). Tech's tyranny provokes revolt. USA TODAY, p. 1A–2A.

National Association of Boards of Pharmacy (NABP). (1999). *VIPPS*. Retrieved January 2, 2003, from http://www.nabp.net/vipps/intro.asp

Nielsen, J. (2000). *Designing Web usability*. Indianapolis, IN: New Riders.

National Science Foundation (NSF). (2002). OMB 515 information quality guidelines. Retrieved January 2, 2003, from http://www.nsf.gov/home/pubinfo/nsfinfoqual.pdf

Office of Management and Budget (OMB). (2002). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. *Federal Register, 66*(189), September 28, 2001, with supplements January 3, 2002. Retrieved May 2, 2003, from www.whitehouse.gov/omb/fedreg/reproducible.html

QFD Institute. (1995). *Quality function deployment*. Retrieved January 2, 2003, from http://www.qfdi.org

Redman, T. (2001). *Data quality: The field guide*. Boston: Digital Press.

Segev, A. (1996). On information quality and the WWW impact, a position paper. Berkeley, CA: University of California at Berkeley, Fisher Center for Information Technology and Management, Haas School of Business.

TARP. (2000). Retrieved January 2, 2003, from http://www.e-satisfy.com

TRUSTe. (1997). Retrieved January 2, 2003, from http://www.truste.org/

Useit. (1996). *useit.com: Jakob Nielsen's Website*. Retrieved January 2, 2003, from http://www.useit.com

Virtual Chase. (1998). *Evaluating the quality of information on the Internet*. Retrieved May 2, 2003, from http://www.virtualchase.com/quality

Integrated Services Digital Network (ISDN): Narrowband and Broadband Services and Applications

John S. Thompson, *University of Colorado at Boulder*

INTRODUCTION

A common topic of discussion in telecommunications today is the notion of converged networks in which one infrastructure is used to carry voice, data, and multimedia content. This concept is often presented as novel and unique to the Internet era. Although the emphasis in today's discussion is on a common transport technology set, converged networks also must include common interface and operational technologies as well. In this sense, integrated services digital network (ISDN) and broadband ISDN (B-ISDN) represent two earlier attempts at converged telecommunication networks.

Although the deployment of ISDN in the world's public telecommunication networks has not reached the levels its developers had anticipated, the technologies that underlie it have had an impact far beyond the number of terminals and switches based on these standards. In fact, the very body of standards that define ISDN and B-ISDN represents a realization that a coherent, standards-based approach to telecommunications technology is the only way to ensure consistent progress and provide fair access to the market for all customers, equipment vendors, and service providers.

The development of standards for ISDN may also represent one approach to this process that can be contrasted with the approach that has characterized that for the Internet. In the ISDN model, based in the International Telecommunications Union (ITU) body of multinational technology experts, standards are created and agreed in total before any deployment generally occurs. This normally means that there is a considerable development period before information is available for guidance on products and services. In the Internet model, managed by the Internet Engineering Task Force (IETF), standards are developed, tested, and deployed incrementally, which allows the process to track more readily rapid changes in technological capability. This latter approach, however, does not always reflect the full impact of new concepts and will often result in the need to render previous standards obsolete with some resulting disruption in the market.

It can be argued that ISDN and B-ISDN are both victims of a slow-paced, deliberative standards process that was overtaken by evolving technology and comprehensive attempts to codify the methods of providing telecommunications to end users that has identified virtually all of the needs of multimedia communication and provided stability to the market. Neither side of the argument is willing to concede much to the other. Perhaps ISDN, especially the narrowband version, is a victim of bad timing. When it was initially defined in the 1970s and 1980s, no one could conceive of a need for end user capabilities as high as 64,000 bits per second (kbps) data communication; once it was ready for full deployment in the 1990s, and the Internet was upon us, no one could imagine a need for such a low rate!

History and Origins

By the end of the 1960s, it was clear that the transition from telephone networks based on analog technology to those based on digital technology was inevitable and would be complete within a few decades. It was equally clear that digital technology options were much more diverse than the analog predecessors and progress could only occur if broad agreements could be reached among all parties with a significant stake in the future of the industry.

The ITU formally assigned the question of how to develop international standards for digital telecommunication networks to study groups in this period, and over the next 24 years a series of standards documents were

Table 1 Principles of ISDN

1.1	The main feature of the ISDN concept is the support of a wide range of voice and nonvoice applications in the same network. A key element of service integration for an ISDN is the provision of a range of services using a limited set of connection types and multipurpose user–network interface arrangements.
1.2	ISDNs support a variety of applications including both switched and nonswitched connections. Switched connections in an ISDN include both circuit-switched and packet-switched connections and their concatenations.
1.3	As far as practicable, new services introduced into an ISDN should be arranged to be compatible with 64 kbit/s switched digital connections.
1.4	An ISDN will contain intelligence for the purpose of providing service features, maintenance, and network management functions. This intelligence may not be sufficient for some new services and may have to be supplemented by either additional intelligence within the network, or possibly compatible intelligence in the user terminals.
1.5	A layered protocol structure should be used for the specification of the access to an ISDN. Access from a user to ISDN resources may vary depending upon the service required and upon the status of implementation of national ISDNs.
1.6	It is recognized that ISDNs may be implemented in a variety of configurations according to specific national situations.

Note: Source ITU-TI.120. Reprinted with permission.

produced that progressively defined the details of design and operation of ISDN. The principles of this work are provided in ITU-T Standard I.120 (1993) and are shown in Table 1.

The standards developers also recognized that the movement to full deployment of ISDN would require a considerable period of time and defined several principles of "evolution" to guide it (ITU-T, 1993). Unfortunately, they did not recognize that the Internet "revolution" would overtake many of their principles. This first effort can be considered narrowband ISDN (N-ISDN) because it is based on a synchronous time division multiplexed structure consisting of increments of communication "bandwidth" of 64,000 bits per second (64 kbps).

ISDN developers also recognized the need for greater transmission rates that could not be accommodated by integral multiples of 64 kbps only and for communication that did not require continuous and dedicated access to any fixed bandwidth value. As a result, broadband ISDN (B-ISDN) was conceived and based on the notion of asynchronous transport using fixed-size data cells (ITU-T, 1997). This asynchronous transfer mode (ATM) is contrasted with the synchronous transfer mode (STM) approach of N-ISDN. ATM held the promise of being the foundation for all digital communication regardless of speed or other characteristics. In fact, it can be considered the second attempt at a common converged infrastructure for telecommunications.

Current Deployment and Significance in Today's Markets

Narrowband ISDN (N-ISDN) is now deployed in almost 60% of the countries of the world, having grown from just over 4% since 1990. This deployment is uneven, however, with the less developed countries generally having little or no ISDN deployment and the wealthy ones having the greatest penetration. The countries of Europe have the greatest ISDN deployment, and the highest proportion of ISDN channels as a fraction of all telephony access channels reaches more than 49% in Norway (other examples include Germany at 35.7% and Netherlands at 30%). Japan (29.7%) and the Republic of Korea (10.7%) are the exception to relatively low deployment in Asia, and the United States (6.7%) and Canada (4.4%) are at the top of the list for the meager deployment in the Western Hemisphere. Nevertheless, these are important, if not overwhelming numbers (ITU, 2002). N-ISDN is certainly a player today in the world telecommunications market, but this is not the end of the story.

When first conceived, N-ISDN, especially the BRI version (described later in the chapter), was designed to provide multichannel, high-speed (64-kbps) connection over single subscriber loops. This, it was argued, would serve the needs of any single-line customer who wanted simultaneous voice and data communication. Since that time, the slow pace of N-ISDN deployment and the development of 56-kbps analog modems drove the market for second-line service in many areas. Furthermore, broader band services such as high-speed asymmetric digital subscriber line (ADSL) on local loops and cable modems on the ubiquitous cable networks have presented a significant challenge to the 64- or 128-kbps capability of N-ISDN for those customers demanding broadband access.

The existence of N-ISDN standards and services also helped support the development and deployment of multimedia communications that includes voice, video, and simultaneous data sharing in a single connection. The ITU-T standard H.320, titled Visual Telephone System, specifies the use of N-ISDN services for such calls. This standard, and others related to it, have helped advance this higher function form of communication in both the switched channel telephone network and in the packet-based Internet.

Although N-ISDN was developed and initially deployed as a service set supported by the conventional, but evolving, public telephony networks, the access technology for it has also seen life as one of a host of digital subscriber line (DSL) services that provide end users with both traditional telephone service and simultaneous wider band data service for Internet access on a single copper wire pair. This service, known as IDSL, uses the same line coding techniques and frame format as the N-ISDN basic rate interface (BRI) service and provides 144 kbps of data communication bandwidth between a user's premises and the central office (CO). At the CO, the connection is diverted directly to an Internet service provider's (ISP) network. IDSL does not provide any switched service, including and in particular, voice service, as found in N-ISDN BRI. This is just the latest of a series of derivative uses for ISDN technology.

N-ISDN also sees service in the video teleconferencing market. The first of a series on ITU-T standards, H.320, defined a Visual Telephone System that operates on multiples of 64-kbps bearer channels up to 1,920 kbps using ISDN call control and management. This standard specifies video, audio, and data services for multimedia communication and also defines the concept of a multipoint control unit (MCU) that permits more than two parties to participate in a conference using all three communication modes. This work has also served as the foundation of one option for call control standards in multimedia communication in packet communication networks such as the Internet.

The significance of ISDN today derives both from its deployment, principally as a narrowband offering, and from the foundation of standards it has spawned in call signaling, frame relay and ATM transport services. For example, frame relay service access ports, as a percent of all commercial broadband access in 2001 was about 35% and was exceeded only by traditional private lines leased largely from traditional telephony providers (38%; E Dunne, 2002; information available at http://www.verticalsystems.com). ATM access service is only a small fraction of this market (less that 1%), but this figure does not begin to represent the importance of the role ATM plays in core networks. The signaling protocols of ISDN not only are redeployed in these derived services for connection setup and management, but they have also been adopted in one of the leading techniques for packetized speech services in the Internet protocol (IP) networks that are taking hold in telecommunications networks today (ITU-T Standard H.323, 1999).

Standards and Organization

Standards development for ISDN was initiated by and has remained largely in the hands of the ITU-T. The "I-series" of standards of that organization are devoted to ISDN issues and include eight categories: general structure, service capabilities, overall network aspects and functions, user–network interfaces, internetwork interfaces, maintenance principles, and B-ISDN equipment aspects (ITU-T Standard I.120, 1993). These standards, and those from other series documents associated with them, generally serve most of the countries of the world for purposes of product and service definition.

In North America, however, regional supplements to international standards were required both to complete definitions left open at the international level and to provide for specific needs in the United State and Canada. These standards are provided through the American National Standards Institute (ANSI). The most significant of these is perhaps that for the physical layer definition for two-wire communication for basic rate service between the customer premises and the attached central office—ANSI T1.601.

Despite all of the intense effort to develop standards, there were still many unanswered questions that had to be resolved before broad-based deployment could occur in North America. The Bell Operating Companies and Bellcore (now Telcordia, see http://www.telcordia.com) developed a series of plans to complete this work and called it National ISDN. Known as NI-1 through NI-4 (for National ISDN-1, etc.), these plans specified details for access, call control, and signaling, uniform interface configurations for basic rate interface, uniformity of BRI services, primary rate interface (PRI) capabilities, data capabilities, operations support capabilities, and billing capabilities (Telcordia, 1995). The National ISDN agreements were crucial to deployment in North America but also introduced yet another delay in the process for that region.

NARROWBAND ISDN

The evolution from analog to digital telephony was based on the concept of synchronous time division multiplexing in which each communication channel was provided a fixed amount of bandwidth suitable to carrying a single, band-limited voice conversation. The standard for such conversations were established with the first digital carrier systems deployed in the early 1960s at an 8 kHz sample rate with 8-bit digital samples based on a nonlinear, compressed code. The basic voice channel rate thus became 64 kbps. Narrowband ISDN is based on providing various combinations of these "bearer" channels between the customer premises and a central office (CO) switch over copper wire access loops.

Basic Rate and Primary Rate Interface Common Features

N-ISDN defines two classes of service and interface: basic rate interface (BRI) and primary rate interface (PRI). Both interfaces include some number of 64-kbps bearer channels and a common signaling and service channel. The bearer channels are called B channels, and the signaling channel is known as the D channel. Both interfaces have a time division multiplexed line format based on a fixed frame structure that has an underlying repetition rate of 8,000, or some submultiple of 8,000, frames per second. This means that the bearer channels, when used for voice transmission, can be synchronized with the telephone network by placing one or more samples in each frame. This supports the end-to-end synchronization of voice communication from one customer premises terminal to another.

The BRI frame has two 64-kbps B channels and one 16-kbps D channel, whereas the PRI has either 23 or 30

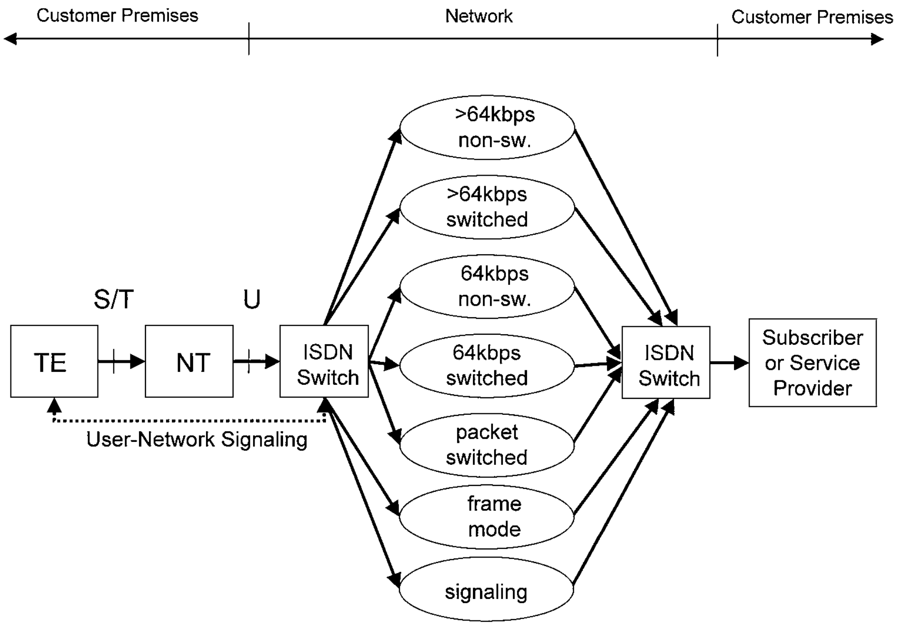


Figure 1: ISDN architectural elements and reference points.

B channels and one 64-kbps D channel. The difference in PRI B channel support depends on the underlying physical layer bit rate of 1.544 Mbps or 2.048 Mbps (derived historically from the T1 or E1 digital hierarchy carrier systems developed in North America and Europe) respectively. BRI service is commonly used for individual subscriber needs, whereas PRI service typically is used to support customer premises switching equipment such as a private branch exchange (PBX).

D channel operations for both BRI and PRI are primarily for signaling to establish and disconnect calls on the B channels associated with the interface. This signaling supports both origination and termination of calls and specifies how the associated B channels will be used including such characteristics as voice coding technique or data multiplexing functions. D channel operations also support direct data communication services as well (see Data Link Layer and Operations later in the chapter).

Physical Layer and Operations

Because BRI is intended to serve as a high-function alternative to conventional analog single-user service, it must be provided on the single wire pair access loop now connecting most users to the central or end office of the public switched telephone network (PSTN). Because BRI service must support at least a $2 \times 64 + 16 = 144$ kbps rate in both directions, this means that the access loop must be capable of supporting a bandwidth much greater than the 4 kHz normally associated with an analog voice connection and that some means must be established to keep the signals traveling in the two directions from interfering with each other.

PRI service is defined as using two wire pairs, but again, this will generally be the same physical wire plant used for single-user service and must support 1.544 or 2.048 Mbps rates. In fact, both BRI and PRI are examples of the general class of access services known as DSL services. All of these services take advantage of the fact that telephone access lines have a much greater effective bandwidth than that needed for simple analog telephone service. The rates that can be carried and the information formats used, however, depend on loop length and a variety of potential impairments that can only be fully determined by direct measurement by service providers.

BRI

Basic rate service is defined in ITU standards to have several reference interfaces both on the customer premises and at the point between the customer premises and the central office (see Figure 1). The “U” reference point is the two-wire connection point and operates at a 160 kbps rate using the 2B1Q (two binary, one quaternary) multilevel line. (This is the line coding used in North America. Different techniques are used in other parts of the world.) This results in an 80-kilobaud signaling rate and minimizes the bandwidth requirement on the local loop. The difference between 160 kbps and minimum 144 kbps occurs because 16 kbps of framing and maintenance information is included in the BRI U reference frame format.

The U reference point is converted to the “T” reference point at the customer premises through the network termination 1 (NT1) functional device. Among other functions, the NT1 device converts two-way, single-pair transmission to one-way, two-pair transmission. A network termination 2 (NT2) device may be realized separately and further defines an “S” reference point where ISDN compliant terminals can be connected. Alternatively, a terminal adapter (TA) device can connect at the “S” reference point to permit support of conventional analog telephones and data communication gear.

The data format at the S/T reference points operates at 192 kbps using a pseudoternary line code. Once again, the difference between 144 kbps and 192 kbps is accounted for by the addition of framing and maintenance information included in the frame format at this point. This physical level detail is specified in the ITU-T Standard I.430.

There is a variety of BRI wiring arrangements that permit up to eight compatible ISDN terminals to connect in parallel at the S/T reference point. Each of these terminals can be assigned its own directory number and operate independently within the limitation that only two B channels are available for use at any one time. It is most common, however, for only a single terminal to be deployed for ISDN BRI service and for that terminal to be a combination telephone hand set and data communication termination for a PC or local area network or as data network router capable of automatic connection to a remote site.

PRI

Because the PRI in North America, Europe, and the rest of the world is already defined to use existing physical level signals currently in place in the digital hierarchy, introducing this capability into the telephone physical plant was somewhat easier than BRI. The specific details for the electrical interfaces and the frame structure are defined in ITU-T standards G.703 and G.704 respectively. As in the case of BRI, the data rate for PRI is slightly greater than the sum of the B and D channel rates to allow for framing and maintenance features.

Regardless of the physical format of the communications link, BRI or PRI, the physical connection must be complete, and end-to-end, bit-level synchronization must be established before any data link level communication can proceed. This step-by-step initialization moving up the levels in the open system interconnection (OSI) reference model is common to all data communication systems, including ISDN.

Data Link Layer and Operations

N-ISDN represents the first time telecommunications standards identified separate, general-purpose user and control "planes" within a single physical transmission facility. Although this notion is now so common within data communication networks such as the Internet that we don't think much about it, in the 1960s, it was quite novel. Within N-ISDN control and user planes exist in separate portions of the transport frame, the B and D channels, but there is also a further separation within the D channel itself that allows that system component to serve as both a signaling and a transport channel.

8	16	Variable	8/16	16	8	bits
Flag	FCS	Information	Control	Address	Flag	

Figure 2: LAPD packet format.

ISDN services, both narrow- and broadband, are most extensively defined at the user–network interface (UNI). This means that while end-to-end information may flow unchanged through the network (e.g., 64-kbps pulse code modulated (PCM) encoded speech), the precise physical, data link, and network level definitions for carrying this information may only be defined at the UNI. Typically, then, ISDN access links terminate at the central office where B channels will be synchronously switched into the larger time division modulated public switched telephone network (TDM PSTN), and the D channel will terminate on CO software control functions that determine when, and how, those B channels will be switched. ISDN standards also define terminations to a variety of other services (see Figure 1), but 64-kbps channelized service is the most common.

The D channel, then, operates as a point-to-point, synchronous data communications link between the customer premises and the CO. At the data link level, the 16- or 64-kbps channel functions as a continuous bit stream even though the individual bits are partitioned and carried in separate frames in the underlying physical layer frame format of BRI or PRI. Information in the D channel is carried in packets delimited by a unique start and stop pattern of bits called a flag (see Figure 2). The structure of this packet, including the flag, is taken from the high level data-link control (HDLC) format that has been used in synchronous data communication from its earliest days (e.g., X.25 and various other standard and proprietary systems).

The D channel packet structure and its operation are known as the link access protocol—D channel, or LAPD, and is one of a class of protocols that has evolved over a number of years to support reliable transport of information across a single data communications link. The key features of the LAPD packet format are, in addition to the delimiting flag patterns, the address, control, information, and frame check sequence (FCS) fields. LAPD is defined in the ITU-T standard Q.921.

The information field is variable in length and carries the actual control information that is interpreted at the network and higher layers of the ISDN protocols and services. The FCS field is a 16-bit cyclic redundancy check code that serves to detect single-bit errors in the overall LAPD packet that can then be used to trigger error control procedures.

The address field is divided into two subfields: the service access point identifier (SAPI) and the terminal endpoint identifier (TEI; see Figure 3). The SAPI value determines what functional service the packet will perform. For example, SAPI = 0 is the value for call control procedures used to set up the use of associated B channels.

10-16	9	3-8	2	1	bits
Terminal Endpoint Identifier	1	Service Access Identifier	C/ R	0	

Figure 3: LAPD address field format.

Other significant values are 16 for X.25 communication on the D channel, 32–61 for frame relay communication on the D channel, and 63 for D channel management functions. The SAPI field is the way LAPD multiplexes the D channel for different functions. Although call control information clearly must terminate on the CO, X.25 and frame relay information would likely terminate with another end user and this implies that the CO would have to include a packet forwarding capability if this service is to operate (see Figure 1).

The TEI field supports the concept of multiple terminal end points at the user premises as in the case of BRI. TEI values can be manually assigned but more likely the CO will automatically assign them as part of an initialization procedure when the ISDN equipment is first connected. Note that the TEI does not relate to B channel use but to specific terminals that may use a given B channel that is established at the time of call setup. The TEI value of 127 is used as a broadcast value when no specific value has been assigned or appropriate.

The control field has several formats depending on the type of information being exchanged. Seven bit send and receive sequence numbers are used to support reliable communication in LAPD by means of acknowledged transfer in which each end of the link is able to track the last packet that was successfully received at the far end and announce what was last successfully received at its end. This technique is used in the basic "information" transfer mode. The "supervisory" mode is used when only one-way acknowledgment is required and the "unnumbered" mode is used when unreliable communication is appropriate, as in the case of the link initialization required when LAPD communication is first established on the D channel. Initialization for LAPD includes assigning a TEI value using the unnumbered information packet type and the set asynchronous balanced mode extended (SABME) packet type in the unnumbered control mode. The D channel is now ready for network layer communication (e.g., call setup).

Network Layer and Operations

The connection-oriented communication model is fundamental to all ISDN services (narrow- and broadband). This means that communication has three distinct phases—setup, information transfer, and disconnect—and that there is some notion of a system resource being reserved for the duration of the communication. The principle resource being reserved in N-ISDN is a B channel. D channel network layer operations serve to exchange information between user end points and the network to establish this reservation for the duration of a call, and the designated B channel resources are then used for communicating information between end points.

Therefore, call control communication in the D channel must be able to identify both the B channel resources in question and the particular call instance of concern. It is also important to be able to describe how the network is to treat the B channels during end-to-end communication in terms of any special coding or other manipulations. All of this capability is contained in the protocols defined for ISDN network layer services (see Figure 4).

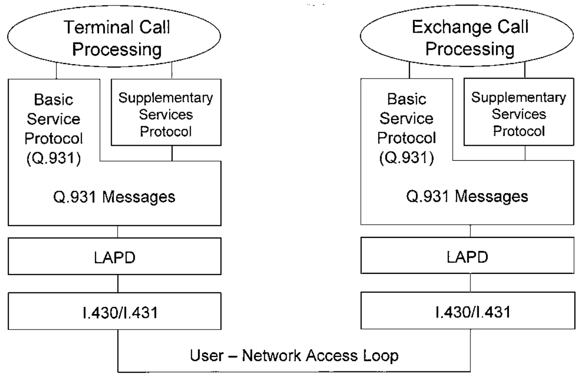

Terminal Call Processing
Basic Service Protocol (Q.931)
Supplementary Services Protocol
Q.931 Messages
LAPD
I.430/I.431
Exchange Call Processing
Basic Service Protocol (Q.931)
Supplementary Services Protocol
Q.931 Messages
LAPD
I.430/I.431
User – Network Access Loop

Figure 4: ISDN protocol and service model.

ITU-T standards Q.930, Q.931, and Q.932 are the principal specifications for N-ISDN call control. These standards and the call management model they represent not only define ISDN but have also provided a call management model for network-wide signaling using Signaling System 7, are the foundation for connection management for B-ISDN (ATM) and frame relay services, and have also been incorporated into one of principal competing standards for voice and multimedia services on IP networks, H.323.

The basic Q.931 call control message is carried as the payload in a LAPD packet between the user end point and the CO (see Figure 5). These messages have four common features:

1. A protocol discriminator that uniquely identifies the payload as a Q.931 defined message
2. A call reference value that uniquely associates all information about one call for its entire duration
3. A message-type indicator that determines which of a defined set of call control messages is being invoked
4. A set of one or more information elements that carry all of the call management detail required, realizing the particular call control message

A typical call control procedure will include a number of common message types. For example, in ISDN, some end user on customer premises equipment always initiates calls. That user's terminal equipment will begin

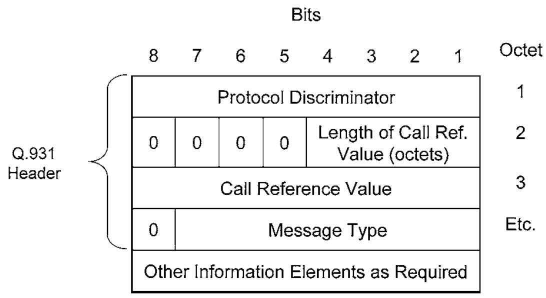


Figure 5: Q.931 call control message format.

by issuing a SETUP message type to the CO. This message must contain a Bearer Capability information element that describes the required bearer channel needs and how they are to be used in the call. The message may also contain such important information elements as calling and called party numbers, call progress information, and display information useful to the human end users.

SETUP messages are acknowledged, as are most ISDN messages, and provide a means to indicate success or failure of the original message. PROGRESS and CALL PROCEEDING messages are included in the Q.931 protocol repertoire to allow the network and terminals to signal that things are progressing as expected. The ALERTING and CONNECT messages allow end points and the network to signal that a connection is being attempted and is complete, respectively.

Finally, either end of a two-party call can initiate termination. This is done with a DISCONNECT message. Recall that Q.931 messages includes a call reference value, and it is this value that allows the terminal equipment and the CO to agree that they are each referring to the same connection at any given time. This value can also be used by terminal and network equipment to maintain records for billing and other purposes, although these features are not defined in the standards. All of the details of call processing in both terminal and network switching equipment beyond the Q.931 protocol definitions are the subject of software that is proprietary to specific equipment vendors.

Extended Bearer Channel Operations

The simplest and most common use of ISDN service is to provide a clear 64-kbps channel to transport either voice or data information between two end points. In such cases, the SETUP message, and its bear capability information element, will define how the B channel will be used. The ISDN standards, however, anticipated a number of other, more complex uses of connection bandwidth.

For example, 64 kbps may not be enough capacity for certain services, such as data or video communication. In this case, capabilities are defined to support fixed multiples of B channel capacity in one connection. A common service might be 2 × 64 kbps to allow the use of both B channels in BRI service to provide a broader band connection for terminating LAN routing gear that must communicate with a remote network location.

Other examples of "$n \times 64$" service are defined for PRI service in which specific aggregations are defined, such as $H_0 = 384$ kbps ($n = 6$), $H11 = 1536$ kbps ($n = 24$, 1.544 Mbps PRI), and $H12 = 1920$ ($n = 30$, 2.048 Mbps PRI). Standard services also support the case for a general value of n for applications in which one of the fixed bandwidth services is not appropriate. In all cases, however, the network resource, in terms of B channels, is reserved for the duration of the call, and this is one of the basic criticisms of N-ISDN: It may not make efficient use of network bandwidth resources for information transport in which the data rate may be irregular during the period of connection.

At the other end of the speed spectrum, there is also a need to provide communication channels with data rates less than 64 kbps. Although this is not as common today as when the ISDN standards were first developed, considerable attention went into creating standard mechanisms for multiplexing several data signals into one 64-kbps B channel. Terminal adaptors at the user premises are responsible for the multiplexing function and, because of the nature of N-ISDN service, individual data signals multiplexed into one B channel must originate and terminate together—there is no "add–drop" capability defined.

Drivers in General Telephony and Data Services

N-ISDN standards developed at a time when there was tremendous growth in digital telephony around the globe. This growth drove an increasing need for uniform and reliable techniques for long distance network signaling. The development of Signaling System Number 7 (SS7; see ANSI standard T.110) almost exactly parallels that for N-ISDN and complements the end user to CO signaling support of ISDN with network to network signaling needed to complete calls between arbitrarily different customer premises.

Even though common channel signaling required for conventional analog telephony represented a major advance in long-distance calling reliability and flexibility, ISDN represented a new richness in call features that placed new demands on the networks. It makes little sense to allow an ISDN end office to support a complex call initiation by an end user if the detail of the required end-to-end connection cannot be communicated across the network to the other party and facilities needed to support it cannot be reserved within the network itself.

An equivalent network layer protocol has been defined in the SS7 protocol system, called ISDN user part (ISUP; ANSI Standard T1.113), for just this purpose. ISUP is not only capable of supporting, through translation, all of the Q.931 messages required in ISDN but it is also a sufficiently significant advance of the earlier SS7 protocol, telephone user part (TUP), that it is becoming the dominant signaling protocol for all service, conventional and ISDN alike.

Signaling is not the only area in which ISDN standards have affected telecommunications beyond the actual deployment of the service itself. Recall that part of the D channel LAPD specification included the ability to support end user to end user data communication using something called frame relay. The 30 values from SAPI = 32 to SAPI = 61 can each designate a unique logical communication connection on the ISDN D channel. Because LAPD operates at Layer 2 of the OSI reference model, frame relay service provides a basic point-to-point transport for arbitrary packet information.

The value of frame relay services is so great that separate distinct standards have been developed by the ITU-T, including Q.922 which is the specification for frame mode bearer services, Q.933, the frame relay call control procedures, and I.122, the overall framework for frame mode bearer services. From a commercial perspective, however, these international standards were not adequate. As a result, the Frame Relay Forum was formed and has provided a set of implementation agreements that has guided the industry since the early 1990s. Details on the forum and the agreements they have produced can be found at http://www.frforum.com.

BROADBAND ISDN

Even as the narrowband ISDN standards were being completed and deployed in the 1980s, standards developers recognized that the restriction of assigning multiples of 64 kbps to connections would not be adequate for the indefinite future. It was also recognized that dedicating transport bandwidth, no matter what amount, to a connection for its entire duration was also not appropriate for a great many, nonvoice services. As a result, development began on standards for broadband ISDN (B-ISDN) as a complement and possible successor to N-ISDN. A basic definition of terms and the overall aspects of B-ISDN can be found in ITU-T standards I.113 and I.121, respectively.

Rationale and Worldview

When measured in bits per second, the total amount of information passing along the world's telecommunications networks that represents voice has continuously decreased, as a proportion of the total, relative to that for other forms of information. At some point in the late 1990s general data began to exceed voice communication as the dominant form of traffic. It is only reasonable to assume that the underlying network for telecommunications would be optimized for the principal type of traffic. In a sense, this was the design principle behind B-ISDN.

Even if one takes a coarse view of classification of telecommunications traffic consisting of voice, data, and video, it is possible to identify several distinct ways to characterize each traffic type and its potential impact on the underlying transport network (DuBoise & Kim, 1992). For example, in addition to the basic average bandwidth required, there is a simple binary distinction of whether the signal requires continuous or time-variable access to this bandwidth. If variable access is required, one can further define the signal in terms of how (proportionately) often access is needed (burst ratio) and how long that access persists (burst length).

Finally, information flows can be characterized in terms of the quality and reliability of the transport they require. Such measures include data loss and error tolerance and overall delay in transport from end to end. The ranges in all these measures are considerable: bandwidth from a few kilobits per second to a few gigabits per second; burst ratios from 1 (continuous transmission such as voice) to 1,000 (Internet World Wide Web access); burst lengths of up to 100 megabits (digital images); loss tolerance from one error in 10^4 bit (speech) to one error in 10^{12} (high-resolution graphics); delay tolerance from less than 10 ms to more than 10 s or more. The B-ISDN standards developers attempted to address this entire spectrum of needs and, in this mission, set about creating a second generation of converged services network.

The B-ISDN Services Vision

Despite the expansive worldview taken in B-ISDN standards development, there is recognition that some form of backward compatibility with and migration from N-ISDN is essential. The vision is that B-ISDN can coexist with N-ISDN and that some subscribers will choose one or the other according to their needs. There is some similarity between the high-level views of architecture for the two classes, however (see Figure 6).

The same concept of architectural reference points exists in B-ISDN to help delineate functions in the network and on the customer premises and help ensure interoperability of service and product components from different providers. A set of standard access capabilities connecting the user to the CO is also provided for B-ISDN as well. In this case, however, the fundamental rates are 150 and 600 Mbps and the access medium is optical fiber instead of copper wire pairs.

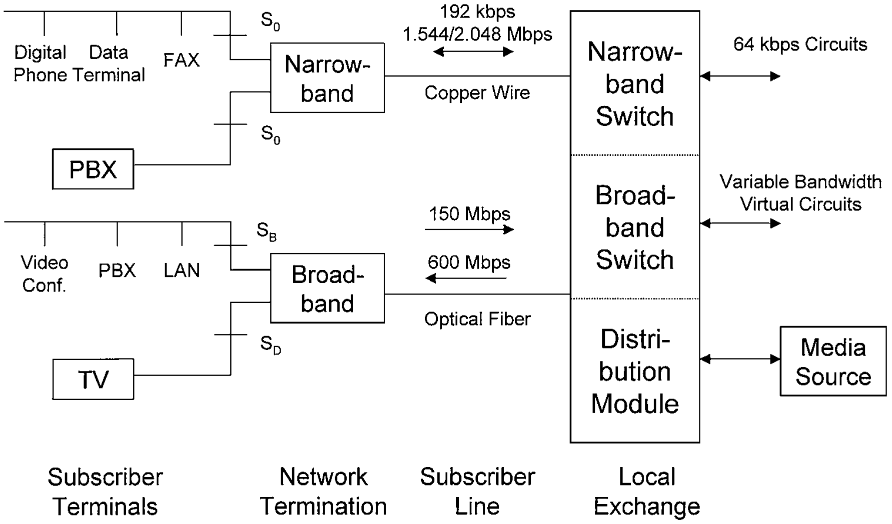


Figure 6: User-network alternatives for broadband ISDN.

B-ISDN services extend from the traditional narrowband functions of telephony to the distribution of entertainment functions we currently associate with cable or satellite television broadcasting. In other words, B-ISDN is to fulfill the promise of N-ISDN by supporting the remainder of telecommunication services not adequately or efficiently supported by 64-kbps transport.

The Decision for Cell Transport and ATM Overview

Perhaps the most significant decision in the design of standards for B-ISDN is the one to abandon STM operation. In STM, all connections, whether single or multiple channels of 64 kbps, are transported and switched in a synchronous manner end to end. This means that every connection has a fixed delay for the duration of the connection. It also means that the entire network must be effectively synchronized to preserve the underlying, common 8-kHz sample rate that is the foundation of the basic B channel service.

B-ISDN had to support not only a wide range of data rates but also the burst ratios that implied that network bandwidth was not continuously required throughout the connection. The solution was to build B-ISDN transport not on a fixed unit of constant bandwidth but to use a fixed unit of information transport size, called a cell. A further decision was made to abandon network-wide synchronization and adopt and ATM of transmission. Actually the two decisions are linked once one considers the network behaviors necessary to switch cells across different network nodes on their way between end points.

In STM, the addressing of channels across the network is separate from the transport of information and switching points preserve the temporal order of the share of time devoted to each connection. In ATM, even though a source may generate information for cells at a constant rate (e.g., voice), other cells for other connections will be inserted and removed as the streams progress through the broadband network. This means each cell must be labeled to associate it with a particular connection.

The cell size chosen was 53 octets of which 5 octets form the header of the cell, which includes the connection identification, and 48 octets contain the information being transported. B-ISDN standards (ITU-T Standard I.121, 1991) define a layered structure in which the physical layer can be a variety of capabilities ranging from 25 Mbps on up. Layer 2 is divided into the ATM sublayer that carries the basic ATM cells and the ATM adaptation (sub-)layer (AAL; see Figure 7). The AAL takes data from higher layers in the OSI reference model and "adapts" it for transport in ATM cells. Currently there are four basic AAL protocols in general use, AAL1, AAL2, AAL3/4, and AAL5, and each one is designed to support a particular class of service including a circuit emulation service (CES in AAL1) that emulates the N-ISDN 64-kbps B channel.

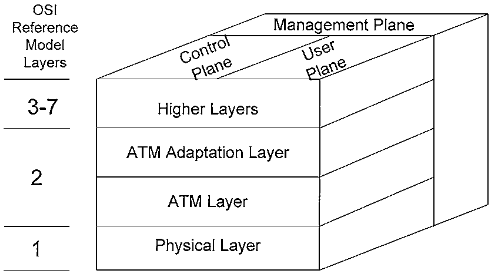


Figure 7: ATM protocol architecture.

An ATM connection is a virtual circuit connection, which means that, although there is no dedicated and reserved bandwidth as in STM systems, there still are logically reserved resources assigned to it. This is accomplished by assigning virtual connection identifier to each ATM cell associated with a connection. This virtual connection identifier actually consists of two parts: the virtual path identifier (VPI) and the virtual channel identifier (VCI) (see Figure 8). This two-part structure gives a logical feel to ATM service similar to the individual trunks (VCI) and trunk groups (VPI) familiar from traditional telephony. In fact, the B-ISDN standards note that ATM switching, maintenance, and other operations can occur at either or both the VCI and VPI logical level.

Because each cell is uniquely identified with its associated connection, it is possible to treat different connections in such a way as to provide better service to one over another. The ATM standards define how this can be done and how switches and call management functions must operate to support the requirements of different types of communication service.

ATM defines both permanent virtual circuits (PVC) and switched virtual circuits (SVC). PVCs are manually administered between end points and are intended to provide dedicated transport for customers. As with frame relay service, which has the same two circuit forms, PVC service is the overwhelming form of ATM service provided today. In effect, ATM service is a private line, dedicated service in the current market (see the following section, Development and Deployment of ATM and B-ISDN).

SVC service, as the name implies, uses standardized call setup and management procedures to establish a virtual circuit. The call management protocol is very similar to the Q.931 standard used for N-ISDN but is sufficiently enlarged that it is defined in a separate set of standards beginning with Q.2931. One other difference in ATM over N-ISDN is that the control channel for signaling is not

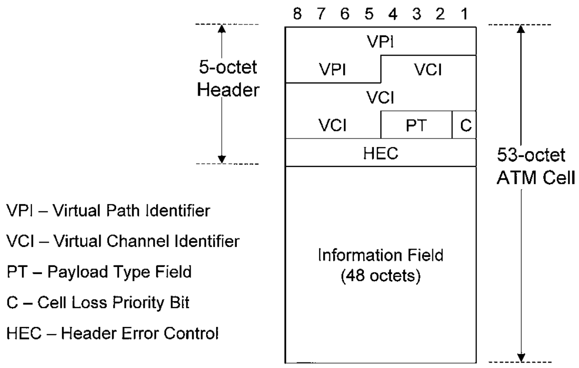

8 7 6 5 4 3 2 1
5-octet Header
VPI
VPI
VCI
VCI
VCI
PT
C
HEC
Information Field (48 octets)
53-octet ATM Cell
VPI – Virtual Path Identifier
VCI – Virtual Channel Identifier
PT – Payload Type Field
C – Cell Loss Priority Bit
HEC – Header Error Control

Figure 8: ATM cell format (internal network version).

physically distinct the way the D channel is but, instead, is simply a dedicated virtual circuit (VCI–VPI value pair) that is only used for this purpose.

Because ATM and B-ISDN were to be carried on optical fiber connections from the beginning, it is logical to assume that one other set of standards being developed at this same time, synchronous optical network (SONET) and the synchronous digital hierarchy (SDH), would also be closely related to this service. SONET/SDH is an 8-kHz framed optical transport system with multiple physical bit rates extending from approximately 50 mbps/ 150 Mbps (OC-1/OC-3) to (currently) nearly 40 Gbps (OC-768) and which can support both STM services as well as ATM (ATM Forum, 2000). As such, it has become the backbone optical service for traditional telephony service providers and customers alike.

Development and Deployment of ATM and B-ISDN

B-ISDN, like N-ISDN, was seen primarily from the perspective of the user–network interface and was driven by an expected flood of demand for new broadband services by end customers. As the market developed, the demand came first from telecommunication service providers who saw ATM (and the conveniently available SONET/SDH physical transport technologies) and not the entire B-ISDN model, as a desirable internal structure for their networks. Just as in the case of frame relay, the demand for completed standards outstripped the capacity of the ITU-T standards development process and the industry turned to a consortium of equipment manufactures, service providers, and customers. This consortium, the ATM Forum, has become a principal source of ATM standards for this service today (http://www.atmforum.com) through a set of implementation agreements that are aligned with ITU-T standards where they exist and are complete.

Although B-ISDN, as such, has not really been deployed, ATM service, including direct customer access service, has been widely deployed. ATM service and networks, based on ITU-T and ATM Forum standards, is used primarily for core network service for service providers, including those providing frame relay service to end users. Direct access to ATM, however, remains at only a few percent of the market for customer service (E. Dunne, personal communication, 2002).

ISDN APPLICATION TO THE INTERNET AND E-COMMERCE

The relationship between ISDN and the Internet divides along a narrowband–broadband fracture line. N-ISDN, with its 64-kbps dedicated bandwidth service, now serves as the low end of the digital access services available to customers who cannot be satisfied with (dial-up) analog modem service. B-ISDN, through the deployment of ATM technology and services largely through traditional telecommunication service providers, makes up a significant, and still growing, part of the backbone transport for ISPs (E. Dunne, personal communication, 2002).

Whether the 64-kbps (or 128-kbps) bandwidth seems to be a limitation, of course, depends on the end user's requirements and the alternatives that are available. The assumption that bandwidth demand for Internet access can only continue to grow dramatically seems to have been challenged a bit by the reverses in the telecommunication industry in 2001 and 2002, but even a resumption of this growth does not necessarily mean an end to N-ISDN BRI service demand. To begin with, both conventional ISDN BRI service and IDSL direct Internet access service using the same line signaling formats can extend to greater distances from the CO than other DSL techniques. This means that ISDN is the only way for some wire-line access customers to gain even modest digital communication rates.

For small commercial offices, ISDN service for local area network routers that can automatically dial central and home office sites have provided a cost-effective alternative to dedicated private lines or other services. In some circumstances, the public network connection flexibility of ISDN is the right solution. In addition, most local and regional ISPs offer ISDN termination options for subscribers to permit direct access to the Internet. For now, this flexibility may be difficult to obtain through other services.

Even though ATM service is a relatively small proportion of the total dedicated data service market, its broadband character and its use internal to networks (as opposed to premises access) promises to position it well for the next few years. Revenue from these applications is expected to grow at a cumulative annual rate of 34% (E. Dunne, personal communication, 2002). Even some of the services competing with N-ISDN, like ADSL, use ATM core network transport to link CO access line terminations to Internet access points. The legacy of B-ISDN lives on in this way.

CONCLUSION AND THE FUTURE OF ISDN

In some sense, one can argue that ISDN has both enabled the Internet and been overrun by it. The core and edge network technologies of ATM and frame relay are outgrowths of ISDN standards efforts and represent both a current and future portion of the transport solutions that connect people to the Internet. Call connection signaling standards for ISDN have been absorbed into at least one solution for converged services using a single, IP-based network. Finally, the comprehensive approach to defining connection management for services like telephony that are inherent to the ISDN standards represents a minimum threshold of excellence that all competing approaches must address.

It is equally clear that neither the narrowband nor the broadband ISDN perspectives for end user service are adequate for the indefinite future. Compared with purely packet-based alternatives, 64-kbps communications channels are simply too restrictive or wasteful. Although an early attempt was made to position ATM services to the general user premises and to individual desktops, this technology has found its place firmly in the core of communications networks. Furthermore, ATM and

B-ISDN are not the same thing; B-ISDN will never see deployment beyond the use of its link layer technology as it is currently applied.

The future of ISDN and its associated technologies is not in question. Eventually it will be displaced, application by application, by the technologies of the Internet. The question is, how long will this take? It is a sound wager that N-ISDN will survive at least as long as the PSTN and will slowly grow as the proportion of service in it. ATM and frame relay continue to have healthy growth rates (E. Dunne, personal communication, 2002) and present some advantages in places where IP-based networks continue to struggle (e.g., Quality of Service guarantees).

For those who would like to explore N-ISDN and B-ISDN and their derivative services in greater depth, there are several resources. Stallings (1999) provides a good, integrated view of all the topics covered in this article in much greater detail. Black (1997) provides a comprehensive view of ISDN and SS7, and Black and Waters (1997) consider SONET and T1 in detail. Miller (2001) considers the issues associated with the protocols in actual application.

GLOSSARY

American National Standards Institute (ANSI) A private, nonprofit organization that administers and coordinates the U.S. voluntary standardization and conformity assessment system. One particular role is to ensure that international standards are complete and appropriate for deployment in the United States and to sponsor additional standards development when necessary. A committee designated as T1 coordinates telecommunications standards. Standards from this committee are designated with the form T1.xxx, where the “xxx” is the specific standard number. Further information is available at http://www.ansi.org.

Asynchronous transfer mode (ATM) A form of physical layer information transport based on fixed length data cells with a 48-octet payload. ATM networks consist of a series of transport links connected through switches that do not retain bit or cell rate synchronism between input and output switch ports. End points in an ATM network are not able to synchronize directly with each other and must use embedded information or an external means to do this. ATM is the transport basis of broadband ISDN standards.

Basic Rate Interface (BRI) The narrowband ISDN user–network interface for individual subscribers. BRI provides two 64-kbps bearer (B) channels and one 16-kbps signaling and control (D) channel in each direction between the subscriber and the ISDN central office.

Broadband integrated service digital network (B-ISDN) The set of international standards that define both the network and operations to support end-user services with both uniform and irregular data rates at speeds up to 600 Mbps between the end user premises and the network. B-ISDN is the extension to (narrowband) ISDN services for broadband applications.

Digital subscriber line (DSL) A generic term that applies to a set of technological solutions for carrying voice and data over the (generally) two-wire subscriber line between a customer premises and the telephone central office. Narrowband ISDN is considered a member of the class and uses strictly digital techniques for transport. Asymmetric DSL (ADSL), another technique commonly available today, provides support for both existing analog telephone service and for simultaneous, broadband digital data service. The term xDSL is often used to reference a member of this class without specifying any one in particular.

International Telecommunications Union Telecommunications Standardization Sector (ITU-T) The ITU, headquartered in Geneva, Switzerland, is an international organization within the United Nations system in which governments and the private sector coordinate global telecom networks and services. The Telecommunications Sector is organized into study groups associated with different aspects of the field. ITU-T standards are designated by an initial letter that identifies a technology area and a three- or four-digit number designating the specific standard (e.g., Q.931). More information is available at http://www.itu.int

Link access protocol (for the) D channel (LAPD) A protocol used on the narrowband ISDN D channel for both BRI and PRI. It is based on high-level data link control protocol and provides the open system interconnection Layer 2 transport for signaling and control for N-ISDN.

Narrowband integrated services digital network (N-ISDN) The set of international standards that define digital, time division multiplexed access for subscribers to a uniform set of telecommunications services based on dedicated 64-kbps bearer channels (often called simply ISDN).

Primary rate interface (PRI) The narrowband ISDN interface definition typically used by enterprises to manage either 23 64-kbps bearer (B) channels and one 64-kbps signaling and control (D) channel (1.544 Mbps) or 30 B channels and one D channel (2.048 Mbps) between the customer premises and the ISDN central office.

Signaling System Number 7 (SS7) The set of international standards to provide out-of-band signaling for the global telecommunications network. In addition to supporting the signaling needs of traditional analog telephony and both digital and analog trunk technology, SS7 has a particular set of signaling capabilities designed especially to interconnect ISDN end offices and support services for ISDN and non-ISDN end points.

Synchronous transfer mode (STM) The time division multiplexed transport technique in which each communication channel on a link is allocated a particular time interval, called a time slot, in a strictly periodic information frame. Time division multiplexed switches that maintain synchronism between input and output ports interconnect links in an STM network. End points in an STM network can thus be directly synchronized with eachother.

CROSS REFERENCES

See *Public Networks; Wide Area and Metropolitan Area Networks.*

REFERENCES

ANSI Standard T1.110 (1992). Telecommunications Signaling System No. 7 (SS7)—General Information. New York: American National Standards Institute.

ANSI Standard T1.113 (2000). Telecommunications Signaling System No. 7 (SS7)—Integrated Services Digital Network (ISDN) User Part (ISUP). New York: American National Standards Institute.

ATM Forum Specificaiton af-phy-0046.000 (1996), 622.08 Mbps Physical layer Specification. Retrieved February 22, 2003, from http://www.atmforum.com

Black, U. D. (1997). ISDN and SS7: Architectures for digital signaling networks. Upper Saddle River, NJ: Prentice-Hall, Inc.

Black, U. D., & Waters, S. (1997). SONET and T1: Architectures for digital transport networks. Upper Saddle River, NJ: Prentice-Hall, Inc.

DuBoise, K., & Kim, H. S. (1992). An effective bit rate/table lookup based admission control algorithm for the ATM B-ISDN. In *Proceedings of the 17th IEEE Conference on Local Computer Networks.* New York: Institute of Electrical and Electronic Engineers.

International Telecommunications Union (2002). *Worldwide telecommunications indicator.* Geneva: International Telecommunications Union.

ITU-T Standard G.703 (1998). Physical/electrical characteristics of hierarchical digital interfaces. Geneva: International Telecommunications Union.

ITU-T Standard G.704 (1998). Synchronous frame structures used at 1544, 6312, 2048, 8448 and 44 736 kbit/s hierarchical levels. Geneva: International Telecommunications Union.

ITU-T Standard H.320 (1999). Narrow-band visual telephone systems and terminal equipment. Geneva: International Telecommunications Union.

ITU-T Standard H.323 (1999). Packet-based multimedia communications systems. Geneva: International Telecommunications Union.

ITU-T Standard I.113 (1997). Vocabulary of terms for broadband aspects of ISDN. Geneva: International Telecommunications Union.

ITU-T Standard I.120 (1993). Integrated services digital network (ISDN)—General structure. Geneva: International Telecommunications Union.

ITU-T Standard I.121 (1991). Integrated services digital network (ISDN) general structure and services capability—Broadband aspects on ISDN. Geneva: International Telecommunications Union.

ITU-T Standard I.430 (1995). Integrated services digital network (ISDN)—ISDN user-network interfaces. Geneva: International Telecommunications Union.

ITU-T Standard Q.921 (1997). ISDN user-network interface—Data link layer specification. Geneva: International Telecommunications Union.

ITU-T Standard Q.930 (1993). Digital subscriber signaling system No. 1 (DSS 1)—ISDN user-network interface layer 3—General aspects. Geneva: International Telecommunications Union.

ITU-T Standard Q.931 (1998). ISDN user-network interface layer 3 specification for basic call control. Geneva: International Telecommunications Union.

ITU-T Standard Q.932 (1998). Digital subscriber signalling system no. 1—Generic procedures for the control of ISDN supplementary services. Geneva: International Telecommunications Union.

ITU-T Standard Q.2931 (1995). Broadband integrated services digital network (B-ISDN) digital subscriber signaling system no. 2 (DSS 2) user-network interface (UNI) layer 3 specification for basic call/connection control. Geneva: International Telecommunications Union.

Miller, M. A. (2001). Analyzing broadband networks (3rd ed.). New York: McGraw–Hill.

Stallings, W. (1999). ISDN and broadband ISDN with frame relay and ATM (4th ed.). New Jersey: Simon & Schuster.

Telcordia (1995). National ISDN 1995 and 1996 (Special Report 3476). Morristown, NJ: Telcordia Technologies.

Intelligent Agents

Daniel Dajun Zeng, *University of Arizona*
Mark E. Nissen, *Naval Postgraduate School*

INTRODUCTION

This chapter is about intelligent agents and their applications in e-commerce. The term *intelligent agents* has many connotations. At one extreme, one can envision competent humans performing in some agent capacity (e.g., real estate, stock transfer), while at another extreme, the term conjures up images of sophisticated software that interfaces with their human users in a natural language and helps them manage various aspects of their personal and professional life (e.g., booking vacation packages and answering business e-mails). The kinds of intelligent agents we discuss in this chapter certainly fall in between these extremes, but articulating a precise, universally accepted definition is infeasible at this early stage of agent development. Nonetheless, drawing from the agent-research literature, we can outline a number of properties generally associated with intelligent agents. For instance, despite ambiguity arising from the term *agent,* we refer to intelligent agents solely in terms of software applications; that is, we exclude all references to people performing in some agent capacity. However, not all software constitutes agents, and not all agents are considered "intelligent."

To be considered an agent, researchers have proposed a number of characteristics, including autonomy, networked interaction, reactivity, and persistence. By "autonomy," we mean the software should be endowed with goals and the mechanisms to pursue such goals without sustained human intervention; some software application that needs to be told what to do each time it is invoked is unlikely to be characterized as an agent. Agents are generally viewed as interacting with other agents in some networked environment; some software that operates alone in some monolithic manner is unlikely to be characterized as an agent. By "reactivity," we mean the software has the capability to sense its environment and respond with different behaviors depending upon the kinds of environmental stimuli received; software that exhibits only a fixed set of behaviors is unlikely to be characterized as an agent. Agents are generally viewed as being persistent in their pursuit of goals and operating for long periods of time if necessary; some batch or user-interactive software that terminates immediately after completing some task is unlikely to be characterized as an agent.

To be considered "intelligent," agent software must exhibit behaviors similar to those that would be exhibited by a person, who is considered to be intelligent, in the same situation. Some well-accepted characteristics of intelligence from a software perspective include goal-oriented automated problem solving, integration of problem solving and execution of decisions, self-awareness, bounded rationality, adaptivity and the ability to learn, flexible exception handling and robustness in an open and distributed environment, proactivity in identifying goals to be accomplished, natural and easy interaction with human users, and long-livedness in a networked environment (Jennings, Sycara, & Wooldridge, 1998; Russell & Norvig 1995).

If these characterizations of an intelligent agent appear somewhat vague, it is because the field of agents research has yet to come to exact terms with aspects such as these. However, the characterizations above are well rooted in the literature (Weiss, 2000), and most researchers would be likely to identify with the majority of them.

Now that we have discussed what it means to be an intelligent agent, we address the e-commerce context of such agent applications. As described elsewhere in this book, recent developments in networking and Internet-related technologies are leading to the emergence of a digital economy. As with agents, many different definitions for e-commerce are possible. For purposes of this chapter, however, we adopt a fairly restrictive definition: e-commerce refers to commercial activities conducted over the Internet. In this sense, we use the terms e-commerce and Internet commerce interchangeably. Given the key role of networks in our characterization of agents above, with the convergence of e-commerce and Internet commerce in this view, it should be clear that intelligent agents represent inherently capable technological enablers.

However, like any software application, agent development consumes precious resources (e.g., development, infrastructure, maintenance), and despite their technological capability, in the commercial setting intelligent agents must also make economic sense. In some respects, this requires considerable scale, for agent development may not be economically viable for small (electronic) marketspaces or those in which the inherent capabilities of software (e.g., processing speed, memory capacity) do not convey sufficient advantage over the fundamental capabilities of people (e.g., flexibility, tolerance of ambiguity). In other words, for intelligent agents to be viable for e-commerce, their domains of application must have sizeable scales in terms of users, transactions, and information content. Otherwise, agents for e-commerce will not produce sufficient return on investment. Despite periodic shakeouts in the "dot-com" industry, even widely differing estimates (Lindberg, 2002) suggest global e-commerce will remain an area of continuing growth through the foreseeable future.

Additionally, we note several reasons why agent technology offers particularly high potential to drive further e-commerce growth. First, as software implementations, agents process information very quickly; hence, they can search through and filter huge information spaces (e.g., e-catalogs, potential customers) much faster than people can. As business, customer, product and service information is increasingly stored online and made Web-accessible, the potential of intelligent agents in this respect will become increasingly pronounced. Second, once a basic agent design has been developed and implemented via software, the capability of one such agent can be replicated by cloning and adapting other agents to perform similar functions; hence, a massively parallel system of intelligent agents can be established quickly and efficiently by replicating agents. Even where the "intelligence" of a particular software agent remains limited (e.g., in human terms), launching a thousand such agents to tackle a search or problem-solving task in parallel can effect great economies in terms of time and other resources.

As an example in the business-to-business (B2B) e-commerce domain, agents for a supplier could perennially monitor sales forecasts of its primary customers and use the results from such monitoring to develop aggregate production schedules. Based on such schedules, another set of agents could in turn interact with lower-tier suppliers and order the necessary parts and supplies, all automatically, without human intervention. Gauging from prototype supply chain agent systems such as the Intelligent Mall (Nissen, 2001), this latter agent scenario is not too distant in terms of realization. Other examples can be envisioned as well—some more futuristic than others—but the applicability and potential of intelligent agents to enable, automate, and enhance e-commerce is profound.

Other e-commerce applications include automation (e.g., performing tasks such as bank transfers without human intervention), decision support (e.g., setting up product comparisons based on consumer preferences; automatic bidding on online auctions), disintermediation (e.g., using agents to replace market intermediaries such as real estate agents), re-intermediation (e.g., using agents to implement new intermediaries in markets such as online information brokers), and others.

Following this Introduction, the balance of the chapter is organized as follows. The next section briefly reviews intelligent agent technology from a technical perspective. E-commerce Agents for Knowledge- and Information-Intensive Processes focuses on the use of agent technology in e-commerce to manage knowledge- and information-intensive processes, and Agent-Based E-commerce Decision-Making discusses agents that can be used to make or aid both operational and strategic e-commerce decisions. In Agent-Based E-commerce System Development and Evaluation, we revisit system development and integration issues related to building effective e-commerce agents, and Roads Ahead concludes the chapter with a summary and look at roads ahead for agent applications in e- commerce.

INTELLIGENT AGENT TECHNOLOGY

The purpose of this section is not to provide a comprehensive survey of agent technology. Such surveys already exist in the agent literature (Jennings, Sycara, & Woodridge, 1998; Weiss, 2000). Rather, we intend to provide a brief overview of various technical aspects of agent technology that have particular relevance to e-commerce applications and e-commerce system development. Whenever appropriate, we suggest further readings that provide detailed in-depth technical discussions.

Intelligent agent technology is highly interdisciplinary and has benefited from a wide spectrum of academic disciplines, including but not limited to philosophy, logic, linguistics, ecology, economics, operations research, distributed systems, computer programming languages, software engineering, and computer networks (Weiss, 2000). However, the main throttle behind the recent rapid development of intelligent agent technology comes from ideas and innovative research originated by the artificial intelligence researchers (Russell & Norvig, 1995).

Here, we briefly summarize the main connections between intelligent agent technology and its related (supporting) disciplines by considering artificial intelligence (AI), emergent computation, distributed systems, economics, and game theory. AI provides the basic automated reasoning capabilities for an agent to be "intelligent." In particular, AI planning techniques, motivated by the needs to achieve intelligent goal-driven behavior through automated construction of an action sequence, are being heavily used in developing agents. In addition, AI equips agents with a logic-based representation framework to model the environment and behaviors of agents.

Recent work in emergent computation is interwoven with intelligent agent research. The basic idea behind emergent computation is that interactions of some simple behaviors in a localized context can lead to sophisticated system-wide behaviors and desirable outcomes. Some best-known examples of emergent computation are genetic algorithms, artificial life, and market-oriented programming (Wellman, 1995).

Regarding distributed systems, many issues occur when a system is composed of a distributed collection of entities and a centralized control mechanism is absent. Distributed systems research provides analytical tools and coordination protocols that facilitate the design and implementation of multi-agent systems.

Finally, a significant portion of economics and game theory studies noncooperative interactions between self-interested decision-making entities (Fudenberg & Tirole, 1991). Many agent-based systems, in particular multiagent systems that consist of a distributed collection of software agents, typically involve agents that represent independent parties that do not necessarily share a common goal. Consider, for instance in a supply chain setting, the buyer's agent and the seller's agent represent their own respective owners whose preferences are in direct conflict when negotiating prices. The models and market mechanisms (as distributed resource allocation mechanisms) developed by economists and game theorists can guide the design of agent-based systems (Wellman, 1995). In effect, the relationship between economics, game theory, and agent technology can be fairly complex and goes in both directions. For instance, intelligent agent technology is also being viewed as a major enabling technology that can reduce economic transaction costs and make various types of markets more efficient.

Because of the highly interdisciplinary nature of intelligent agent technology, it is not surprising to see much confusion concerning the agent-related terminology and many different views on what agent technology can offer. Next, we summarize some of the widely accepted views particularly relevant to e-commerce applications.

Agents as a System Analysis Tool

Agents can be used as a system analysis tool in the conceptual design phase of a software development project. They are particularly useful for applications that involve a distributed set of autonomous decision makers/execution units. In an agent-based system analysis approach, autonomous entities are modeled as agents. Interactions among these entities are modeled as interagent communications in the syntactical level, and as agent coordination protocols in the semantic and application level. Depending on the nature and complexity of the target application, the inner structure of an agent can be modeled recursively as another team of agents or an atomic decision-making unit or data source. Compared to traditional system analysis tools, this agent-based approach has many advantages such as providing a natural mapping of the distributed nature of data and control; providing a powerful abstraction and encapsulation mechanism to model complex and dynamic behavior; and facilitating modeling of scenarios that involve parties with conflicting goals.

Agents as a Programming Paradigm

Agents represent a natural evolution from object-oriented programming (OOP) as a programming paradigm (Shoham, 1993). Software engineering research has recognized that interaction among software modules is the most important aspect of complex software systems. Agents provide a richer, more independent software building block than OOP to achieve better data independence, control independence, interaction independence, behavior encapsulation, and software reuse. Agents can be particularly useful in developing interorganizational applications where decisions on various aspects of software development such as the computing platform and programming environment are made independently. Agents also operate in a robust manner in "open" computing environments, which are characterized by interaction between diverse computational artifacts (e.g., computer hardware, operating systems, agent designs) and dynamic agent participation. This highlights a particularly challenging aspect of an open environment, in that agents might join and leave some virtual environment (e.g., marketspace) or organization (e.g., supply chain federation) repeatedly and without notice.

Agents as a Distributed Decision-Making Paradigm

In many applications, decisions must be made in a distributed manner. For instance, in a supplier/buyer negotiation case, the selling price of the product under negotiation must be jointly decided by all parties involved. These parties are self-interested and have conflicting goals. In situations like this which involve strategic interactions, a distributed decision-making paradigm is a must, and agents provide a convenient modeling alternative that is flexible and extensible.

In another example of a more collaborative nature, an online retailer is making inventory decisions for its regional warehouses. These inventory decisions can be potentially made by a centralized decision-making mechanism or jointly by all regional warehouses. Although this type of decision-making does not involve self-interestedness, a distributed paradigm still offers many potential advantages over its centralized counterpart. For instance, a centralized approach needs access to centrally maintained data sources that consolidate data from all regional centers, whereas a distributed approach does not have this data requirement. In addition, from a pure computational standpoint, an agent-based, distributed decision-making model can solve satisfactorily (in a heuristic sense) some problems that are too large for centralized approaches (e.g., Liu & Sycara, 1997).

E-COMMERCE AGENTS FOR MANAGING KNOWLEDGE- AND INFORMATION-INTENSIVE PROCESSES

In this section, we discuss agents for information and process management, address key technical issues associated with such agents, and feature one real-world implementation of e-commerce agents developed specifically for managing knowledge- and information-intensive processes. Each of these three topics is discussed in turn.

E-commerce Agents for Information and Process Management

Despite inexpensive electronic access to information, effective use of the Internet by people and decision support systems has been hampered by some dominant characteristics of the Web. First, information available from the Web is typically unorganized, multimodal, and distributed on servers all over the world; that is, the right information can be difficult to find. Second, the number and variety of data sources and services are increasing dramatically every day, and the availability, type, and reliability of information services are constantly changing; that is, once found, accessing and retrieving information can be challenging. Third, the same piece of information can be accessible from a variety of different information sources, but with varying degrees of "freshness"; hence, understanding how much confidence to place in specific information content requires judgment. Thus, information can be quite difficult for a person or machine to collect, filter, evaluate, and use, so the critical problem of locating, accessing, filtering, and integrating information in support of e-commerce has become a very challenging task.

Alternatively, intelligent agent technology provides a novel and promising approach to address this e-commerce challenge. For instance, e-commerce agents can act on behalf of their human users in order to perform laborious information gathering tasks. Examples include locating and accessing information from a wide diversity of online information sources, resolving inconsistencies in the retrieved information, filtering out irrelevant or unwanted information, integrating information from heterogeneous information sources, and even adapting over time to their human users' information needs (Sycara & Zeng, 1996).

Consider a specific application—comparative shopping—to illustrate how agent technology is being applied to address this e-commerce challenge. Recent years have seen a blossom of shopping agents that facilitate consumer and organizational purchasing activities (e.g., mysimon.com, shopper.com, bizrate.com). Most of these agents are hosted by third-party e-commerce information intermediaries. Given one or several user-provided keywords, these systems search through an internal, dynamically maintained database for potential vendors. Information items such as real-time price quotes, vendor ratings, product reviews, and shipping costs are then displayed to the user. Such e-commerce agents are providing increasingly impressive services to potential buyers, because they make sophisticated use of many underlying information technologies.

For instance, information technology has developed considerably in the area of knowledge representation and domain ontologies; that is, formally or heuristically defined domain knowledge enables shopping agents' intelligent behavior (e.g., resolving synonyms, tolerating user-input errors, suggesting relevant products and search terms). Another instance pertains to information retrieval and digital library technologies; that is, Web-based information collection, storage, and search capabilities have advanced greatly. Further, technologies associated with natural language understanding and content parsing have recently come of age in terms of e-commerce agent applications, as some agents apply sophisticated language understanding techniques to parse vendors' Web pages directly and extract fresh information on the fly. Additionally, machine learning and personalization technologies now enable some shopping agents to learn user preferences and present personalized interfaces and information search experience. In addition, collaborative information retrieval and recommendation technologies enable several shopping agents and recommendation systems to leverage information gleaned from prior user sessions when performing new tasks.

Technical Issues

As with any software development, technical issues associated with agents for e-commerce abound. However, such issues are particularly demanding in terms of intelligent agent development. One reason for this technical difficulty is that the agents field is comparatively young with respect to other software areas (e.g., transaction processing systems, database applications, client/server architectures). However, another reason stems from the power of intelligent agents and their focus on knowledge. Two key technical issues must be addressed: (1) how to represent process knowledge via agents, and (2) how to develop capable agent applications for managing knowledge- and information-intensive processes.

Knowledge representation (KR) has long attracted the attention of AI researchers, and a wide variety of KR tools and techniques have been developed over the past three decades (e.g., rules, frames, scripts, semantic nets). One aspect of KR for agents that leads to difficulty stems from our discussion above, in that agent applications are generally composed of many individual agents that operate together in some networked, organized manner. Thus, the representation of knowledge via agents must be distributed.

One approach to this distributed KR problem is the use of agent-development tools. Although specific tools come and go, the use of tools in general has proven effective in addressing the distributed-KR problem. For instance, Nissen (2001) employs Grafcets to define the structure and behavior of intelligent agents. Grafcets are derived from work on Petri Nets and have been accepted as an international standard (IEC 848 and IEC 1131–3) for specification of programmable logic controllers (David, 1995). Grafcets are used to outline the key steps and transitions associated with process performance, along with internal rules and procedures, external events and alternative system states that affect reasoning, decisions, and behaviors represented for a process. This technique facilitates mapping between the knowledge and activities required for effective enterprise performance, at the process level, and the behaviors of an agent application, at the technology level. As a method supporting abstraction and successive refinement, it facilitates agent design by allowing the developer to concentrate on high-level process knowledge and activities before diving into details of objects, messages, and code.

Nissen (2001) approaches the technical issue of agent development by using a research application called ADE (Nissen, 2000). ADE is a layered application that runs on top of G2, an environment for developing real-time intelligent systems (http://www.gensym.com). It provides integration between the activities and required knowledge captured from an enterprise, at the process level, and the technology-level transitions and steps that define agent behaviors and communications. These latter agent behaviors and communications are in turn implemented through objects, methods, messages, and rules; that is, below the agent level of abstraction, familiar object-oriented analysis, design and programming approaches and techniques apply. ADE provides its own environment for simulation and execution of agents, and it contains a translation mechanism for implementing agents in Java.

A number of other agent-development tools are also available, but as discussed in greater detail in the section Agent-Based E- commerce System Development and Evaluation, the relative value of agent-specific tools appears to be diminishing. Regardless of the specific tool (e.g., agent-specific, general) used, however, perhaps the greatest technical challenge associated with agent development is anticipating the emergent behavior of many agents operating simultaneously. Because of the properties from above generally ascribed to agents (e.g., autonomy, persistence), each individual agent is often developed to pursue its own goals (e.g., those specified by the agent's user/principal), and the aggregate of such goal-oriented behavior by many agents can be very difficult to anticipate during the agent-design stage of development.

For instance, in the Intelligent Mall discussed in the next section, agents for the same enterprise (e.g., different procurement agents of the same company) can find themselves competing with one another to purchase high-priority parts or supplies from a particular vendor, which can have an adverse impact of driving up prices. We label this behavior as "emergent," because it is not specifically designed into the agents; rather, it simply emerges through the interaction of networked agents working toward individual goals. Alternatively, if this emergent behavior can be anticipated (e.g., through networked-agent simulation), then more appropriate behaviors—such as agents combining their orders to negotiate price discounts—can be defined through Grafcets or whatever agent- development approach is being employed.

Feature: Real-World Agent Implementation

A number of different agent systems for managing knowledge- and information-intensive processes have been developed and implemented in the modern enterprise. For instance, Nissen (2000) developed an agent system to aid the design of business processes in a reengineering context. This process-design agent was implemented and then tested through laboratory experimentation, in which its performance surpassed that of people in terms of process-design efficacy and efficiency. As another instance, the Intelligent Mall (Nissen, 2001) was developed to automate and support the procurement and order fulfillment processes of the supply chain, and this agent system was later adapted to match people with jobs in an e-employment market. Further, the ADEPT system (Jennings, Norman, & Faratin, 1998) makes a good case for an agent-based approach in this domain, showing how agent technology can improve efficiency by ensuring that business activities are better scheduled, executed, monitored, and coordinated. Here, we briefly feature the Intelligent Mall, characterizing its role and operation in the context of knowledge-work required for performing supply chain processes. As an operating e-commerce application, its discussion provides a concrete exemplar of current agent technology applied to this area.

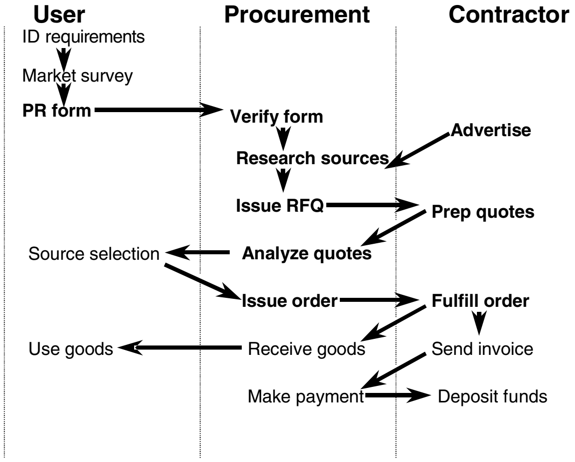


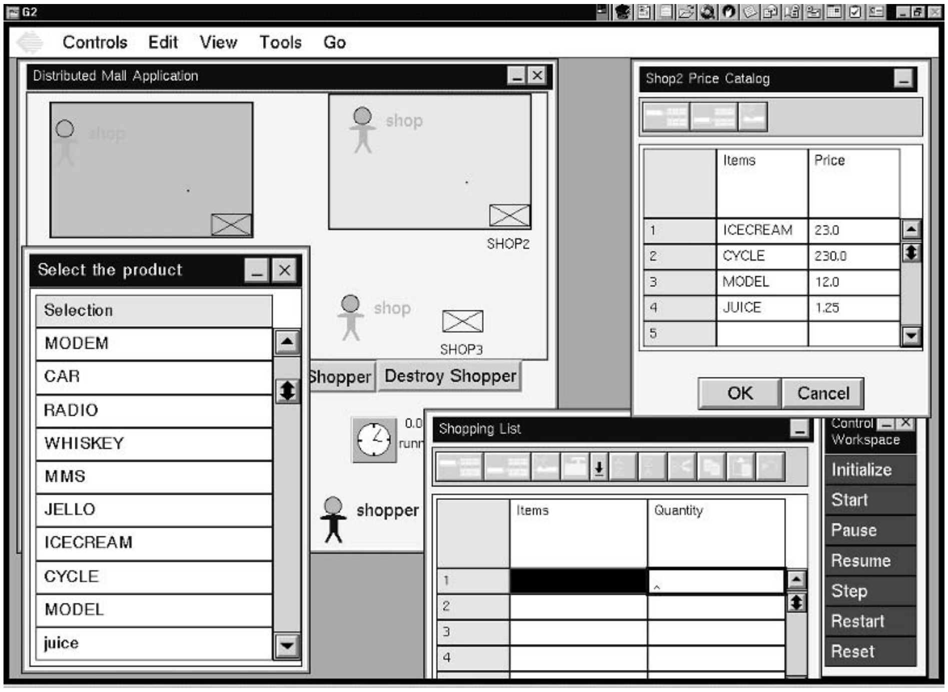
Figure 1: Integrated supply chain process.

The Intelligent Mall was developed to automate and support two complementary processes of the enterprise supply chain: procurement and order fulfillment. Specifically, the enterprise procurement process pertains to work done by the supply department at a medium-size government facility in the United States. The commercial order fulfillment process pertains to work done by a leading-edge U.S. technology development company.

To help understand the Mall application, the associated supply chain process activities and flow are presented in Figure 1. Note that this process involves three participants: the user, a procurement organization, and a vendor/contractor. This represents a representative link of the supply chain, as the same kinds of structures, relationships, and process flows exist between the vendor/contractor and its lower tier suppliers, between the latter's still-lower tier suppliers, and so forth to the end of the chain. Of the major process activities diagrammed in the figure, 16 activities (e.g., advertise items for sale, request vendor quotations, analyze quotes, pay for items) are automated through a small federation of agents that cooperate in a virtual mall environment.

In this environment, sellers establish agents to advertise and sell products, and buyers establish agents to search for and purchase products. Once the agents—both seller and buyer—have been launched, they operate autonomously and persistently until they either accomplish their goals (e.g., purchasing a specific item) or are halted by their users/principals (e.g., when an item is no longer offered for sale).

A screenshot of the Intelligent Mall interface is presented in Figure 2. This figure shows a number of animated agents for visualization as they perform procurement and order fulfillment process activities. For instance, the animated agents labeled "shopper" represent buyers in this virtual mall environment and seek out purchase items specified by their users/principals. Likewise, the animated agents labeled "shop" represent sellers and offer items for sale in the Mall. In terms of behavior, shop agents



Shop2 Price Catalog

	Items	Price
1	ICECREAM	23.0
2	CYCLE	230.0
3	MODEL	12.0
4	JUICE	1.25
5		

Select the product

Selection
MODEM
CAR
RADIO
WHISKEY
MMS
JELLO
ICECREAM
CYCLE
MODEL
juice

Shopping List

	Items	Quantity
1		
2		
3		
4		

Figure 2: Intelligent mall screenshot.

advertise items for sale (see example in upper right corner of the figure) that the Mall makes available through electronic Yellow Pages (i.e., what is for sale; see example on left side of the figure) and White Pages (i.e., what customers and vendors are participating in the Mall). Shopper agents take their lists of items to purchase (i.e., "shopping lists"; see example in lower right corner of the figure) and send messages requesting price quotations from all shop agents that offer such items for sale. When quotation requests are received, shop agents respond with individualized messages back to each shopper agent indicating the price (and any other pertinent terms) of each item. Each shopper agent then analyzes all quotations received from shops and selects the best buy (e.g., based on lowest price). Shopper and shop agents then pass several successive messages (e.g., with payment information, a purchase receipt) to one another to complete the purchase transaction.

Each shopping agent is capable of interacting with many different shop agents, and each shop agent can interact with many different shopper agents. Additionally, all shop and shopping agents behave autonomously and independently, and processes that require several hours—or even days and sometimes weeks—to perform through people in the enterprise can often be completed in a minute or less. Further, the Intelligent Mall environment is truly virtual, so the physical locations of various shops and shoppers are irrelevant; only a computer and Internet connection are required to launch agents that participate in the Mall.

The architecture of this agent federation is flat and distributed. It is flat in that all agents and hosts operate at the same, nonhierarchical level in the federation and on a network. It is distributed in that diverse agents in a federation can operate, communicate, collaborate, and move across various host machines and locations. Agents communicate and collaborate through messages in a peer-to-peer manner, but some communication support and like services (e.g., agent registry, Yellow Pages, message forwarding) are provided by a specialized class of agents across distributed hosts. These specialized agents help balance the low overhead achievable through a flat architecture with the need for organization and structure among agents as a distributed application scales to a large enterprise federation.

AGENT-BASED E-COMMERCE DECISION MAKING

At the present time, the use of agent technology in the realm of information and knowledge management dominates agent application in e-commerce. We believe, however, that the next wave of innovation will rely heavily on using software agents in making and executing e-commerce decisions in an automated or semi-automated (e.g., in the decision-aiding context) manner. Agent-based e-commerce decision making and automated execution can lead to significant improvement in decision quality and reduction in commerce transaction costs. This argument is essentially in line with the vision behind the Semantic Web, XML, and more recently Web services: the Web should be not only a medium through which human users can obtain and publish information but also a distributed, scalable, interoperable platform for software programs to exchange information and leverage each other's capabilities.

Many types of decisions need to be made in e-commerce practices. We use a simple criterion to divide them into two groups: operational and strategic decisions. By operational decisions, we mean the set of intraorganizational decisions that in principle can be made by a single decision-making entity. Examples of this type of decisions are coordinating an internal fleet of delivery trucks, deciding whether an order request made online should be accepted, and making a production schedule for a floor shop. By strategic decisions, we mean the decisions that cannot be made without explicit involvement from other self-interested parties. Obvious examples of decisions of the strategic nature include online auctions and negotiations where binding economic deals are made jointly by several active parties, each of which may have conflicting interests.

Agent-Based Operational Decision Making

Enterprise operational decisions are currently made largely in a centralized manner. Mathematical models and decision-making frameworks proposed in operations management, industrial engineering, and management information systems serve as the theoretical foundation for this type of centralized operational decision making. Existing enterprise-computing platforms such as ERP systems provide the corresponding IT infrastructure and execution environment.

Many facets of today's enterprise management in the e-commerce context, however, pose significant challenges to this centralized approach and promote a distributed framework to which agent technology contributes as both a system analysis and implementation methodology, as well as an effective computational paradigm. For instance, the modern enterprise comprises a network of distributed decision entities, both manual and automated. These entities usually have their own unique information channels, perspectives, and incentive structures. Furthermore, the supply chains and trading partnerships consist of many loosely connected companies and consumers with completely different objectives and constraints. In this case, it is unlikely that one can impose some common decision structures as necessary in a centralized data management and decision-making approach.

As another instance, it has been a clear trend in many industries that enterprises or production units focus on their respective core competencies (Kalakota & Robinson, 1999). They are agile organizations that may dynamically team up to form virtual enterprises to deliver value. In this type of setting, centralized mechanisms fail to capitalize on agility, flexibility, and quick response provided by this type of organization. Dynamic teaming and efficient coordination are also difficult to come forth in centralized frameworks.

From the viewpoint of implementing enterprise decision support systems, high modularity, incremental system development, and evolutionary deployment are usually preferred. Given the open nature of the environment, it is important for a system to expand and grow in various dimensions. These extensibility issues are difficult to resolve in traditional frameworks.

Further, the size of the decision problems that an enterprise system must face increases dramatically given the new development on the e-commerce front and growing interdependencies among supply chain partners. Scalability issues regarding the problem size are a major stumbling block for centralized systems based on traditional optimization formulation and computation.

Moreover, quick responsiveness is viewed as one of the key success factors for an enterprise due to the high degree of uncertainty and rapid advances of technology. It is difficult to develop a monolithic and responsive system given the size and the complexity of the problems and decision scenarios that a total enterprise system is designed to deal with.

An agent-based, distributed approach offers many advantages that alleviate the problems already mentioned. Each agent corresponds to a real entity, such as a decision-making unit, an activity, a resource, and so forth. This provides closer management of an entity in its situated environment and therefore facilitates potential real-time responsiveness. Each agent can make decisions based on its own strategies and its local situation. This provides modularity for constructing the system. As well, such an agent-based model provides the ability to decompose the system into appropriate levels of granularity, to disentangle the problem interdependencies, and to analyze how they interact with each other at each of these levels.

From a computational standpoint, an agent-based approach promotes the idea of "emergent" computation. Instead of attacking a problem directly in its entirety from first principles, an agent-based approach first uses an appropriate decomposition scheme that divides various aspects of the original problem into relatively independent components. This scheme uses different criteria (e.g., decision variables, constraints, resources, data access, control structure) to decompose the problem depending on its nature and characteristics. Each component is then assigned certain designed behaviors. The problem solving is in turn accomplished through iterated, structured interactions of these components according to the selected agent coordination protocols (Weiss, 2000). Effectiveness of agent-based decision making is largely decided by the appropriateness of problem decomposition and agent interaction design.

Significant progress has been made in the development of agent-based decision making for enterprise and e-commerce applications. Examples include distributed agent-based production scheduling and constraint satisfaction (Liu & Sycara, 1997), and agent-based supply chain management (Swaminathan, Smith, & Sadeh, 1998; Zeng & Sycara, 1999).

Agent-Based Strategic Decision Making

Agent-based systems that aim to facilitate practical strategic decision making are starting to emerge. Well-known examples include systems that provide strategic decision support in international trade negotiation, labor negotiation, and bargaining (Jennings, Sycara, & Wooldridge, 1998). Most of these current systems are programmed with rules-of-thumb distilled from intuition about good behavioral practice in human strategic interactions.

Positive conclusions regarding effectiveness of these systems have been reported in the literature. Nevertheless, almost all these evaluative studies were conducted via computer simulation with limited, rigid problem scenarios. Significant research efforts are called for to address issues relating to how to prevent one's agents from being badly exploited by aggressive new agents that have been programmed to take advantage of their weaknesses (Binmore & Vulkan, 1999). Another interesting and important research challenge in this area is to explore the dynamics of human/agent and agent/agent interactions.

In this section, we first briefly discuss the theoretical foundation for making e-commerce strategic decisions. We then summarize three common types of e-commerce strategic decision-making agents offering varying degrees of decision-making sophistication, modeling, and computational capabilities. We conclude this section by presenting several emerging research topics closely related to the development of these e-commerce agents. These topics are been actively pursued by researchers.

It is not surprising that the academic discipline most relevant to the study of e-commerce strategic decision-making is economics, and more particularly, microeconomics, a branch of economics focusing on the interaction of individual economic entities such as consumers and firms through well-defined market mechanisms. Although consumers and firms' commercial behaviors in the e-commerce realm may be significantly different from those in traditional, nonelectronic settings, underlying market mechanisms (i.e., the communication and transaction protocols governing selling and buying activities) remain largely the same in both electronic and nonelectronic settings. In addition, there are no fundamental changes in the end goals of market participants' involvement (e.g., profit and utility maximization). Thus, the basic analytical framework developed in microeconomics is still largely applicable to e-commerce. This framework includes expected utility theory, risk preferences, mechanism design, Nash equilibrium and its refinements, and their applications in various types of markets including auctions among others (Fudenberg & Tirole, 1991; Pratt, 1964). We caution the reader, however, that although we argue that microeconomics provides an underlying modeling framework and a wide choice of relevant models, we are not stating that e-commerce strategic decision-making is a solved problem. On the contrary, it poses many new challenging problems of practical significance to a number of related fields including economics, and in recent years has motivated fruitful collaboration and synergetic research efforts among economists, artificial intelligence researchers, and applied mathematicians, some of which are discussed at the end of this section.

We choose auction theory to illustrate the applicability of microeconomics theory to e-commerce strategic decision making. This choice is made because not only auction theory is a leading, mature field of microeconomics but also online auction is currently the most important form of market mechanism used in e-commerce. For instance, most of online auctions are the ascending price, real-time action, called English auction, in which the bidder with the last highest bid buys the auctioned item at a price equal to his or her bid. Auction theory has a lot to say about how bidders should (or are predicted to) behave in various auction types including the English auction and what the expected seller payoff should be (McAfee & McMillan, 1987). A prevalent solution concept is the Bayesian-Nash equilibrium in which each bidder adopts a bid function that is the best strategic reply to the bid functions adopted by rival bidders. Following such a strategy, a bidder will bid "straightforwardly" in any English auction selling a single item: This bidder will outbid any standing bid that another bidder holds by the smallest allowable bid increment so long as the new bid does not exceed his or her valuation of the item for sale (also called reservation price). In effect, this strategy is a dominant strategy in the sense that it does not depend on how other bidders behave or the bidder's risk preferences. In other common types of auctions such as first-price sealed-bid auctions (every bidder submits his or her bid in a sealed envelop, and the highest bidder wins and pays the highest bid) and Dutch auctions (the auctioneer announces descending prices, and the first bidder buys the auctioned item at a price equal to the accepted price), which are being practiced mostly through traditional means rather than through an electronic marketplace, figuring out the optimal bidding function can be significantly harder. This is because in these auctions, optimal bidding hinges on information regarding other players, typically in the form of probabilistic distributions of their valuations as well as the bidder's risk preferences. However, in principle, the defining mathematical properties of optimal bidding are known and sometimes closed-form solutions can be derived under certain restrictive assumptions regarding value distributions and risk preferences. The reader is encouraged to read (McAfee & McMillan, 1987) for an excellent introduction to basic auction theory that can serve as a foundation for analyzing and designing online market mechanisms.

Agent technology can be applied to facilitate strategic decision making on online markets in several ways. From an implementation standpoint, agents can be used to automatically collect information necessary to make informed decisions (e.g., comparative shopping agents as discussed under E-commerce Agents for Knowledge- and Information-Intensive Processes and auction information and consolidation agents such as http://www.auctiontamer.com). From a computational perspective, agents can be used to "crunch" the numbers as significant computational support is needed to solve for some of the auction-theoretic (and in general game-theoretic) solution concepts relevant to e-commerce. From the point of view of transaction, agents can also be a major technology enabler in automating many aspects of e-commerce transactions, freeing people's time and reducing related information and cognitive overload. Last but not least, from an evaluation standpoint, agents can be used as part of the simulation and experimental framework to evaluate at both individual and overall societal levels the effectiveness and efficiency of the bidding strategies under study.

We now discuss three common types of e-commerce strategic decision-making agents:

1. *Simple execution agents*. This type of agent does not have an explicit model of the underlying strategic

decision task. Nor does it provide active decision support. The sole functionality of a simple execution agent is to automatically execute its human user's trading strategies, which are fully specified by the user at the beginning of trading through a limited number of parameters. Popular proxy bidding agents (e.g., the eBay proxy bidder) that function on online markets operated under English auctions fall into this category. These agents are initialized with their user's reservation price. Then without user intervention, they will monitor the progression of an auction and place a bid when the current highest bid is below the prespecified reservation price and is held by another agent. For other types of online markets, such as bargaining, some simple analogous strategy exists where a proxy negotiation agent submits its price proposals based upon a simple function specified by the user at the outset of the negotiating process (e.g., Kasbah; see Chavez & Maes, 1996).

2. *Behavioral agents.* This type of agent possesses explicit knowledge of the e-commerce trading institutions and provides active decision support. Computational models of auctions and negotiations from the economics and AI literature (e.g., the auction-theoretic ones already mentioned) serve as the basis for constructing such agents. To be computationally feasible, these models typically ignore certain aspects of strategic interaction. Examples of the models that have been utilized to develop behavioral agents are optimization-based strategies for competitive bidding that assume known distributions of values and the number of bidders (McAfee & McMillan, 1987), AI logic and rule-based negotiation strategies (Kraus, Sycara, & Evanchik, 1998), automata play, and adaptive auction and bargaining strategies using machine learning techniques (Wellman & Hu, 1998).
3. *Agents with human-like properties.* All the decision-making models already mentioned assume that the sole purpose of the agent is to maximize its user's utility or profit. In commercial and other social activities conducted through either traditional or online channels, however, rich patterns of interaction among participants cannot be fully captured by this utility-maximizing, self- regarding assumption. This prompts recent studies aimed at modeling other regarded properties and developing agents with such human-like properties. For instance, theoretical models proposed by Cox and Deck (2001) readily serve as the foundation for programming the behaviors of agents that demonstrate properties such as seeking fairness and being cooperative. Recent work has also started to equip such agents with human-like appearance and communication capabilities such as natural language input/output and facial expressions.

We conclude this section by pointing the reader to two ongoing lines of research driving the current development of strategic decision-making agents for online markets.

Purchasing bundles of goods is much more common than purchasing one unit of a single item in real-world applications, especially in industrial procurement. Different mechanisms have been proposed and studied in the literature. For instance, there has been a substantial interest in a particular class of auctions called combinatorial auctions that allow the bidders to submit bids for a bundle of goods as opposed to single items. Selecting a winning bid from these bundle offers turns out to be nontrivial. Significant effort from researchers in a number of related fields including economics, operations research, and multiagent systems has been invested to study combinatorial auctions in recent years (Sandholm, 2002). Another mechanism called the simultaneous ascending auction has also been proposed to deal with bundle purchases and shown to have a number of desirable properties (Milgrom, 2000). In effect, the simultaneous ascending auction was successfully applied to sell licenses to user bands of radio spectrum in the United States.

The third approach is a "meta-auction" based on a collection of simultaneous single-item auctions. In this approach, each auction is conducted according to its own rules and run independently from one another. Bidders who may need to purchase a bundle are responsible for coordinating their own bidding activities across these independent markets to achieve the best overall profit. Since there is no theory prescribing bidding strategies in this heterogeneous market environment, heuristic or expert system-based approaches are being actively developed. These approaches are agent-based e-commerce strategic decision-making systems in the strong sense. One significant issue with developing such agents is evaluation: How would different strategies fare against each other in this meta-auction context? In response to this evaluation issue, the multiagent systems research community has been organizing an annual tournament called "Trading Agent Competition." In this tournament, agents, designed and implemented by different research teams using a common programming interface, function like a human travel agent and attempt to purchase travel packages that include airline tickets, hotels, and entertainment (baseball, ballet) tickets, sold through independent auction markets. These agents are matched against each other in the profits they are making by finding the right packages for a given set of customers with specific (randomly generated) travel needs (Greenwald & Stone, 2001).

Another important research direction concerns the multidimensional aspect of the goods or services being auctioned or sold through other markets. By multidimensional, we refer to product characteristics besides price, such as delivery method and quality among others. It has been shown that multidimensional auctions, among other multidimensional markets, are hard to design in principle. Computing bidding strategies for such markets is also particularly hard, making an agent-based approach a viable modeling, computational, and implementional alternative (Zheng, 2000).

AGENT-BASED E-COMMERCE SYSTEM DEVELOPMENT AND EVALUATION

In this section, we discuss system architecture and development issues and evaluation of agent-based e-commerce systems.

System Architecture and Development Issues

In the early years of agent technology, software development tools and communication protocols specifically designed for agents received a lot of attention. For instance, KQML, which was developed prior to the Web protocols, coupled with KIF, had been accepted as one of the de facto agent communication languages. Integrated agent development environments had also attracted considerable attention from both the research community and industry.

In recent years, however, the use of agent-specific development tools and protocols has been waning. Alternatively, generic software development environments, especially those that have been motivated for Web-based applications, are being increasingly used to develop agent-based systems. We suspect the following reasons contribute to this interesting phenomenon.

First, due to lack of industry support, none of the agent-specific tools and protocols has become sufficiently stable and feature-rich to be adopted widely. Second, more and more applications are starting to adopt some aspects of agent-based computing ideas, although they remain far from being full-fledged intelligent agents. Using an environment motivated for the development of sophisticated agents is an overkill and can cause serious system maintenance and integration problems. Third, because of the significant academic and commercial interest in distributed Web-based computing, mainstream programming environments are now equipped with sophisticated support for knowledge representation and manipulation (such as XML), messaging, mobility and basic distributed coordination mechanisms (such as JavaSpace and Web services). Fourth and last, as agent ideas are being applied in highly specialized decision making, it is unlikely that any programming environment will provide readily reusable software modules. We envision that this trend of relying on generic programming platforms and toolsets will continue, and the role of agents as a system design methodology and a distributed decision-making framework will increasingly dominate the role of agents as an implementation environment.

Evaluation of Agent-Based E-commerce Systems

Existing research on agent-based e-commerce systems remains largely in the technological realm. Evaluation of such systems is also confined to a large extent in the algorithmic and technical contexts. For instance, in the Trading Agent Competition discussed previously, performance and efficiency of trading agents employing different meta-auction algorithms are evaluated through a simulation-based tournament.

As these systems are maturing and being accepted increasingly in practice, issues beyond technology development such as adoption and evaluation are also becoming more pressing and warrant significant research efforts. In this section, we emphasize evaluation of agent-based e-commerce systems in light of human–agent interactions as opposed to agent–agent interactions, which have been extensively discussed in the multiagent systems literature (Weiss, 2000). We also ignore a large body of literature studying human bidding behavior in both laboratory environments and real-world online markets, since software agents do not play a pivotal role there.

Studies on the evaluation of comparative shopping system designs based on research methodologies ranging from marketing, consumer behavior, human—computer interaction, and industrial psychology are starting to emerge. Also, the literature on general technology adoption has significant relevance to such evaluative studies, not limited to comparative shopping. Historically, experimental economics has focused almost exclusively on human decision making. Smith (1982) explains the theory underlying the method of experimental economics and how microeconomics can be an experimental science. In the laboratory, the experimenter controls the environment (e.g., the costs to a seller and the values for a buyer). The experimenter also controls the institutions or the rules by which the human agents can take actions and send messages in the marketplace. Controlling these two components of the microeconomic system, we then observe the behavior of the human agents. By holding the environment constant in the laboratory but altering the institution, behavior may change such that the resulting market outcomes are markedly different.

One overlooked characteristic in agent-based e-commerce is the human-agent interaction and the mixing of human play with agent play. Direct human behavior may change in a nontrivial way if mixed with human-to-agent play on the same side or other side of the market (i.e., buyer or seller). Zeng, Sheng, and Wilson (2000) followed an experimental economics-based research methodology to evaluate agents for Web-enabled auctions. In their pilot study, they focused on proxy bidding agents, used by many auction sites such as eBay. Using these proxy agents, bidders specify the maximum bidding amount, which is kept private. The proxy agents then automatically bid as the auction proceeds, bidding only enough to outbid other bidders. Although these bidding agents sound very straightforward, the interactions between agents and their human users are intriguing.

To explore the behavior of markets with a proxy-bidding agent, Zeng et al. recruited 20 University of Arizona undergraduates to participate in a one-hour experiment using Java applet auction software that they developed. As is standard in laboratory auction experiments, the subjects were informed that if they purchased a unit for less than their value, they would earn the difference as profit, which would be converted into cash. One bidder in each market was first instructed to submit a single proxy-bidding agent to the market. Once submitted, the remaining three bidders were free to bid in real time for the next two minutes. The theoretical prediction is that the highest value bidder will win the auction, paying a price equal to the second highest value. In their experiments, however, they observed that the presence of a proxy bidder consistently leads to prices on average less than the second highest value. One possible explanation is that the presence of the proxy-bidding agent changes the nature of the auction (e.g., from an English auction into a last second first price auction). The implication of this behavior is that the auctions can be inefficient when a simple automatic execution mechanism is introduced. Although preliminary, this study argues strongly that deploying agent technology in e-commerce settings does not necessarily lead to more efficient and effective outcomes. Similar experiments have been conducted at the IBM T. J. Watson Research Center for other types of auctions more commonly seen in stock exchanges. In summary, agent behavior could inadvertently change the nature of the market mechanism and thus result in unexpected inefficiency. Careful examination, possibly through controlled experiments, is called for before the deployment of any agent-based e-commerce systems.

ROADS AHEAD

This chapter discusses the application of agent technology in e-commerce applications. We argue that agent technology has great potential in guiding e-commerce systems analysis, design, and implementation. We focus on specific applications of agent technology in two major types of e-commerce tasks: managing knowledge- and information-intensive processes, and operational and strategic decision making. Interested readers are encouraged to browse a comprehensive agent technology portal at http://agents.umbc.edu for a wide array of ongoing agent research projects and industrial applications including e-commerce.

Agent-based e-commerce systems are still in their early stage of development. Significant technological advances need to be made to make agent-based e-commerce a reality. In the area of managing knowledge- and information-intensive processes, knowledge representation challenges related to building complex product and service catalogues and resolving vocabulary ambiguities need to be addressed. Challenges also exist in the area of user interest solicitation and system personalization. In the area of decision making, issues such as supporting scalable, real-time operational decision making and providing computational support for strategic decision making (e.g., for coordinating bidding activities across markets as discussed under Agent-Based E-commerce Decision Making) call for new breakthroughs.

Agent-based e-commerce systems do not exist in isolation. Challenging system integration issues need to be addressed in interoperating agent-based and non-agent-based (such as transactional) parts of the system. Other significant technological challenges include, but are not limited to, how to architect and implement agent-based systems in a wireless and mobile-commerce setting; how to make the system scalable and robust in an open environment; how to address security and authentification issues (e.g., a shopping agent may possess all kinds of sensitive information including credit card numbers); and how to resolve the real-time issues (e.g., an auction site needs to decide how and when to end an auction in face of network delays). Nontechnological issues such as legal implications (e.g., when automated execution systems make deals on behalf of their organizational users), adoption, and possible organizational changes also warrant significant new research efforts. These are exciting times in terms of intelligent agents for e-commerce.

GLOSSARY

Agent-oriented programming A special case of object-oriented programming where computer programs are written as a collection of interacting agents, which in turn are programmed directly using generic constructs such as belief, desire, and intention.

Auction A market institution with an explicit set of rules governing selling and buying activities including the determination of trading prices, given the bids from buyers and/or sellers.

Combinatorial auction An auction that sells many assets or goods simultaneously by allowing buyers to bid for asset or good bundles.

Distributed artificial intelligence A subfield of artificial intelligence studying intelligent agents and multiagent systems.

Electronic commerce The application of information and communications technologies to facilitate all or some aspects of commerce-related activities, including but not limited to online marketing and trading, products and services delivered through electronic means, and business transaction automation.

Electronic marketplace A technology-enabled virtual market through which sellers and buyers share information, trade, and conduct other business transactions. Internet and other communications technologies such as wireless technology are the main driving forces behind the development of the electronic marketplace. Examples include online auction sites.

Emergent computation A distributed computing paradigm where solutions to the problem under study emerge as a side effect of multiple autonomous computational entities interacting with one another following certain social norms.

Experimental economics A branch of economics studying the principles of using experimental methods (including both laboratory and field experiments) in economics and testing the validity of underlying economic theories. A significant portion of experimental economics focuses on the study of different forms of market mechanisms including auctions.

Intelligent agent A computational entity that is capable of autonomous actions, manifests nontrivial behavior, is self-aware, and can improve its performance through experience.

Knowledge representation The art and science of making knowledge explicit and amenable to automated inference through computational models.

Multiagent system A distributed computational system consisting of multiple autonomous agents that operate asynchronously, interact with one another through well-defined protocols, and may represent conflicting interests.

Process management The activities associated with planning, organizing, and monitoring systems of recurring processes composed of a set of interdependent tasks in the enterprise.

Proxy bidding A simple automated bidding service available from many online auction sites. After receiving a user-specified maximum willingness to pay, the proxy bidding system monitors the ongoing auction session and automatically increases bids (if needed) till the maximum willingness to pay is reached.

Supply chain management Coordinating enterprise activities concerning order generation, order taking, order fulfillment, procurement, and distribution of products, services, and information, across organizational boundaries.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Online Auctions; Rule-Based and Expert Systems; Supply Chain Management*.

REFERENCES

Binmore, K., & Vulkan, N. (1999). Applying game theory to automated negotiation. *Netnomics, 1*(1), 1–9.

Chavez, A. & Maes, P. (1996). Kasbah: An agent marketplace for buying and selling goods. In *Proceedings of the First International Conference on the Practical Application of Intelligent Agent and Multi-agent Technology.* (pp. 75–90). London, UK: Practical Application Company.

Cox, J., & Deck, C. (2001). On the nature of reciprocal motives. Discussion paper, Department of Economics, University of Arizona.

David, R. (1995, September). Grafcet: A powerful toll for specification of logic controllers. *IEEE Transactions on Control Systems Technology, 3*(3), 253–268.

Fudenberg, D., & Tirole, J. (1991). *Game theory.* Cambridge, MA: MIT Press.

Greenwald, A. R., & Stone, P. (2001). Autonomous bidding agents in the trading agent competition. *IEEE Internet Computing, 5*(2).

Jennings, N., Norman, T. J., & Faratin, P. (1998). ADEPT: An agent-based approach to business process management. *ACM SIGMOD Record, 27*(4), 32–39.

Jennings, N., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-agent Systems, 1*(1), 7–38.

Kalakota, R., & Robinson, M. (1999). *E-business—Roadmap for success.* Reading, MA: Addison Wesley.

Kraus, S., Sycara, K., & Evanchik, A. (1998). Argumentation in negotiation: A formal model and implementation. *Artificial Intelligence, 104*(1–2), 1–69.

Lindberg, B. C. (2002). The growth of e-commerce. Retrieved from http://home.earthlink.net/~lindberg_b/GECGrwth.htm

Liu, J., & Sycara, K. (1997). Coordination of multiple agents for production management. *Annals of Operations Research, 75.*

McAfee, R. P., & McMillan, J. (1987). Auctions and bidding. *Journal of Economic Literature, 25*(2), 699–738.

Milgrom, P. (2000). Putting auction theory to work: The simultaneous ascending auction. *Journal of Political Economy, 108.*

Nissen, M. E. (2000, Winter). An experiment to assess the performance of a redesign knowledge system. *Journal of Management Information Systems, 17*(3), 25–44.

Nissen, M. E. (2001). Agent-based supply chain integration. *Journal of Information Technology Management, 2*(3), 289–312.

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica, 32,* 122–136.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach.* Englewood Cliffs, NJ: Prentice Hall.

Sandholm, T. (2002). Algorithm for optimal winner determination in combinatorial auctions. *Artificial Intelligence, 135,* 1–54.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence, 60*(1), 51–92.

Smith, V. (1982). Microeconomic systems as an experimental science. *American Economic Review, 72,* 923–955.

Swaminathan, J., Smith, S., & Sadeh, N. (1998). Modeling supply chain dynamics: A multiagent approach. *Decision Sciences, 29*(3).

Sycara, K., & Zeng, D. (1996). Coordination of multiple intelligent software agents. *International Journal of Cooperative Information Systems, 5*(2&3), 181–211.

Weiss, G. (Ed.). (2000). *Multiagent systems: A modern approach to distributed artificial intelligence.* Cambridge, MA: MIT Press.

Wellman, M. (1995). The economic approach to artificial intelligence. *ACM Computing Surveys Symposium on Artificial Intelligence, 27*(3).

Wellman, M., & Hu, J. (1998). Conjectural equilibrium in multiagent learning. *Machine Learning, 33,* 179–200.

Zeng, D., Sheng, O., & Wilson, B. (2000). The design and experimentation of agent-based procurement systems. In *Proceedings of the Third International Conference on Telecommunications and Electronic Commerce.* (pp. 364–375). Dallas, TX: The Cox School of Business, Southern Methodist University.

Zeng, D., & Sycara, K. (1999). Dynamic supply chain structuring for electronic commerce among agents. In M. Klusch (Ed.), *Intelligent information agents.* New York: Springer.

Zheng, C. (2000). *Optimal multidimensional auctions.* Discussion paper, Northwestern University.

Interactive Multimedia on the Web

Borko Furht, *Florida Atlantic University*
Oge Marques, *Florida Atlantic University*

INTRODUCTION

Personal computing, multimedia, and computer networks have experienced a tremendous boom during the past decade. From expensive, stand-alone MPC-1-compatible PCs that landed in the marketplace in the early 1990s to the popularization of the Internet, particularly the World Wide Web, the production, archiving, and distribution of multimedia has evolved from a few, technologically limited devices to a plethora of hardware and software available for these tasks. It is now possible to develop higher quality multimedia content, in less time, and distribute it to a much wider audience (using the Web as opposed to CD-ROMs) than anyone could have imagined a decade ago.

The Internet has become ubiquitous and is quickly extending its arms to the wireless arena. The hunger for multimedia-rich content over the Web has fostered developments in many fields, from the technological infrastructure necessary to carry multimedia data, to the tools used to create, manage, and distribute multimedia content. Interactivity is the key concept in many such applications, such as games, maps, and distance learning applications.

In the first part of the chapter, we define the concepts of multimedia and interactivity, particularly on a Web-based context, and present a summary of the network technologies that enable transmission of interactive multimedia information over the Web. In the next section, we show the process and related languages and tools for creating interactive multimedia Web pages. Finally, we present an overview of current interactive multimedia applications and services.

WHAT EXACTLY IS INTERACTIVE MULTIMEDIA?

We start by conceptually defining "multimedia" and "interactivity" in the context of Web-based applications. The term multimedia has been frequently used to mean the combining of text, graphics, images, sounds, and video on one page or presentation. According to this definition, most of the pages currently available on the Web are inherently multimedia. For the discussion that follows, it is important to know the following. (a) Multimedia content that combines only text, images, and animations (for which there is a built-in HTML support) is different from applications that use video clips, animations, and sounds (for which third-party products—usually available as browser plug-ins—are required). A multimedia-rich HTML page consisting exclusively of text, JPEG images, and animated GIFs would belong to the first category, whereas a page that loads a Macromedia Flash movie would belong to the second group. (b) Multimedia information that is delivered without user interaction is different from multimedia contents that change dynamically in response to the user's inputs. The streaming of a prerecorded video clip would be classified under the first group, whereas the constant updating of a screen in a multiplayer Web game would qualify as an example of the second case.

MULTIMEDIA NETWORK TECHNOLOGIES

There are five main types of communication network technologies used to carry multimedia traffic (Halsall, 2001): telephone networks, data networks (including the Internet), broadcast television networks, integrated services digital networks (ISDN), and broadband multiservice networks, such as B-ISDN. In this chapter, we focus mostly on the Internet. Besides enabling a number of interpersonal communication tools, such as e-mail, the Internet also hosts a large number of interactive multimedia applications stored in Web servers.

Characteristics of Multimedia Traffic

Interactive multimedia may be real-time or non-real-time information. Real-time, or *continuous,* information is transferred from source to destination as it is generated and is played out at the destination as it is received. This mode of operation is commonly known as *streaming*. Audio and video streaming over the Web have become very popular over the past few years. Examples include radio and TV broadcasting over the Web. In non-real-time, or block-mode, media, the information is generated and stored in a way that is not time sensitive and is downloaded by request by interested users.

Multimedia traffic can be classified into two main types: *constant bit rate* (CBR) and *variable bit rate* (VBR). Audio streams, for example, are typically generated at a constant rate, determined by the sampling frequency and number of bits per sample used to convert the audio information into digital format. Video streams, on the other hand, exhibit VBR, even though the frames that make up a video program are generated and displayed at a constant rate. This is due—among other things—to the fact that most video compression algorithms use differential techniques that employ fewer bits on frames that resemble previous frames.

The transfer of multimedia information streams between any source and destination of can occur in a number of different ways.

Simplex: Information flows in one direction only.

Half-duplex: Information flows alternately in both directions.

Full-duplex: Information flows in both directions simultaneously.

Unicast: Information flows from the source to a single destination.

Multicast: Information flows from the source to multiple destinations, comprising a subset of the nodes connected to the network.

Broadcast: Information flows from the source to all destinations connected to the network.

In the case of half- or full-duplex communications, the flow of information can be *symmetric,* where each party sends a comparable amount of information to the other, or *asymmetric,* where the flow of information is much higher in one direction than in the other.

The Internet uses packet-switching techniques to transfer information between two end computers using intermediate nodes—routers—along the way. This mode of operation is also sometimes referred to as *store-and-forward,* indicating that there is a built-in delay in the process that occurs every time a packet arrives at a router and waits for the router's decision (upon looking at its routing tables) as to where the packet must be forwarded next. The sum of the store-and-forward delays in each router contributes to the overall end-to-end delay between source and destination. The average of this delay is known as *mean packet transfer delay* and the variance is usually known as *delay variation* or *jitter*. The inevitability of these delays is one of the technical challenges behind enabling interactive multimedia applications on the Web.

Besides the delay limitations, the Internet has another characteristic that makes for less-than-ideal conditions for carrying multimedia traffic: It offers a best-effort service to its end users, which means that the TCP/IP-based transfer of packets is subject to errors (damaged packets, lost packets, timeouts, etc.), and the typical way to handle such errors is to discard packets and (optionally) request/wait for retransmission. Because much of multimedia traffic may be sensitive, retransmissions are usually not desired, because they add up to traffic inside the network and also because by the time the retransmitted packet arrives, it is no longer needed. A common way of handling with lost/retransmitted packets while providing the user with the illusion of continuous flow of information is the use of buffers at the receiving side.

Designers of networked multimedia applications usually specify the minimum network requirements that must be met to enable the application to work. These requirements are collectively known as network *Quality of Service* (*QoS*) parameters. Examples of QoS parameters are maximum packet size, average bit rate, average packet error rate (PER), mean packet transfer delay, worst-case jitter, and transmission delay. In addition to the network QoS parameters, the application itself has QoS parameters associated with it, such as the required bit rate or mean packet transfer rate, the maximum start-up delay, the maximum end-to-end delay, the maximum jitter, and the maximum round-trip delay. To simplify the process of determining whether a network meets the QoS requirements of a particular application, several *service classes* have been defined, each of which contains a specific set of QoS parameters. A network may support two or more service classes, in which case different classes are usually assigned different priorities.

Standards for Multimedia Communications

Multimedia information is typically transmitted in compressed form over the Web. Many standards have recently been developed for image, audio, and video compression and transmission. The most important standards are as follows.

MPEG-1: The first widely deployed standard proposed by the Motion Pictures Expert Group (MPEG). It aims at audiovisual encoding at a bit rate of 1.5 Mbps, which allows for VHS-like video with stereo audio to be encoded and which is stored in a CD-ROM. The best-known contribution of the MPEG-1 standard is its layer III audio compression, popularly known as MP3, which caused an entire revolution in the music industry.

MPEG-2: Enables the transmission of audio and video over broadband networks. It aims at encoding digital TV (and HDTV) signals and has been widely used in the consumer electronics field, from delivery of TV programming via personal satellite (dish) systems, to DVDs, to personal digital video recorders such as TiVo.

MPEG-4: Follows an object-based representation approach to audiovisual scenes, where each scene is composed of objects, both natural and synthetic. MPEG-4-based systems are not yet fully deployed at a consumer level at the time of this writing. It is expected, though, that the object-based nature of the standard will significantly enhance interactive multimedia applications.

MPEG-7: Standard that describes multimedia content, enabling users to search, browse, and retrieve multimedia

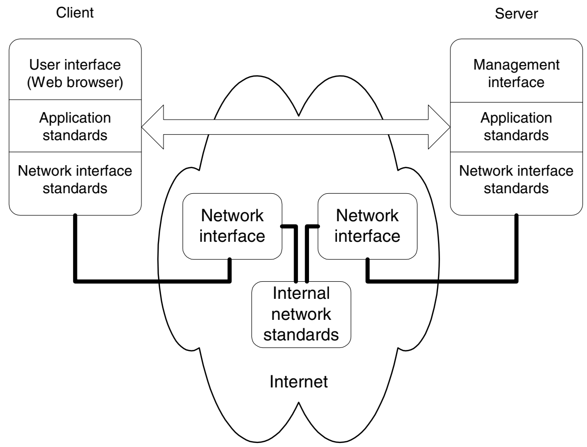


Figure 1: Standards requirements for interactive multimedia applications over the Internet.

contents more effectively than current text-based methods allow.

MPEG-21: The latest effort of the group focuses on providing a multimedia framework that should enable transparent use of multimedia resources across a wide range of networks and devices.

JPEG and JPEG 2000: The JPEG (Joint Photographic Experts Group) compression standard for still images was developed in the late 1980s and has become tremendously popular. JPEG 2000 is a new standard that uses Wavelet subband coding techniques and allows for numerous compression options. It has recently been approved and is starting to enjoy some popularity among Internet imaging users.

In the context of interactive multimedia applications on the Web (Figure 1), standards are required both at the application and at the networking level. Examples of application-level standards include the JPEG and MPEG compression standards and the hypertext transfer protocol (HTTP) implemented by all Web browsers to allow HTML pages to be displayed at the client's side. Network-level standards include the TCP/IP protocol suite.

CREATING INTERACTIVE MULTIMEDIA WEB PAGES

Contemporary tools for designing Web pages (such as Macromedia Dreamweaver and Microsoft FrontPage) allow the inclusion of multimedia objects, such as videos, images, animation, and audio, in Web pages. To play sound, animation, and video, the user system may require plug-ins, which are small programs that work in cooperation with a Web browser. Some plug-ins have become very popular, so Netscape and Microsoft browsers have built them into their latest versions. Popular multimedia plug-ins include RealNetwork's Real Audio and Real Video, Apple's Quick Time, and Macromedia Shockwave and Flash.

Besides plug-ins, which allow multimedia objects to be included in Web pages, more sophisticated Web pages also include Java Applets and VRML (virtual reality modeling language) three dimensional (3D) images. *Java Applets* are programs embedded in Web pages (in HTML code) and used to create multimedia Web pages that include animation, sound, images, and videos. The Applet runs locally on the client machine, which is running a Web browser. *VRML* is a generic text-based language used to construct 3D images and 3D text for Web pages. A VRML document is an ASCII file, much like an HTML document. In VRML, the Web designer specifies objects—called *nodes*—that define shapes and their attributes. These nodes create 3D scene graphs, which a VRML browser renders. Examples of nodes include cube, sphere, cone, cylinder, rotation, and scan.

Although including all these components makes Web pages multimedia rich, it does not make them interactive. The complete process of creating interactive multimedia Web pages, and the related tools and languages, is shown in Figure 2. This process is described next.

The original (home) Web pages, hosted at a server, include multimedia objects (such as music, sound, images, animations, and video clips), Java Applets, and VRML images. To provide interactivity, a Web designer should also apply client-side and/or server-side scripting.

Client-side scripting includes CGI (common gateway interface) scripts that provide interactivity by processing and responding to user inputs. When a user submits an HTML form, a program that resides on the server system receives and processes the form's content. This program then creates an HTML document dynamically, which the server displays back to the user on the client machine. CGI scripts are usually created using Perl, JavaScript, or VBScript.

Perl is an interpreted scripting language for processing data collected from HTML forms, which is well suited for text manipulation. Its features include flexibility, compact, form, security, and support for file and database operations.

JavaScript is another scripting language that is used to create and process Web pages with HTML forms. However, the browser, and not a program on the server, executes JavaScript, adding interactivity and dynamic content to Web pages.

Java Media Framework (JMF) tools provide the integration into Java applications and Applets of a wide range of audio and video formats, advanced imaging, animation, 2D and 3D graphics and modeling, as well as speech and telephony support. The JMF consists of a suite of Java media APIs (application programming interface).

JMF player API is used by applications and Applets to create and control media playback. Java 2D API (J2D) and 3D API (J3D) provide sets of classes for writing 2D and 3D graphics applications and Applets, respectively. Java Advanced Imaging API allows sophisticated image processing to be incorporated into Java Applets and applications.

VBScript is a scripting language from Microsoft that is based on Visual Basic, which is used to create interactive multimedia Web pages. Like JavaScript, VBScript statements are embedded in an HTML code, which a browser executes. VBScript provides access to OLE (object linking and embedding) objects, called ActiveX controls.

ActiveX is an OLE object, which is a class library that contains a complete set of functions for manipulating an

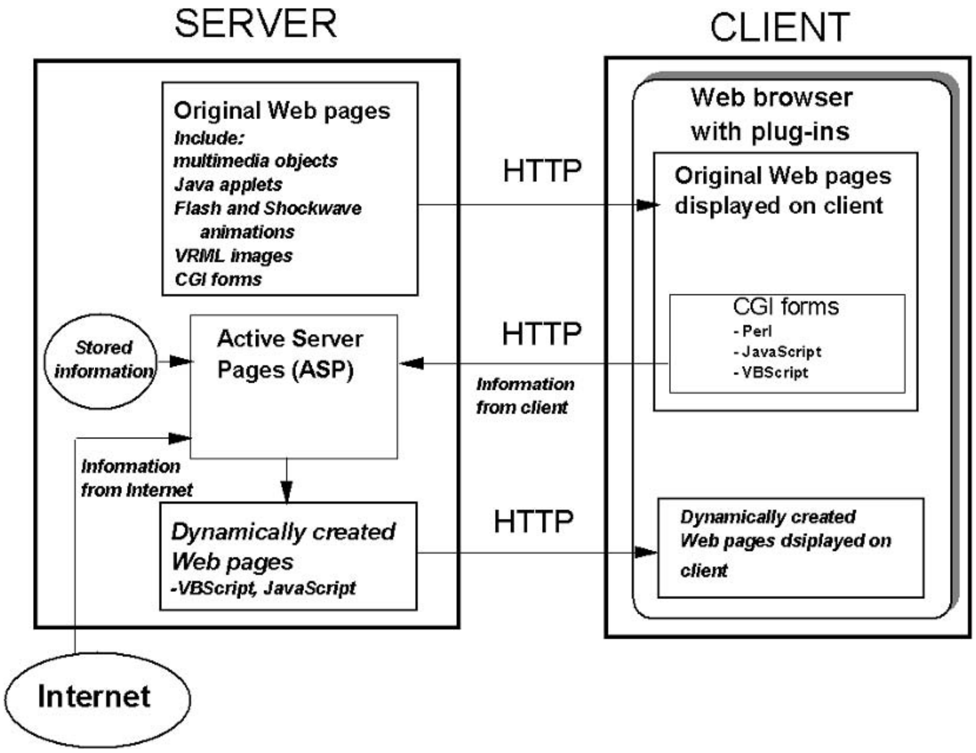

SERVER
CLIENT
Original Web pages
Include:
multimedia objects
Java applets
Flash and Shockwave animations
VRML images
CGI forms
HTTP
Web browser with plug-ins
Original Web pages displayed on client
Stored information
Active Server Pages (ASP)
HTTP
CGI forms
- Perl
- JavaScript
- VBScript
Information from client
Information from Internet
Dynamically created Web pages
-VBScript, JavaScript
HTTP
Dynamically created Web pages dsiplayed on client
Internet

Figure 2: A process and tools in creating interactive Web pages.

object. The code for an OLE object already exists, so programmers can include the objects in applications quickly and easily. ActiveX controls provide programmers with a way to extend browser capabilities and create a wide range of Web site capabilities. Some examples of ActiveX controls include button (control to display various frame sequences), chart (enables the programmer to draw various types of charts with different styles), and gradient (shades the area with a range of colors).

Server-side scripting provides an effective way to dynamically build documents in response to client (or user) requests. As shown in Figure 2, server-side scripting uses information sent by clients, information stored on the server, and information from the Internet to dynamically create Web pages (in XHTML or XML), which are then sent back to the client.

Active Server Pages (ASP) is a popular server-side technology from Microsoft, which is processed by an ActiveX component (server-side ActiveX control) called a scripting engine. An ASP file has an extension .asp and contains XHTML tags and scripting code. The VBScript is most widely used for ASP scripting, but other languages, such as JavaScript, can be also used.

Scalable Vector Graphics (SVG) is a Web standard that specifies the language for describing two-dimensional graphics in XML. It will provide the next-generation Web browsers with native support for multimedia. SVG allows for three types of graphic objects: vector graphic shapes (e.g., paths consisting of straight lines and curves), images, and text. Graphical objects can be grouped, styled, transformed, and composited into previously rendered objects. Text can be in any XML name space suitable to the application, which enhances searchability and accessibility of the SVG graphics. The feature set includes nested transformations, clipping paths, alpha masks, filter effects, template objects, and extensibility. SVG drawings can be dynamic and interactive. The Document Object Model (DOM) for SVG, which includes the full XML DOM, allows for straightforward and efficient vector graphics animation via scripting.

In summary, a number of tools, languages, and components, which are introduced in this section, are used to create contemporary and effective interactive multimedia Web pages.

TOOLS, SERVICES, AND APPLICATIONS

Today, there is a wide range of interactive multimedia services and applications on the Web. In this section, we discuss the basic services and applications as well as related tools. The list of these services and applications is given in Table 1.

Multimedia Portals

Market environment is changing rapidly and therefore new functionality is important for gaining competitive advantage. Time-to-market is critical, and businesses need to integrate systems, automate processes, and provide each other with access to key functionality. Recently, Web portals evolved as a single, integrated point of access of information, applications, and people. Web portals are Web sites that offer a great deal of

Table 1 Interactive Multimedia Services and Applications on the Web

APPLICATION	DESCRIPTION
Multimedia portals (sample URL: http://www.eUniverse.com)	Give users single, customizable access to multimedia services, including video, audio, animation, interactive banners, etc; can be classified as entertainment, educational, and information portals.
Interactive e-learning and distance learning (sample URL: http://www.click2learn.com)	Applications include a wide variety of media files to interactively explain complex concepts; distance learning allows customers to subscribe to online courses taught at remote sites.
Media sharing (sample URL: http://www.ic.media.mit.edu)	Provides collaborative music and video production by a group of artists from dispersed locations.
Interactive multiplayer games (sample URL: http://www.sierra.com)	Games in which multiple players participate over the Internet.
Web 3D virtual environments (sample URL: http://www.parallelgraphics.com)	3D content that allows users to fly through virtual space.
Participatory publishing	Web content that users, along with authors, can contribute to and influence.
Interactive Web maps (sample URL: http://www.archaeology.usyd.edu.au/research/time_map)	Time-enabled maps that typically display archeological and historical data.
Multimedia search over the Web (sample URL: http://www.fastsearch.com)	Web-based search systems that provide search results as images and audio and video files on a single result page.
Integration of the Web with traditional media (sample URL: http://www.broadcast.com/television)	Applications include radio and television broadcasting on the Web, video-on-demand, and interactive TV applications.

content and services. They integrate new content with existing content, server-side applications, and Web-based services. There are two basic Web portals: (a) public portals, such as Yahoo, which bring together information from various sources, applications, and people, and which offer personalized Web sites; and (b) enterprise, or usiness, portals, which give employees access to organization-specific information and applications.

In another classification, a horizontal portal provides a gateway to a broad range of information, whereas a vertical portal provides in-depth information about a specific topic.

Multimedia portals are special Web portals that offer multimedia services, such as streaming video, audio, interactive banners, and animation (Dimitrova et al., 2000). Multimedia portals can be classified as belonging to one of the following groups. (a) *Entertainment portals* typically provide customizable music, television, and movie news. For example, a portal from Warner Brothers, shown in Figure 3, provides features such as an e-mail center; customizable access to music, DVD videos, television, and movies; entertainment news; a chat room; and interactive advertiser banners. (b) *Educational portals* include online learning applications that provide access to diverse sets of content. An example of a tool for developing educational portals is the Aspen Learning Content Management Server from click2learn (http://www.click2learn.com), which allows a team of instructional designers, media developers, subject experts, and content reviewers to work together effectively through project management, workflow, and collaboration. (c) *Information portals* provide users with the means to find information on any topic (such as Yahoo).

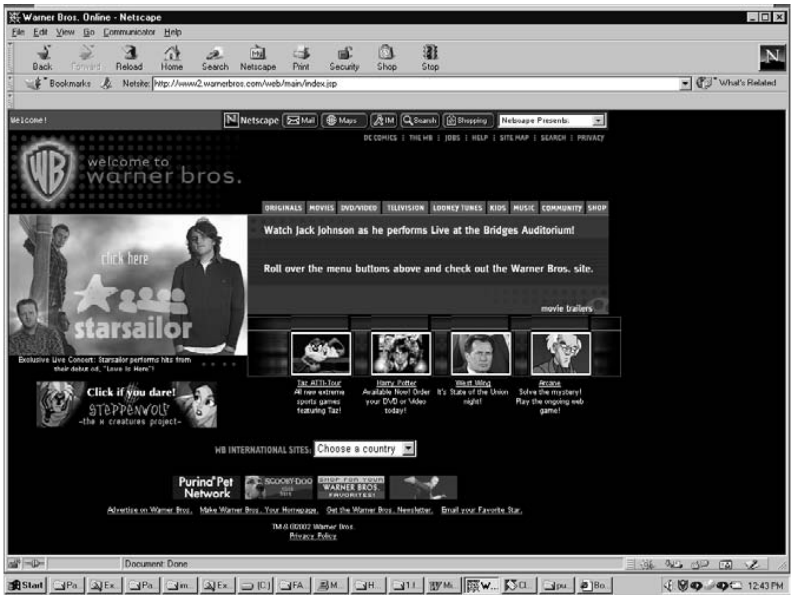
Interactive E-learning Tools and Distance Learning Applications

The Web coupled with multimedia content has the potential to change how people are educated. Contemporary e-learning tools allow the development of online interactive learning applications. An example of such a tool is ToolBook Instructor and Assistant (http://www.click2learn.com), which supports a wide variety of media files, including sound, animation, streaming media, and still images. Some interactive learning applications using ToolBook include innovative technical and maintenance training, electronic service manuals, and software training courses. Figure 4 shows a screen from the interactive Boeing 777 course, showing how skilled aircraft mechanics share techniques for installing safety hardware in confined spaces.

Distance learning allows remote participants to receive classes over the Internet. Advanced distance learning systems include a real-time interactive virtual classroom, which allows a remote participant not only to receive a live class feed but also to interact in a live class using audio and video over the Internet (Deshpande & Hwang, 2001).

Media-Sharing Applications

Media sharing includes the design of tools that support collaborative music or video production by a group of artists from geographically dispersed areas. A sharable media project team at MIT Media Lab, led by Glorianna Devenport (http://www.ic.media.mit.edu), has developed a set of tools that allows online groups of filmmakers to collaborate on the same film project by contributing


Warner Bros. Online - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Netsite: http://www2.warnerbros.com/web/main/index.jsp
What's Related
Welcome!
Netscape Mail Maps IM Search Shopping Netscape Presents
DC COMICS | THE WB | JOBS | HELP | SITE MAP | SEARCH | PRIVACY
welcome to warner bros.
ORIGINALS MOVIES DVD/VIDEO TELEVISION LOONEY TUNES KIDS MUSIC COMMUNITY SHOP
click here
starsailor
movie trailers
Click if you dare! STEPPENWOLF -the x creatures project-
WB INTERNATIONAL SITES: Choose a country
Purina Pet Network
SCOOBY-DOO
SHOP FOR YOUR WARNER BROS. FAVORITES!
Document: Done
Start
12:43 PM

Watch Jack Johnson as he performs Live at the Bridges Auditorium!

Roll over the menu buttons above and check out the Warner Bros. site.

Exclusive Live Concert: Starsailor performs hits from their debut cd, "Love Is Here"!

Taz ATTi-Tour All new extreme sports games featuring Taz!

Harry Potter Available Now! Order your DVD or Video today!

West Wing It's State of the Union night!

Arcane Solve the mystery! Play the ongoing web game!

Advertise on Warner Bros. Make Warner Bros. Your Homepage. Get the Warner Bros. Newsletter. Email your Favorite Star.

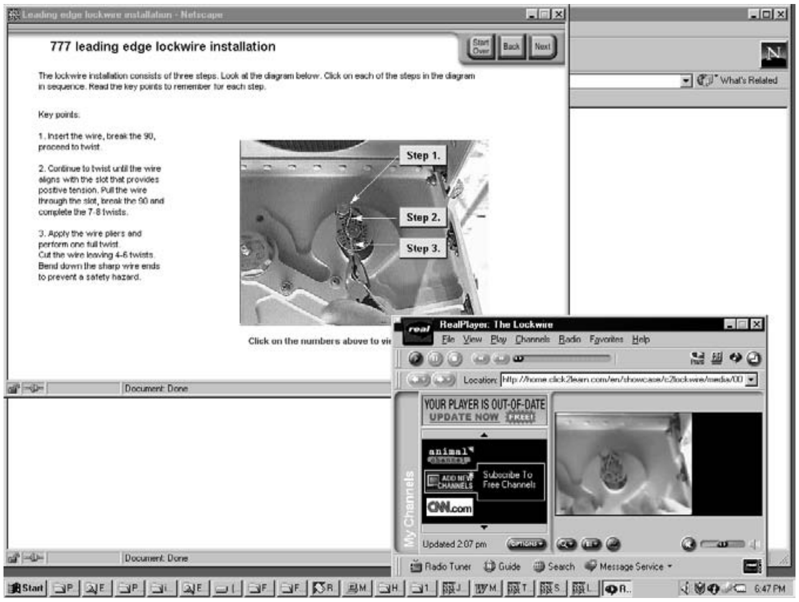
TM & ©2002 Warner Bros.
Privacy Policy

Figure 3: An example of the entertainment portal from Warner Brothers (from http://www.entertaindom.com).

and sharing material, and to comment on one another's creations (Kelliher et al., 2000). An example from this project is PlusShorts application, which is implemented as a Java Applet, and which allows distributed groups of users to contribute and collaborate in the creation of shared movie sequences.

RocketNetwork (http://www.rocketnetwork.com) has developed a tool for online musicians that provides simultaneous access to audio files and allows the collaboration of musicians and producers. This system, referred to as "global production network," allows the dynamic updating of files from anywhere in the world. Applications include sound track production for film, TV, and radio, and sound tracks for broadcast media.

The system is based on a RocketNetwork's central server, which coordinates sharing, dynamically updates, and distributes audio and MIDI (musicial instrument digital interface) files.


Leading edge lockwire installation - Netscape
Start Over
Back
Next
Step 1.
Step 2.
Step 3.
Document: Done
RealPlayer: The Lockwire
File View Play Channels Radio Favorites Help
Location: http://home.click2learn.com/en/showcase/c2lockwire/media/00
YOUR PLAYER IS OUT-OF-DATE UPDATE NOW FREE!
animal channel
ADD NEW CHANNELS
Subscribe To Free Channels
CNN.com
My Channels
Updated 2:07 pm
Radio Tuner Guide Search Message Service
What's Related
Start
6:47 PM

777 leading edge lockwire installation

The lockwire installation consists of three steps. Look at the diagram below. Click on each of the steps in the diagram in sequence. Read the key points to remember for each step.

Key points:

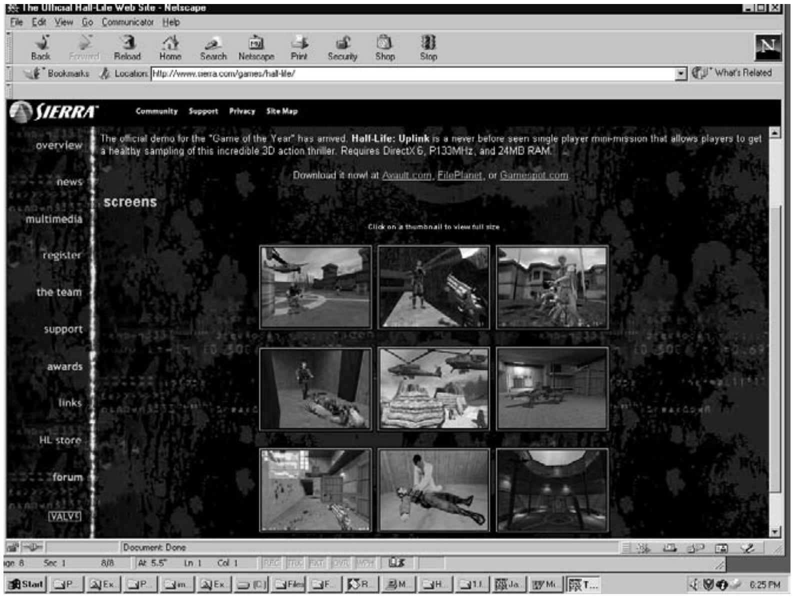
1. Insert the wire, break the 90, proceed to twist.

2. Continue to twist until the wire aligns with the slot that provides positive tension. Pull the wire through the slot, break the 90 and complete the 7-8 twists.

3. Apply the wire pliers and perform one full twist. Cut the wire leaving 4-6 twists. Bend down the sharp wire ends to prevent a safety hazard.

Click on the numbers above to vi

Figure 4: Boeing 777 interactive maintenance course created using the ToolBook e-learning tool (from http://www.click2learn.com).


The Official Half-Life Web Site - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.sierra.com/games/half-life/
What's Related
SIERRA
Community Support Privacy Site Map
overview
news
multimedia
register
the team
support
awards
links
HL store
forum
VALVE
Document: Done

The official demo for the "Game of the Year" has arrived. **Half-Life: Uplink** is a never before seen single player mini-mission that allows players to get a healthy sampling of this incredible 3D action thriller. Requires DirectX 6, P133MHz, and 24MB RAM.

Download it now! at Avault.com, FilePlanet, or Gamespot.com

screens

Click on a thumbnail to view full size

Figure 5: Half-Life from Sierra.com—illustration of a multiplayer Web game (from http://www.sierra.com).

Multiplayer Web Games

Interactive multiplayer games allow multiple players to participate in games over the Internet. The challenge of Web-based multiplayer games is in the quality of the connection and the ability of the designer to implement the plot in a cyberspace way. Real-time games also require that players visit a chat room and register and that all clients be synchronized. Figure 5 illustrates an example from the game Half-Life, created by Sierra.com (http://www.sierra.com).

Web 3D Virtual Environments

Support for VRML in Web browsers allows developers to present their content in 3D rather than just as documents. The user can "fly" through virtual space and experience a real car, house, or town. The browser requires a plug-in, a Web browser that provides complex VRML rendering. Figure 6 illustrates the Cortona VRML plug-in, by ParallelGraphics for a 3D model of a sport car.

The same company, ParallelGraphics, offers Outline3D, an interior design application that enables users to quickly customize a virtual environment specific to their design needs (http://www.parallelgraphics.com). Potential applications include (a) a 3D room for potential buyers to view the finished interior design of an office or a home, and (b) real estate applications, for real estate agents to show customers the look of their future home.

Participatory Publishing

Participatory publishing refers to Web pages where readers can contribute and influence the content along with authors.

An example of participatory publishing is WebTour, an application developed by the Siemens Research Center in Princeton, in which viewers can make verbal comments and dynamic annotations or drawings to illustrate some points. The system is capable of recording, playing back, storing, searching, and distributing personalized dynamic multimedia annotation on Web documents in the form of guided Web tours.

Potential applications of participatory publishing tools, including WebTour, include the following.

E-commerce: When a customer enters the Web site of a retail store, a Web clerk greets the customer and guides him/her through various Web pages explaining key products and adding advertisements when necessary.

Corporate Intranet: Different users, such as employees and/or customers, can make annotations on the same Web document and share ideas.

Distance learning: A teacher can make annotations on Web pages explaining complex problems. The annotations could then be retrieved by students and played back.

WebTour is implemented using open Web technologies described earlier—standard browsers, Java, dynamic HTML and JavaScript, and Active Server Pages.

Interactive Web Maps

Interactive Web maps are time-enabled maps created from distributed databases. Interactive maps typically display archaeological and historical data, such as historical maps and images, the growth of cities, the spread of dynasties and empires, or the distribution of archeological sites. However, there is a wide range of other applications, which could include time-stamped data. Interactive Web maps could also link map objects to images, other Web pages, multimedia, or a database. In addition, interactive

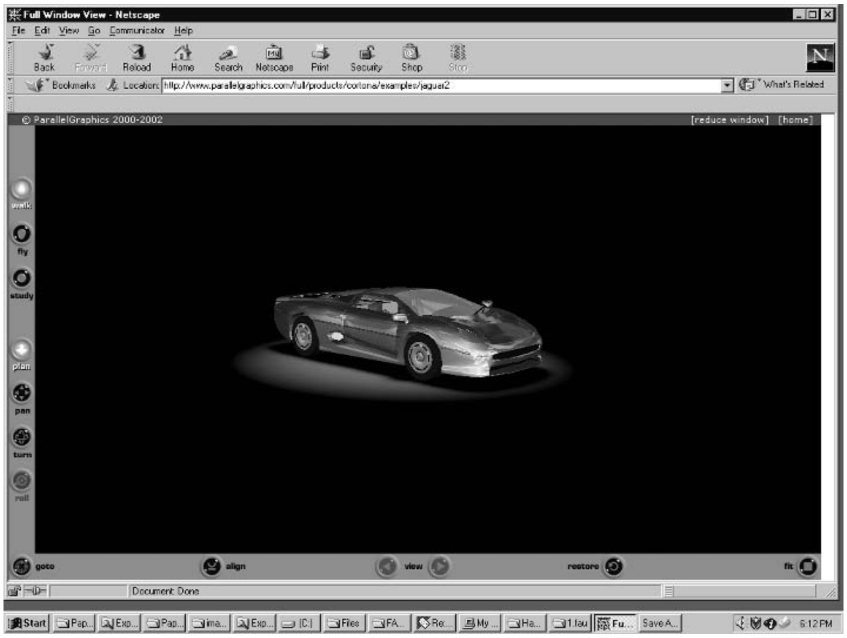

Full Window View - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.parallelgraphics.com/full/products/cortona/examples/jaguar2
What's Related
© ParallelGraphics 2000-2002
[reduce window] [home]
walk
fly
study
plan
pan
turn
roll
goto
align
view
restore
fit
Document: Done
Start
6:12 PM

Figure 6: VRML model of classic sport cars. The navigation buttons allow the user to change views, open the door, and change the car's movement (from http://www.parallelgraphics.com).

features include zooming and panning, temporal filtering, data-driven hotlinks, and other functionality.

An example of a tool for creating interactive Web maps is TimeMap, from the University of Sydney (http://www.archaeology.usyd.edu.au/research/time_map). Figure 7 illustrates an interactive map developed by the TimeMap team for the international Dunhuang project and the British Library. By selecting an area and clicking the mouse, the viewer can access another interactive screen (Figure 8). This interactive map allows the viewer to select and view archeological sites, monumental remains, routes, and so forth.

Multimedia Search Over the Web

Today, there are more that 275 million multimedia files on the Web. On average, Web surfers interact with sites containing multimedia content up to 25 times longer than with sites with static content.

FAST Multimedia Search (http://www.fastsearch.com) is currently one of the largest integrated multimedia search products on the Internet, allowing users to receive results for image, audio, and video files on a single, results page. Users simply search by keyword, and as a result, all retrieved multimedia files are accessible from the results

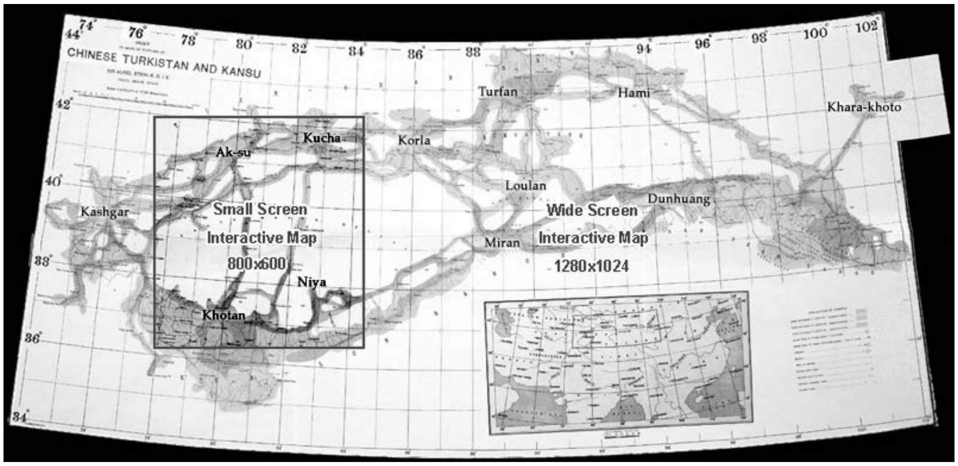
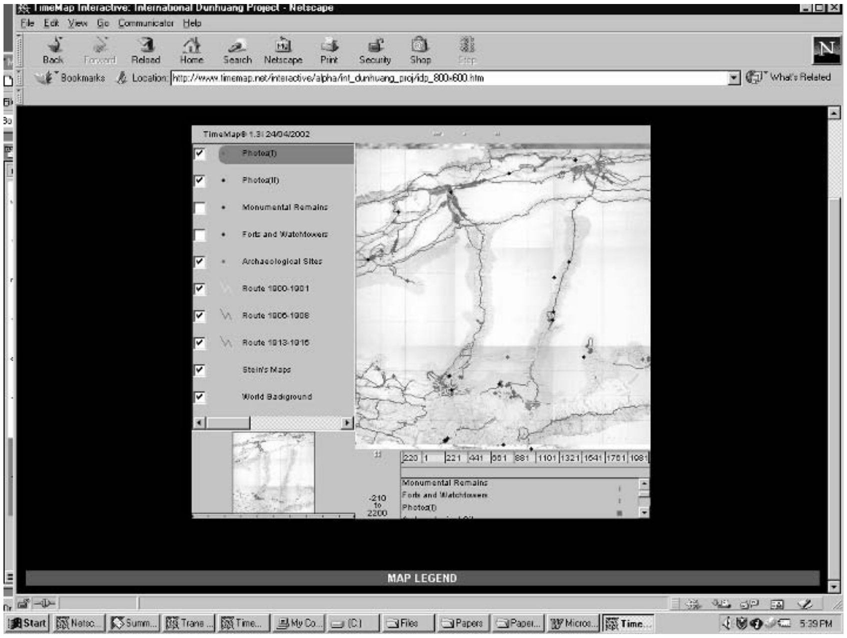


Figure 7: An example of an interactive Web map. The user can click and select an area that brings the next interactive screen (from http://www.archaeology.usyd.edu.au/research/time_map).



TimeMap Interactive: International Dunhuang Project - Netscape

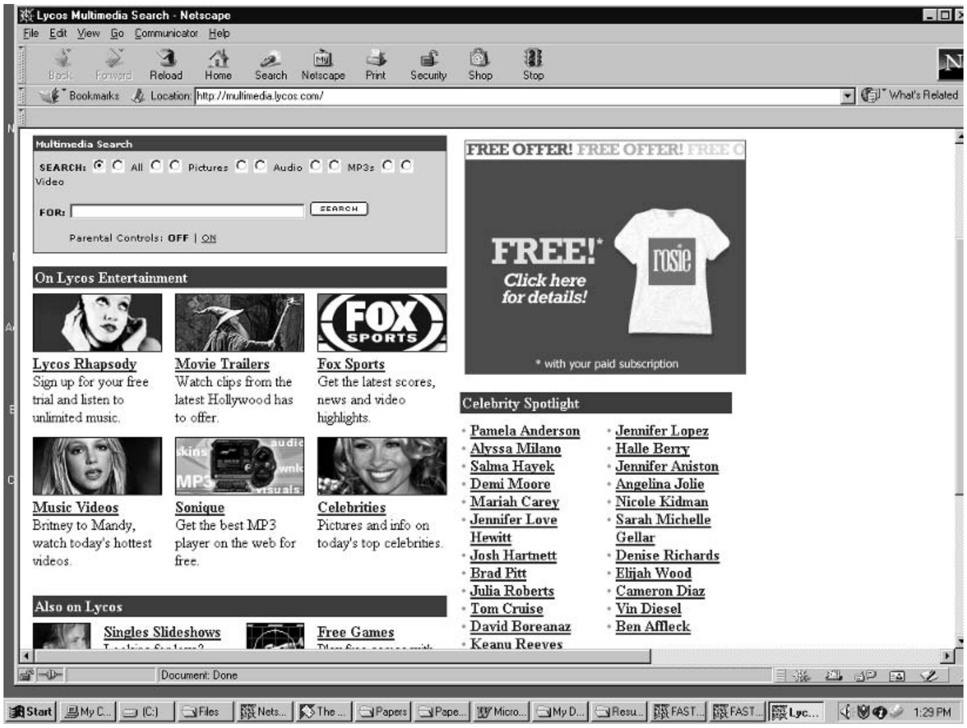
Location: http://www.timemap.net/interactive/alpha/int_dunhuang_proj/idp_800x600.htm

TimeMap® 1.3: 24/04/2002

☑ Photos(I)
☑ Photos(II)
□ Monumental Remains
□ Forts and Watchtowers
☑ Archaeological Sites
☑ Route 1900-1901
☑ Route 1906-1908
☑ Route 1913-1916
☑ Stein's Maps
☑ World Background

File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks
What's Related
-210 to 2200
220 1 221 441 661 881 1101 1321 1541 1761 1981
Monumental Remains
Forts and Watchtowers
Photos(I)
MAP LEGEND
Start Netsc... Summ... Trans... Time... My Co... (C:) Files Papers Paper... Micros... Time...
5:39 PM

Figure 8: The user can select to view archeological sites, monumental remains, routes, and so forth (from http://www.archaeology.usyd.edu.au/research/time_map).



Lycos Multimedia Search - Netscape

Location: http://multimedia.lycos.com/

Multimedia Search

SEARCH: ● All ○ Pictures ○ Audio ○ MP3s ○ Video

FOR: [SEARCH]

Parental Controls: OFF | ON

On Lycos Entertainment

Lycos Rhapsody
Sign up for your free trial and listen to unlimited music.

Movie Trailers
Watch clips from the latest Hollywood has to offer.

Fox Sports
Get the latest scores, news and video highlights.

Music Videos
Britney to Mandy, watch today's hottest videos.

Sonique
Get the best MP3 player on the web for free.

Celebrities
Pictures and info on today's top celebrities.

Also on Lycos

Singles Slideshows

Free Games

FREE OFFER! FREE OFFER! FREE O

FREE!* Click here for details!

* with your paid subscription

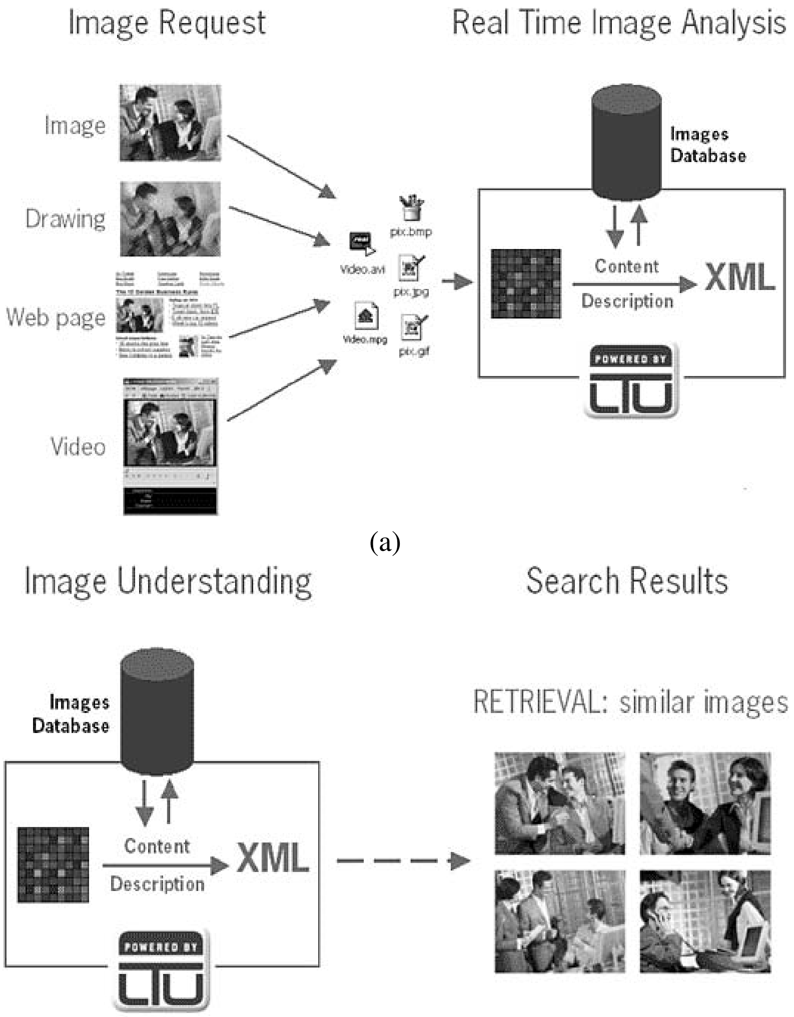
Celebrity Spotlight

- Pamela Anderson
- Alyssa Milano
- Salma Hayek
- Demi Moore
- Mariah Carey
- Jennifer Love Hewitt
- Josh Hartnett
- Brad Pitt
- Julia Roberts
- Tom Cruise
- David Boreanaz
- Keanu Reeves
- Jennifer Lopez
- Halle Berry
- Jennifer Aniston
- Angelina Jolie
- Nicole Kidman
- Sarah Michelle Gellar
- Denise Richards
- Elijah Wood
- Cameron Diaz
- Vin Diesel
- Ben Affleck

Document: Done

File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks
What's Related
FOX SPORTS
rosie
skins MP3 audio wnk visuals
Start My C... (C:) Files Nets... The ... Papers Pape... Micro... My D... Resu... FAST... FAST... Lyc...
1:29 PM

Figure 9: Lycos offers the largest integrated multimedia search engine on the Web. It includes 60 million images, sounds, and videos (from www.multimedia.lycos.com).



(a)

(b)

Figure 10: (a) An image is presented to the system and real-time image analysis is performed by extracting features of the image. (b) The system then matches these features with the features of images in the database and retrieves similar images (from http://www.ltutech.com).

page. This tool has been used by Lycos to create their multimedia search engine (http://www.lycos.com), which is illustrated in Figure 9.

Another, more advanced search of the multimedia repository is content-based retrieval, where the user provides information about the actual contents of the image, audio, or video data, rather than just keywords (Marques & Furht, 2002). A content-based search engine translates this information to query the multimedia database and retrieve the candidates that are likely to satisfy the user's request.

An example of a commercial content-based image retrieval system is a product suite from LTU Technologies (http://www. ltutech.com), which includes five products: Image-Filter, Image-Indexer, Image-Seeker, Image-Shopper, and Image-Watcher. The core of the system is a high-level perceptual image analyzer that is capable of indexing, recognizing, and describing images according to their visual features. Figure 10 illustrates the operation of the system.

Integration of the Web With Traditional Media

There is a growing trend in integrating the Web with traditional media, such as radio and television. As the streamed multimedia protocols for audio and video have matured, a number of radio and television stations have begun offering live and near-live audio and TV content over the Web. Examples include Yahoo! Broadcast radio and television Web sites (http://www.broadcast.com/radio and http://www.broadcast.com/televsion).

In addition, some programs are capable of supporting active user feedback, and even participation, so we can expect a number of new video-on-demand and interactive TV applications.

CONCLUSION

The synergy between new multimedia technologies and the Web have brought new delivery mechanisms and new types of applications. We are still in the first phase of this marriage, where we are faced with such technical challenges as how to efficiently transmit multimedia data over the Internet and how to best use our new tools. However, the focus is already moving to new areas, such as a new way to refer to publishing, how mass media can better reach and involve their audiences, and how to users can better access the Web's wealth of multimedia information (Wynblatt et al., 2000).

APPENDIX

The following is a list of Web resources.

Multiplayer Games
http://www.sierrastudio.com/games/half-life
http://www.microsoft.com/games
http://www.uo.com
http://www.cryo-networks.com/uk/cryonetworks.htm
http://www. idsoftware.com
http://www.epicgames.com

Web Multimedia Portals
http://www. eUniverse.com
http://www. tollywoodlive.com
http://www.oz.net/blam
http://www.click2learn.com

Interactive Web Maps
http://www.eparka.com
http://www.hexatech.com
http://www.archaeology.usyd.edu/au/research/time_map

Web Multimedia Search Engines
http://www.fastsearch.com
http://www.multimedia.lycos.com

Content-Based Search Engines
http://www.ccrl.com/amore
http://www.qbic.almaden.ibm.com
http://www.ctr.colombia.edu/WebSEEK
http://www.convera.com
http://www.virage.com
http://www.cobion.com
http://www.lutech.com

Radio and Television Broadcasting Over the Web
http://www. broadcast.com/television
http://www. broadcast.com/radio

GLOSSARY

Active Server Pages (ASP) Server-side technology from Microsoft that is used to dynamically create documents in response to user requests.

ActiveX An OLE object, which is a class library that contains a complete set of functions for manipulating an object. It is used to create a wide range of Web site capabilities.

Client-side scripting Scripts embedded in Web pages that provide interactivity by processing and responding to user inputs.

Constant bit rate (CBR) A type of network service in which the amount of network traffic follows a regular pattern, proportional to the amount of information that is being transferred.

High definition television (HDTV) An imaging system with a resolution of approximately twice that of conventional television in both the horizontal and vertical dimensions, and a picture aspect ratio of 16:9.

Java 3D (J3D) API Provides a set of classes for writing 3D graphics for Java applications and Applets.

Java Applets Programs that are embedded in Web pages and can be used to create multimedia Web pages.

Java Media Framework (JMF) Tools that provide the integration into Java applications and Applets of a wide range of audio and video formats, advanced imaging, animation, 2D and 3D graphics and modeling, as well as speech and telephony support.

JavaScript A scripting language that can be used to create interactive multimedia Web pages.

Jitter Refers to the variance of the overall end-to-end delay between source and destination in a packet-switched network.

Joint Photographic Experts Group (JPEG) Responsible for definition of a standard widely adopted for encoding and compression of digital images.

Mean packet transfer delay The average value of the overall end-to-end delay between source and destination in a packet-switched network.

MPC-1 Part of a set of standards that defined minimum hardware and system software specifications for PC hardware for the earliest multimedia PCs. MPC-1, particularly, was based on Intel's 386SX processor. The last MPC standard published was MPC 3. There are currently no plans for the publication of any additional MPC standards.

Motion Pictures Expert Group (MPEG) Responsible for definition of several standards widely adopted for encoding, compression, and description of audiovisual contents.

Packet switching Refers to a switching technique by which the information to be transferred between two end computers is divided into packets and relayed through intermediate nodes (routers) along the way.

Quality of service (QoS) A collection of application and/or network requirements that must be met to enable an application to work on a specific network.

Server-side scripting Provides a way to build Web documents in response to user requests on the server side. Besides the information sent by users, it also uses information stored on the server and from the Internet to dynamically create Web pages.

Streaming Refers to the continuous transfer of information from source to destination.

Transmission control protocol/Internet protocol (TCP/IP) A standard network protocol stack used on the Internet and on many private networks.

Variable bit rate (VBR) A type of network service in which the amount of network traffic may follow an irregular pattern, mainly because of the nature of compression algorithms used at the source.

VBScript A scripting language from Microsoft that is based on Visual Basic. It is used to create interactive multimedia Web pages.

Virtual reality modeling language (VRML) A text-based language used to construct 3D images and 3D text for Web pages.

CROSS REFERENCES

See *Active Server Pages; ActiveX; Circuit, Message, and Packet Switching; Java; JavaScript; Multimedia; Video Streaming; Virtual Reality on the Internet: Collaborative Virtual Reality; Visual Basic Scripting Edition (VBScript).*

REFERENCES

Deshpande, S., & Hwang, J.-N. (2001). A real-time interactive virtual classroom multimedia distance learning system. *IEEE Transactions on Multimedia, 3*(7), 432–444.

Dimitrova, N, Yu, H., Galliano, F., Koenen, R., Zakhor, A., & Bouman, C. (2000). Entry into the content forest: The role of multimedia portals. *IEEE Multimedia, 7* (3), 14–20.

Halsall, F. (2001). *Multimedia communications: Applications, networks, protocols, and standards*. Harlow, U.K.: Addison-Wesley.

Kelliher, A., Seo, J.J, Pan, P., Lin, C., & Davenport, G. (2000). Visual interfaces for shareable media. *ISEA 2000 International Symposium on Electronic Art*.

Marques, O., & Furht, B. (2002). *Content-based video and image retrieval.* Norwell, MA: Kluwer Academic Publishers.

Wynblatt, M., Benson, D., & Hsu, A. (2000). Multimedia applications on the internet. In B. Furht (Ed.), *Handbook of Internet Computing* (pp. 307–331). Boca Raton: CRC Press.

International Cyberlaw

Julia Alpert Gladstone, *Bryant College*

INTRODUCTION

The economic benefits that can be derived from business transactions conducted on the Internet have been recognized for several years. Cyberspace has evolved from a basic network established to accommodate the needs of government and academic researchers to a domain of glitzy advertisers and international financial machines. Some forecasters estimate business-to-business e-commerce will generate $8.5 trillion by the end of 2004 (Pastore, 2002). The true potential of the Internet as a commercial venue remains elusive, however, because of the legal uncertainty surrounding what laws are appropriate for governing cyberspace activity.

Questions arise as to whether land-based laws should apply to cyberspace or whether regulations that account for the borderlessness, or the virtual nature, of the medium are more appropriate (Gladstone, 2000). Cyberspace is global; therefore transactions cross national borders. As each nation seeks to protect its citizens and preserve its sovereignty, the nation establishes its own set of rules and regulations that it applies to cyberspace. This creates a conflict as to which nation's laws to apply to a global Internet transaction. This chapter reviews the major national and international laws in the areas of jurisdiction, privacy, electronic signatures, and encryption as well as copyrights and patents that presently impact cyberspace activity, and suggests that there is an incongruence between nations in many areas and that global attention must be focused to harmonize the laws in these areas to provide certainty and consistency. These particular topics of law were selected for examination because they create the foundation of the legal structure of the Internet upon which other, more-country-specific legal issues, such as consumer protection, taxation, gambling, and obscenity, are derived. Predictability and reliability are critical for the growth of commerce, and the law can play a key role in providing this element of certainty.

This chapter begins by exploring the fundamental question of who, if anyone, has the legal authority to regulate the Internet. In the context of transnational law, this is, in essence, the issue of personal jurisdiction. The author reviews the challenges that are often raised against a nation's exercise of power beyond its borders and moves on to review the current status of jurisdiction jurisprudence for cyberspace. Concern about the protection of personal information threatens the growth of commerce on the Internet. In the section on privacy, the author examines the national laws and international laws and treaties that loosely form a privacy policy in cyberspace. Commerce is the major growth area on the Internet that depends on reliable contracts. The use of encryption and digital signatures, which can insure the enforceability of electronic contracts, is discussed in the section on encryption, which also analyzes the current global status of electronic signature legislation.

Our definitions and understanding of rights to ownership in intellectual property have been challenged in response to the movement of information onto the Internet, and nations have responded to these changes in various ways. The section on intellectual property reviews the principal international and national rules in the areas of patent and copyright infringement as they have been influenced by and responded to the activity in cyberspace. Trademark issues are not included because this is a matter that is closely tied to the issue of registration of domain names, which with the development of the Internet Corporation for Assigned Names and Numbers (ICANN) has become its own specialty, and it is best left for a separate chapter.

The lack of uniformity in approach to legislation in many substantive areas is evident from the analysis presented in this chapter, and the author suggests that policy makers remove their national blinders in order to foster e-commerce.

JURISDICTION

Jurisdiction—An Overview

An essential element for the profitable evolution and efficiency of electronic commerce is for business and consumers to be able to act with confidence, knowing the laws that will apply to transactions in which they are engaging. Compliance with rules is impossible without an understanding of whose law is applicable, and the corollary to this is where the legal dispute, if it arises, will be resolved (ABA Project, 2000). These are questions of personal and prescriptive jurisdiction, which a century and a half ago were resolved easily when people lived and died in small geographic areas and the law of that place applied. The sovereign power of any state, which legitimizes the enforcement of its laws, is based upon territoriality. A general correspondence exists between borders drawn in physical space and those drawn in law space. National borders have formed the sovereignty paradigms for regulatory authority and decision making. When parties to the action lived, and the activities in dispute occurred, in a single state, that state's courts and laws were the only obvious and uncontested jurisdictional choice (Mody, 2001).

Long before the Internet, however, social changes challenged the territorial jurisdictional principles across all borders, and the concept of contacts between a foreign defendant and the foreign state was invoked to establish principles of fair and just jurisdiction. The additional jurisdictional challenges that the Internet has presented are due to the perception of the Internet as its own distinct "cyber sovereignty," which would thus make it impervious to real-world legal rules. Some people have argued that the lack of geographical borders in cyberspace turned cyberspace into its own distinct sovereignty beyond the scope of territorial law (Johnson & Post, 1996). However, for the purpose of applying the law, these separate realties need not be seen as distinct. The technology does change how the parties communicate and their state of mind, but it does not change the fact that the parties still exist in physical space, which is the critical point for jurisdictional analysis.

In order to proceed with determining jurisdictional issues on the Internet, the question of whether the law should view cyberspace as a place, a means of communication, or a technological state of mind needs to be resolved. Initially, the number of participants were small, the activities were limited, and the view of cyberspace as a state of mind or a place beyond the rule of law was appealing and even plausible, but in the year 2002, the number of cyberspace transmissions are diverse and they extend well beyond a technological phenomenon and into the real world. Cyberspace, or the Internet, represents merely another means of communication, where a terrorist can promulgate bomb making or kidnapping tips, merchants can conspire to fix prices by e-mail, a corporation can issue fraudulent securities, or a pornographer can sell child porn. Simply stated, there is a tremendous amount of previously unavailable information that is now available free of charge, which makes the Internet susceptible to real-world legal rules.

Generally speaking, a state can only enforce its laws against a defendant when there is a local presence or when there are assets within the local jurisdiction (ABA Project, 2000). The Internet's architecture allows information to flow without bounds, and therefore, the individual or organization supplying the information cannot control where that content will end up. There are no physical barriers or cues to notify the provider; therefore, an activity that may be lawful in Belize, for example, may be easily accessed in New York where it is unlawful (Mody, 2001). Individual states invest time, money, and effort to protect the welfare of its citizens and, naturally, legislate accordingly. There is an apparent conflict between a state's sovereign interest to protect its citizens and a foreign content provider's ability to carry on its lawful activity. Imposing the state's law on the content provider, or transnational regulation, raises objections regarding the violation of jurisdictional norms. The main criticism regarding transnational cyberspace breaking jurisdictional rules is based upon the notion that a state may not act beyond its own territorial borders.

Fundamental Jurisdictional Principles Under International Law

Jurisdiction can be broken into two categories: prescriptive jurisdiction, which addresses the authority of a state to apply its own laws to regulate conduct; and enforcement jurisdiction, which is the executive's authority to compel compliance with these laws. The focus of this chapter is on prescriptive jurisdiction, although issues of enforcement jurisdiction will be raised in the discussion of the Yahoo! case (*Yahoo! Inc., v La Ligue Contre Le Racisme Et L'Antisemitisme,* 2001).

The threshold matter of "jurisdiction to prescribe" means that the substantive laws of the forum country are applicable to the particular persons and circumstances (Restatement, 1987).

When a country has jurisdiction to prescribe, it can appropriately apply its legal norms to conduct. Simply stated, a country has jurisdiction to prescribe law with respect to (a) conduct that, wholly or in substantial part, takes place within its territory; (b) the status of persons, or interests in things, present within its territory; (c) conduct outside its territory that has or is intended to have substantial effect within its territory; (d) the activities, interests, status, and relations of its nationals outside as well as within its territory; and (c) certain conduct outside its territory by persons who are not its nationals that is directed against the security of the country or against a limited class of other national interests (Restatement, 1987).

Jurisdiction to adjudicate or to enforce means that the tribunals of a given country may resolve a dispute to enforce a judgment where the country has jurisdiction to prescribe the law. The exercise by a country of jurisdiction to enforce is subject to the requirement of reasonableness. States exercise jurisdiction to adjudicate on the basis of various links, including the defendant's presence, conduct, or, in some cases, ownership of property within the country. Exercise of judicial jurisdiction on the basis of such links is on the whole accepted as "reasonable"; reliance on other bases, such as the nationality of the plaintiff or the presence of property unrelated to the claim, is generally considered "exorbitant" (Restatement, 1987).

A country may employ judicial or nonjudicial measures to induce or compel compliance or punish noncompliance with its laws or regulations, provided it has jurisdiction to prescribe. Thus, a country may not exercise authority to enforce a law that it had no jurisdiction to prescribe. A country may employ enforcement measures against a person located outside its territory (a) if the person is given notice of the claims or charges against him and they are reasonable in the circumstances; (b) if the person is given an opportunity to be heard, ordinarily in advance of enforcement; and (c) where enforcement is through the courts, if the country has jurisdiction to adjudicate (Restatement, 1987).

The extraterritorial objections to cyberspace regulation reflect a basic misunderstanding of modern jurisdictional doctrine. Although historically a sovereign's power was tied to its geography, a state's power to regulate activity that originates outside the country but causes local harms has been recognized for many years. One of the first cases to recognize that a state's regulatory authority may extend to an extraterritorial activity was decided in the Permanent Court of International Justice (PCIJ) in 1927. In the case of *S. S. Lotus,* (Fr.v. Turk.) 1927 P.C.I.J. (Ser A) No. 10 (19227), the court held that the state of Turkey could apply its criminal law to a foreigner who acted outside of Turkey when committing the offense so charged and which prejudiced Turkey and its citizens, provided the foreigner was arrested in Turkey. This was an application of the law to a person or act outside the territory of Turkey that had an effect on and/or in Turkey. This early case applied the "effects principle" to expand the authority of a state. This is a logical extension of control by a state outside its territory to protect its citizens from actions taken by a defendant that had an "effect" or impact within the state. A state's territorial borders were no longer the sole determination of rule-making authority.

In the more famous case of the *United States v. Aluminum Company of America* (ALCOA), 148 F.2d 416 (2d Cir.1945), the United States Second Circuit Court of Appeals considered whether the United States could apply the antitrust provisions of the Sherman Act to a Canadian company. The anticompetitive acts took place in Canada, but the material effects were experienced in the United States, and the Court did apply United States rules to this Canadian company. Thus, with the ALCOA case, the presumption of extraterritoriality had been overcome.

Fundamental Principles of Jurisdiction Under European Law

In the European Union, the primary source of jurisdiction law has been the Convention on Jurisdiction and the Enforcement of Judgments in Civil and Commercial Matters, September 30, 1968 (the Brussels Convention). In general, jurisdiction is based upon the defendant's domicile, and alternative jurisdictional grounds are available if there is a close link between the court and the action or if the "sound administration of justice" would be facilitated. In contract matters, the place of performance would govern the jurisdiction decision. In tort matters, jurisdiction would lie in the place where the harmful event occurred or where there was a risk of it occurring.

The Brussels Convention was modified, effective March 2002, as the European Union issued the so-called Brussels Regulation (the Brussels Regulation). In contrast to a convention or directive, a "regulation" of the European Union becomes binding immediately after its adoption by the 15 member states without the need for further implementation. The economic drive of electronic commerce created the need for certainty and uniformity of jurisdictional rules early on; therefore, the European Union found it efficient to proceed quickly with a mere regulation. Although the Brussels Regulation does not alter the main structure of the Brussels Convention, it effectuates certain changes that take account of the new technological developments that result from e-commerce. Most important, the Brussels Regulation, which is consumer centered, establishes the fact that the courts of the consumer's domicile will have jurisdiction over a foreign defendant if the latter "pursues commercial or professional activities in the Member State of the consumer's domicile or, *by any means,* directs such activities to that Member state . . . and the contract falls within the scope of such activities." This language expands the range of situations in which the consumer can sue in his or her place of domicile.

The phrase "by any means" was included to broaden the scope of jurisdiction to reach internet-based transactions (Explanatory Memorandum, July, 1999). The Brussels Regulation equates doing business or offering goods and services via the Internet with an invitation or with advertising by businesses that "by any means" directs their activities toward that member state.

In essence, the Brussels Regulation stipulates that an unintended effect in a member state can be the basis for jurisdiction. Because jurisdiction in European countries is not limited by U.S. constitutional principles of due process, as it is in the United States, the Brussels Regulation does not require notice or "minimum contacts."

The Brussels Regulation was controversial; the negotiations reflected the tension between business and consumer groups. Industry groups claimed that it would hinder the growth of e-commerce, by making small- to medium-size businesses reluctant to set up Web sites for fear of being subjected to the jurisdiction of the courts of too many countries. Alternatively, the EU Commission believed that without strong consumer protection, the negative impact on consumer confidence would hurt the unified European market (Explanatory Memorandum, October 2000). If, as a result of reluctance to venture into the World Wide Web to shop, consumers were to stay within their own country, the Commission believed the EU e-commerce sector would be put at a significant competitive disadvantage to the United States; this is based upon the belief that the United States has stronger consumer protection laws.

The EU Parliament has also passed its Electronic Commerce Directive, which applies a more restrictive jurisdictional doctrine to legal disputes, restricting plaintiffs to a "country of origin" approach. Under the Electronic

Commerce Directive, "the law of the country of origin (which is the seller's place of business) would govern the cross-border disputes." Although this may provide more certainty for businesses, it provides consumers with much less confidence in their business dealings (Boam, 2001). There are several exclusions from the Electronic Commerce Directive including "contractual obligations concerning consumer contracts;" therefore, it is not know how these exclusions will work with the Brussels Convention as amended, and in the short run, case by case examination of jurisdictional issues will be required.

Classic United States Jurisdiction Principles

In the United States, assertion of jurisdiction over a person must satisfy the standard of constitutional due process. States exercise jurisdiction over nonresidents under their respective long-arm statutes. First, to establish personal jurisdiction over a defendant, a United States court will apply the relevant long-arm statute to see whether it permits the exercise of personal jurisdiction. Second, the court will apply the precepts of the Due Process Clause. A standard inquiry for whether due process has been satisfied focuses on whether the defendant has "minimal contacts" within the forum such that assertion of jurisdiction does not offend the "traditional notions of fair play and substantial justice" (*World-Wide Volkswagen Corp. v. Woodson,* 444 U.S. 286, 1980).

Under the Due Process Clause of the Constitution, one must look at the relationship between "the defendant, the forum and the litigation" (Rice, 2000). Physical presence is not required; rather the plaintiff must show that the defendant has purposefully directed its activities toward the forum state, or otherwise "purposefully availed itself of the privilege of conducting activities within the forum State, thus invoking the benefit and protection of its laws" (*Hanson v Denckla,* 357 U.S. 235, 1958).

Questions of jurisdiction in cyberspace have generated lengthy court decisions and caused much global debate because of the unique nature of Internet information dissemination. The current global jurisdictional case law suggests that two principal tests are being used to ascertain jurisdiction (Rice & Gladstone, 2002).

One is the "Zippo test," which is named after the case that first articulated it (*Zippo Manufacturing Co.v Zippo Dot Com,* Inc., 952 F. Supp. 119, W. D. Penn., 1997), and the other is the "effects test," which is based upon a standard developed from a U.S. Supreme Court case that arose in the context of print media (*Calder v Jones,* 465 U.S. 783, 1984). Both of these tests constitute a course correction from the earliest U.S. decisions which had based specific personal jurisdiction over a nonresident defendant on mere accessibility in the forum state of the defendant's Web site.

The Zippo test establishes jurisdiction over a nonresident defendant based upon the degree of interactivity between the Web site and the forum. Mere access to the nonresident's Web site is the least interactive; under the Zippo test, a passive Web site is not sufficient to establish specific jurisdiction. The district judge explained that the decision of whether jurisdiction could be properly asserted in a case was to be based upon the nature and quality of the commercial activity that an entity conducted over the Internet. Jurisdiction cases fall somewhere on a "sliding scale" or "spectrum," where the likelihood that personal jurisdiction can be constitutionally exercised is directly proportionate to the nature and quality of the commercial activity that an entity conducts over the Internet.

The primary difficulty in applying the Zippo sliding-scale standard to jurisdiction cases in cyberspace has been determining the degree of "interactivity." Similar fact patterns have led to different jurisdictional findings in different courts. In addition, determining whether a Web site has been integral in a forum often hinges more on a court's perception than on real differences in the manner in which the user employs the Internet. For example, in the year 2000, a judge in the Southern District of New York found that the mere availability of the defendant's Web site in New York made it "intuitively apparent" that defendant's services were used by New York residents (*Cable News Network, L. P. GoSMS.com, Inc,* 2000 WL 1678039, Southern District of New York).

The judge, therefore, found grounds for jurisdiction even though he acknowledged the plaintiff's allegations that the defendants' mobile telephone and two-way e-mail services allegedly used in New York were "factually unsupported."

Jurisdiction Based on "Effects"

The first instance when "purposeful direction" arose was in the context of traditional media, and it has become the basis for all United States cases applying the "effects test." In *Calder v. Jones,* 465 U.S. 783 (1984), Florida residents who had essentially no physical contact with California wrote and edited an article in the *National Enquirer* that defamed Jones, who was a well-known movie actress residing in California. The *Enquirer* had greater circulation in California than in any other state, and the material was based upon California sources. The U.S. Supreme Court found jurisdiction holding that California was the focal point both of the story and the harm suffered. The court held that the defendants' acts were intentional, that they were aimed at California, and that its effects took place in California.

The first use of the effects test in asserting jurisdiction against a defendant in an Internet context was in *Panavision Int'l. L.P v. Toeppen,* 141 F.3d 1316 (9th Cir. 1998), which was a "cyber-squatting" case. Toeppen, an Illinois defendant, had intentionally registered the California plaintiff's trademark as his domain name, namely Panavision. When attorneys for Panavision contacted Toeppen to demand that he stop using the name, he responded by offering to sell the name for $13,000 and promising not to acquire any other, similar names. Unwilling to be bribed, Panavision sued Toeppen in California District Court, where Toeppan objected on jurisdictional grounds. The district court found jurisdiction appropriate under the effects test because the defendant had intentionally directed its conduct toward California, knowing that the effect of his registering the domain name would be felt in California. The Ninth Circuit agreed, analogizing cyber-squatting to an intentional tort and found that "the brunt of the harm to Panavision was felt in California."

In a recent Australian libel case, the effects test was applied to assert jurisdiction to protect a Melbourne citizen. The Victorian Supreme Court found jurisdiction over Dow Jones, a United States company, based upon its *Wall Street Journal* Web site, which carried an allegedly libelous article about Joseph Gutnick, an American businessman who lived in Melbourne, Australia. The Australian Court found that the publication occurred and, thus, had its impact whenever the article was downloaded, thereby dismissing the defendant/Dow Jones' argument that the court lacked jurisdiction because the information was not published in Australia. Again, the impact/effect was felt in Australia (*Gutnick v. Dow Jones,* VSC 305, 2001).

Several leading cyberspace commentators have suggested that the effects test is a more useful mechanism than the Zippo test for establishing jurisdiction in Internet cases. In fact, the effects test has been refined into two parts: a "strict effects" test, which looks to the intent of a defendant acting outside the jurisdiction to establish a connection, and a looser, "soft effects" test, which focuses more simply on the impact within the jurisdiction (Rice & Gladstone, 2002). The former use of the effects test has been employed mostly in tort and intellectual property cases. Indeed, cases of defamation lend themselves easily to finding jurisdiction under the strict effects test because intent to harm is an essential element of the underlying cause of action. Currently, it is likely that a court dealing with the issue of jurisdiction over a nonresident based on his online activity will start its inquiry by using the Zippo test but will continue the analysis by applying the effects test. Therefore, attorneys who advocate jurisdiction in a particular forum would be best advised to use both tests, because the effects test may apply where the Zippo test does not.

Enforcement Jurisdiction and the Yahoo! Case

The recent lawsuit by the International League Against Racism and Anti-Semitism and the Union of French Law Students against Yahoo! (the Yahoo! Case), which has received so much attention in the popular press, summarizes many of the fundamental principles and issues that remain to be resolved in the area of international jurisdiction. In April 2000, two French groups, namely the Union of French Law Students and the International League Against Racism and Anti-Semitism, filed suit against Yahoo! for hosting auctions that displayed and sold Nazi propaganda. The memorabilia auctions were accessible only via the English language site, Yahoo.com. Direct access through Yahoo.fr was not possible. Yahoo! argued in French court that the French court did not have jurisdiction over Yahoo!. That plea was denied, and in November 2000, a French court ruled that Yahoo! must put filtering systems in place to block users in France from access to the Nazi-related goods area or pay fines of approximately $13,000 per day. Only a watered down version of the soft effects test could be seen to apply to the French court's decision in this case, and because Yahoo! was not targeting France, which is a key element in the effects test, the assertion of jurisdiction arguably violates the due process requirement of U.S. law (Rice & Gladstone, 2002).

Yahoo! chose not to appeal the French court's judgment. Rather, it challenged the enforcement of the order in the United States. In December 2000, Yahoo! filed a lawsuit in the United States District Court of Northern California, seeking a declaratory judgment that any final judgment of a French court would be enforceable in the United States. Before the California court could address the merits of the case, in a bit of an ironic twist, the French defendants motioned the California court to dismiss the declaratory judgment suit due to lack of jurisdiction. The U.S. court denied the motion to dismiss, finding jurisdiction based upon the effects theory. The court ruled that the defendant knowingly engaged in the activities and intended to have an effect on United States citizens (e.g., the use of U.S. Marshals to serve Yahoo! officers in California). Clearly, the French citizens purposely availed themselves of the benefits of the United States.

A state can only enforce its laws against a defendant in a forum where the defendant can be found or where there are assets belonging to the defendant. Enforcement of a judgment rendered by another forum requires its recognition by another court to enforce it. If it is the judgment of a court in a state in the United States, the Full Faith and Credit Clause of the Constitution requires that it be recognized by another state. Recognition sought in the United States of a judgment of a foreign court depends on the principle of "comity" (ABA Project, 2000). Comity is not a matter of absolute obligation. It is the recognition that one nation allows within its territory the legislative, executive, or judicial acts of another nation (ABA Project, 2000). National procedures required for recognition and enforcement of judgments vary widely around the globe. In the United States, comity is upheld unless doing so would violate due process, personal jurisdiction, or some public policy.

In trying to determine the enforcement jurisdiction of the French court over Yahoo!, the Federal District Court for the Northern District of California found the issue to be whether it was consistent with the Constitution and the Laws of the United States for another nation, namely France, through their court order, to curtail the Yahoo! Web site. The French, thereby, would be regulating the speech of United States' residents within the United States on the basis that such speech could be accessed by Internet users in France. The court was mindful of the extent to which the United States is governed by the "comity of nations" but did not believe that comity was a matter of absolute obligation. The court decided the case in accordance with the Constitution, finding that the French Order violated the Constitution of the United States, thereby recognizing that it was necessarily adopting the position that given "certain judgments embedded within this enactment including the fundamental judgment expressed in the First Amendment that it is preferable to permit nonviolent expression of offense viewpoints than to impose viewpoint based government regulation upon speech."

The court rendered judgment in favor of Yahoo! in a summary judgment motion that Yahoo! requested on the declaratory judgment action to find the French Order in violation of the First Amendment. This finding of a threat to Constitutional rights by the court was the grounds by which it effectively rendered the Order unenforceable, and which demonstrates the limits of perspective jurisdiction. This case suggests the disharmony that continues to exist among nations on questions of jurisdiction.

It appears that courts and legislatures have found legitimate grounds for asserting prescriptive jurisdiction over defendants based upon actions taken in cyberspace, but that may have little importance when the plaintiff seeks a restorative remedy. Enforcement jurisdiction, which requires the injured party to attach either the defendant or his tangible assets, becomes an issue of comity or state's recognition of its obligation to enforce a law. Questions of comity have not been resolved sufficiently to assure smooth enforcement on the Internet. Policy makers and governments will need to address this higher level of enforcement jurisdiction to foster the predictability and certainty necessary for the growth of commerce on the Internet.

PRIVACY

The technological efficiencies that characterize the Internet, namely transparent dissemination, collection, and aggregation of information, have not previously been experienced by society. Although the facility of data collection has economic benefits, it compromises the individual's right to privacy. During the past half decade, many sectors of the global community have been directing time and resources to resolve the inherent tension that has developed between seeking the economic benefits from modern data collection practices and ensuring human dignity, which is threatened by modern surveillance (Gladstone, 2000). The Internet is the largest electronic infrastructure that allows public access to a nearly infinite resource of information, and it holds the current greatest threat to personal privacy.

The mechanisms used to protect that information fall into three categories, reflecting either a self-regulatory, statutory, or technology approach. The self-regulatory, or market-dominated, approach, which is adopted in the United States, is based upon industry-developed norms, policies, and contracts, rather than statutory legal rights, to protect the privacy interests of its citizens. The statutory, or rights-dominated, approach, which was developed in the European models and recently adopted in Asia, New Zealand, and India, relies on statutory and common laws to establish rights to information privacy (Reidenberg, 2000). Technology is used globally in varying degrees in different sectors. Data flows on the Internet are international, and these divergent data protection policies and rules confront each other with increased frequency. Attempts at harmonization have been enacted, most notably the International Safe Harbor rules, or Safe Harbor Rules, which were adopted to implement the EU Privacy Directive, but further work toward uniformity is needed.

This section begins with a description of two examples of how one's privacy is threatened on the Internet. The first example reviews data collection practices that threaten consumer privacy. It is followed by a review of the laws in the United States that have been passed to prevent inadvertent exposure to pornography on the Internet. The section continues by exploring the differences that underlie the approach to privacy of the United States versus that of other nations, which suggests reasons for difficulties in finding a common ground. The section on EU privacy discusses the background and scope of the Council Directive 95/45/EC (the EU Privacy Directive), including a discussion of the strategies for U.S. compliance, namely the Safe Harbor rules, the Model Contract Terms, and derogations, or exceptions, allowed under national law. The technology that drives the Internet is based upon an open architecture, the natural default of which exposes information about people's actions on the Internet, whether through the World Wide Web or e-mail. Several surveillance initiatives, or technologies, by the U.S. government have drawn on this open architecture to retrieve personal information about citizens. The recent enactment of the Uniting and Strengthening America by Providing Appropriate Tools to Intercept and Obstruct Terrorism Act of 2001, Pub. L. No. 107–56, 115 Stat 272., (the USA PATRIOT Act), which has broadened the surveillance powers of the United States, is reviewed toward the end of this section. The passage of the USA PATRIOT Act suggests that the differences in the data protection policies between the United States and Europe may be getting wider.

Data Collection, Pornography, and Privacy Invasions

The expansion of digital computers and networking technology, with the Internet being the most prime example, has moved much of our social, educational, and commercial activities into an electronic environment. The technology that has enabled integrated global networks facilitates the creation of digital records showing what people spend time looking at, how much time they spend at particular sites, what messages they send, and what purchases they make. This "electronic footprint" is created by a variety of processes that can be generated by simply browsing the Web (Piera, 2001). Activity of the user within the Web site provides "click-stream data," which includes the time spent on each page and the information retrieved. There is technology that can collect and organize all this information into data packets; these are referred to as "cookies" and they are stored on the user's hard drive. The cookie is assigned a unique identifying code, and each time the user goes back to the Web site, the cookie is retrieved. It then tailors that visit to the previous behavior at the Web site. This invisible data collection is primarily conducted by business. The warehousing of transaction information and profiling of online users has become a critical component for e-commerce business models. The behavioral information enables sites to characterize users and offer them content of personal interests. Internet revenue is generated through target advertising that has become especially efficient as a consequence of the design of technological infrastructures that enables the global network.

This economic efficiency comes at the cost of lost privacy to the consumer. Studies indicate that Internet users do choose not to make online purchases because of privacy interest issues. In fact, some people argue that consumers have lost total control over their personal privacy choices.

Inadvertent exposure to online pornography presents another example of how privacy is threatened by the borderless and anonymous world of cyberspace. Internet pornography is big business, comprising 11% of the entire $9 billion e-commerce revenues in 1998, with industry experts projecting that e-porn will generate over $3 billion by 2003 (Alexander, 2002).

Many of the sites are free and serve as teasers to lure people to commercial sites. Consequently, children and adults alike may enter a pornographic Web site inadvertently. Such accidental exposure is common, particularly because users inputting Web site addresses often misspell them. In addition, the very open architecture of the Internet, where all material is equally accessible, means that sex-related materials are not segregated, as they may be in the material world. This exposure violates one's right to be let alone. Sexually explicit materials on the Internet range from the commonplace pornographic still frame, to live broadcasts of couples losing their virginity.

The general public in the United States has voiced a desire to restrict the dissemination of pornography on the Internet, and Congress has responded with at least three major legislative reforms. Congress' top-down attempts to regulate have not been successful, because in each instance a federal court has invalidated the statute based upon constitutional grounds.

The first and most significant act that Congress passed to eradicate online pornography was the Communications Decency Act (CDA), which covered a wide variety of activity. The operative provisions found in Section 502(a) prohibited the "knowing transmission of obscene or indecent messages to any recipient under 18 years of age" and criminalized the "knowing, sending or displaying of patently offensive messages in a manner that is available to a person under 18 years of age." In *ACLU v Reno,* 521 U.S. 844 (1997), in upholding the decision of the three-judge panel, the Supreme Court found that provisions of the CDA were impermissible content-based restrictions on speech. In addition, the court found that the language was facially overly broad and vague, which would create an unacceptable chilling effect on the speech of adults using the Internet.

The subsequent two laws that Congress passed were more specifically directed toward children. The Child Online Protection Act (COPA) followed the tone of the CDA by criminalizing Web publishers who used the World Wide Web to make harmful material available to minors. COPA provided broad affirmative defenses for Web publishers with restricted access to minors regarding the use of credit cards or other age verifying technology.

COPA was found to be unconstitutional because of the impermissible burden placed on protected adult speech. The district court found that COPA was neither narrowly tailored nor was it the least restrictive means used to protect children from harmful materials. Finally, the Children's Internet Protection Act (CIPA) required that all public schools and libraries with Internet access install filtering software to block access to sexually explicit Web sites. A U.S. District Court in Philadelphia found the CIPA unconstitutional because the filtering software would "block access to a substantial amount of speech that was both constitutionally protected and fails to meet even the filtering companies' own blocking criteria."

As of the time of this writing, the regulation of Internet pornography continues to present a challenge to Congress that remains unresolved. Access to the Internet, which is truly a unique marketplace of ideas, presents tensions among several fundamental rights. Our freedom of expression must be protected while one's right to be let alone or one's right to privacy must also be respected. In order to balance the right of privacy against other rights or practices, such as data collection, one must first examine a much more important question: What is the value of the right to privacy?

Views of Privacy

The American approach to privacy has evolved as one of restraint, whereas many other nations take an omnibus approach, giving the state a proactive and preventative role. The fundamental difference in the American versus the European conception of privacy can be seen most basically in the words used to describe data privacy. Americans use the term "privacy," which can refer to the right to be free from the gaze of a "peeping Tom" or the right not to disclose one's name on a Web site. Europeans use "data protection," which very specifically addresses information generated by an individual's overt activity. In the entire history of the United States, the position of Chief Counselor for Privacy lasted for 2 years only, whereas in Europe there are entire parliamentary departments and "privacy czars"devoted to data protection and privacy concerns.

As colonies of individuals who fled the tyranny of a controlling government, Americans feared a strong central government and established a Constitution that grants primary power to the states. The first ten Amendments (Bill of Rights) assure Americans additional freedoms, which they covet. The protection of privacy, which is not a fundamental right stated in the U.S. Constitution, can impose restrictions on other fundamental rights. In nearly every country within the European Union, the right to privacy is expressly stated in a constitution.

The First Amendment free-speech guarantee limits government regulation of the flow of information, including personal data. Indeed, by securing the freedom of speech, the First Amendment limits the protection of privacy. The American reluctance to grant power to the government in this area is also reinforced by the country's laissez-faire market economy. The recent threats to national security that resulted from the September 11, 2001, bombings of the World Trade Center in New York and the Pentagon in Washington, D.C. resulted in the passage of legislation, namely the USA PATRIOT Act, which suggests that Americans continue to not recognize the importance of privacy (Godwin, 2001).

The first major international accord that addressed personal data protection was the Organization for Economic Cooperation and Development Privacy Guidelines of 1980 (the OECD Privacy Guidelines). The OECD Privacy Guidelines have been adopted by all twenty-five member nations, and although they are not binding, they serve as suggestions for member countries and others in developing their domestic legislation. The OECD Privacy

Guidelines established eight basic principles that govern the handling of personal information, which are referred to as "fair information principles." They are collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability.

These are internationally accepted principles, which pertain to all types of data processing by both the public and the private sector. Application of the OECD Privacy Guidelines is based upon a loose reciprocity model of enforcement whereby personal data can flow freely between countries that provide equivalent protection.

EU Privacy Directive Explained

The most recently enacted and extensive statement of information privacy principles is the EU Privacy Directive. The EU Privacy Directive is a fully fledged system for the protection of personal data, which requires the establishment of rights for data subjects and obligations for those who process personal data. It also provides monitoring by an independent body and sets out sanctions for offenders. The goal of the EU Privacy Directive is to increase the free flow of information, and it is designed to allow personal data to be sent or processed on the same terms within the European Union and throughout the world. Each of the 15 member countries has had to pass its own legislation to implement the EU Privacy Directive, and to date, only France, Ireland, and Luxembourg have not passed such laws. The EU Privacy Directive is broken into seven chapters, which contain a total of 32 articles. The following overview of key provisions of the EU Privacy Directive illustrates the intensive focus on information privacy taken by the member states and also reflects a concern for data protection greater than that typically found under United States law.

Article 2 of the EU Privacy Directive contains the operative definitions of the EU Privacy Directive, which include personal data, processing of personal data, personal data filing systems, controllers, processors, third parties, and recipients. The EU Privacy Directives employs the terms "controller" and "data subject," which creates a top-down assumption of computer networks, not necessarily personal or individual use of computers. These assumptions are less applicable or useful in a world of personal computers where persons are browsing on the Internet.

Articles 6 and 7 of the EU Privacy Directive provide the general rules on the lawfulness of the processing of personal data. Article 6 establishes data quality principles by requiring that personal data must be processed fairly and lawfully and that such data be accurate and kept up to date and in a form that permits identification of data subjects for no longer than necessary. Lawful and fair processing is further defined by requiring that the data be collected for explicit, legitimate purposes. In addition, there can be no "secondary use" or "sharing" of data. Affiliate sharing or secondary use of personal data is allowed in the United States under the Financial Services Modernization Act of 1999 (Gramm-Leach-Bliley Act, Pub.L.No. 106–102, 113 Stat 1338, 1999), which provides the most comprehensive privacy protections for consumers to date (Gladstone, 2000). Once again, this highlights the differences in privacy policies between the United States and Europe.

There are provisions in the EU Privacy Directive that allow data processing if the data subject "unambiguously gives consent." A data subject can give his consent either by "opting in" or "opting out." A data processor may disclose on its Web site that personal data disclosed by the consumer may be further distributed for purposes of research or marketing. The Web site may offer the viewer a choice of not allowing the Web site to engage in this practice. This is called opting out. It is unclear whether opting out satisfies the criteria of "unambiguously giving consent" or whether only an opting-in alternative, whereby the Web site disseminates the data only if the viewer clicks "yes" to such a practice, is sufficient (Gladstone, 2000).

Article 7 also allows the processing of data that is needed in order to execute a task carried out in the public interest or for the exercise of the official authority vested in the controller. As in Article 3, the focus of the EU Privacy Directive is to protect data from abuse by private hands: The EU Privacy Directive deference to the government or an official authority is unlike the American approach, which seeks to minimize government involvement.

Articles 25 and 26 govern the transfer of personal data outside the EU to third countries such as the United States. The EU Privacy Directive bars the export of European personal data to countries that do not have "adequate" personal data protection regimes. The EU Privacy Directive sets out specific derogations whereby personal data may be exported despite adequate protection. These exemptions include situations where there has been consent by the subject, where it is necessary to the completion of a contract, and where it is in the public interest or in the vital interest of the subject (EU Privacy Directive). Information on the Internet crosses geographical and political borders on a continuous basis; therefore, all countries connected to the Internet are subject to the extraterritorial application of the EU Privacy Directive. Many U.S. companies have been concerned about the impact of the EU Privacy Directive's "adequacy standards" on their privacy policies and practices.

Safe Harbor Compliance

The initial response of the United States to comply with the "adequacy standards" of the EU Privacy Directive resulted in the establishment of the International Safe Harbor Principles. After protracted negotiations, the U.S. government and the European Commission jointly agreed to satisfy the "adequacy standards" of the EU Privacy Directive. Organizations that seek to benefit from the Safe Harbor Principles must self-certify their compliance with the U.S. Department of Commerce, thereby agreeing to terms of data handling practices, compliance, and dispute settlement. The Safe Harbor mechanism allows voluntary commitment by U.S. companies, which builds on the self-regulatory approach the United States takes toward privacy. To date, only 148 firms have joined the Safe Harbor Program (Aaron, 2002).

The Safe Harbor Principles are composed of two documents: a list of seven "critical" elements dealing with data processing, and a list several frequently asked questions

(FAQs). Organizations may choose to precisely adhere to the specific provisions of the Safe Harbor Principles to obtain the benefits of the Safe Harbor and publicly declare that they do so. Alternatively, organizations may develop their own self-regulatory privacy program, provided it conforms to the Safe Harbor Principles, or join a self-regulatory privacy program that adheres to the Safe Harbor Principles. Organizations that have agreed to comply with the Safe Harbor Principles are subject to the Section 5 "unfair and deceptive" practices of the Federal Trade Commission Act. Air carriers are subject to the equivalent statute of the Department of Transportation. There are several industries, such as the telecommunications and financial services industries, that are not eligible for the Safe Harbor or whose business practices are incompatible with the provisions.

The seven critical elements of Safe Harbor are notice, choice, onward transfer, security, data integrity, access, and enforcement. At least five of the critical elements directly address the provisions in Articles 2 through 14 of the EU Privacy Directive. From the perspective of assuring the same high level of privacy protection as the EU Privacy Directive, the Safe Harbor Principles appear to fall short in several areas, which again reflects the tradeoff between privacy and other fundamental rights in the United States. Under the Safe Harbor criteria of providing notice, individuals must be informed in clear and conspicuous language of the following: (a) the purpose for which an organization collects and uses information; (b) the types of third parties to which an organization discloses the information; and (c) how to make inquiries or complaints. This Safe Harbor requirement follows the requirements of Articles 10 and 11 of the EU Privacy Directive, but Articles 10 and 11 require that information be given to the subjects *before* the collection of personal information. The Safe Harbor notice requirement provides more leeway, because it stipulates that such notice may be given before personal information is received or "as soon thereafter as practical." This delayed notice was undoubtedly fashioned to encourage easy compliance with the EU Privacy Directive, but abuse of this option could effectively negate one of the key privacy protection mechanisms of the EU Privacy Directive. If an organization finds that prior notice is too costly, under the Safe Harbor Principles, they may opt routinely for giving notice after the fact. In addition, the Safe Harbor notice requirement does not require that data subjects be explicitly informed of their right of access to personal data (Gladstone, 2000). This, again, presents a disparity with the EU Privacy Directive, and it diminishes the underlying goals of the EU Privacy Directive.

Article 14 of the Directive grants the data subject rights to object to the processing of data that the controller anticipates as being processed for direct marketing and is incompatible with the purpose for which it was collected. The Safe Harbor Principles give individuals the opportunity to opt out in deciding how personal information will be disclosed to third parties for a purpose other than that for which it was originally collected. Under EU Privacy Directive, the data subject must be given the opportunity to object before the data is disclosed for the first time to a third party. The Safe Harbor opt-out choice falls short of the prior choice requirements in Article 14 of the EU Privacy Directive (Gladstone, 2000). Article 6 of the EU Privacy Directive sets out the principle of the quality of the data that may be collected and prohibits further processing of that personal data in a way that is incompatible with the "legitimate purpose for which it was collected." A parallel provision of Safe Harbor Principles entitled "Data Integrity" allows an organization more freedom in collecting and maintaining data, because it stipulates that an organization is not prohibited from the further processing of data as long as the organization takes "reasonable steps to ensure the data is accurate, complete and current."

Article 25 of the EU Privacy Directive prohibits the transfer of personal data to a third country that lacks an adequate level of protection, and Article 26 offers certain limited exceptions to this prohibition. The onward transfer provision of the Safe Harbor requires that disclosure of personal information to a third party be consistent with the principles of notice and choice; an organization is liable if it knows or has reason to know that a third party will process the information improperly. Under the Safe Harbor Principles, however, an organization is in compliance when transferring data to a third party even if the third party does not subscribe to the Safe Harbor Principles or the EU Privacy Directive as long as that third party signs an agreement to protect the data. This flexibility could easily create of a data haven and effectively subvert the EU Privacy Directive even if one were to comply with the Safe Harbor, thereby essentially directly frustrating the purpose of the EU Privacy Directive (Reidenberg, 2000).

Model Contract Clauses

In an effort to offer more flexibility for compliance with the EU Privacy Directive, in December 2001 the EU Commission adopted the Commission Decision 2002/16 (Contract Clause Decision), which sets out standard contract clauses for the transfer of personal data to processors in non-EU countries that have not been recognized as providing "adequate protection" for data. Under these standardized contract clauses, an EU company exporting data can and must treat the data with the full respect of the EU data protection requirements. The terms of these standard privacy contract clauses can be appended to existing licenses or contracts and offer a guarantee that the necessary security measures for privacy protection are in place.

The companies that choose to comply with the Safe Harbor provisions are subject to the enforcement jurisdiction of the U.S. Federal Trade Commission. Under the standard contract clauses, an entity becomes subject to the European Data Authority and the standard contract creates a private right of action in Europe. In addition, the terms of the standard contract are rigid and may impose criminal penalties.

The EU Privacy Directive, which has set the standard for privacy protection for international transfer of data, can be satisfied in several ways. The means of compliance that is adopted, and which can vary from joining the Safe Harbor to drafting model contract clauses, will depend on the nature of one's organization and how one's data is collected.

Technological Responses to Privacy Protection

For the part they play in the effort to protect privacy, the legal and self-regulatory instruments described above have legitimately received substantial public support. The fundamental flaw with these privacy programs and legislations is that most users of the Internet lack an understanding of the premise underlying the technological structure of computer networks, in particular the Internet. Loss of control over one's personal data or relinquishment of one's privacy is a direct result of the technology or infrastructure of the Internet. Network computer systems are designed to have identifiable transactions; every time a user logs onto the Internet, an electronic record is created (Lessig, 1991).

The ease with which privacy is sacrificed as a result of the openness of the infrastructure of the Internet was recently brought to the attention of the public, as the FBI has widely implemented a very efficient surveillance technology, Carnivore, that captures Internet conversations. Carnivore is a software technology that was developed to intercept e-mail messages based upon code words (Electronic Privacy Information Center, 2000). Carnivore is the term used for an entire system: a computer running on a Microsoft's Windows 2000 operating system and software that scans and captures packets, the standard unit of Internet traffic, as they travel through an Internet service provider's network. The FBI can install a Carnivore unit at the network station of an internet service provider (ISP) and configure it to capture e-mail going to or from the person under investigation. Under the USA PATRIOT Act, discussed below, in order to obtain a court order to install Carnivore, a law enforcement agent must simply certify to a judge that the information is "relevant to an ongoing criminal investigation." In addition to the fact that under the USA PATRIOT Act the warrant requirements are lowered, the packet-switching technology that drives the Internet has allowed the FBI to gather more information than a prescribed search warrant would allow. This compromises the privacy of all persons who have any interaction with the targeted suspect. Questions have been raised regarding the mechanics of the Carnivore system. When the Electronic Privacy Information Center filed a Freedom of Information Act (FOIA; 5 U.S.C. 552 1994 & Supp. 1999), request for Carnivore's source code, the inner workings of how the device functioned, the FBI refused to disclose information about how Carnivore worked. The persistent refusal for full disclosure by the FBI has led to several lawsuits against the them, which has created the impression in the eyes of the public that law enforcement is taking away rights rather than protecting them (Van Bergen, 2002).

Any discussion of privacy must recognize that a balance must be struck between interest of privacy and security. The swiftness with which the USA PATRIOT Act recently passed through Congress illustrates that Americans do not highly value their privacy when national security is threatened. The U.S. Department of Justice has been arguing with civil liberties groups and privacy advocates for several years over amendments to federal statutes that would expand law enforcement wire tapping and electronic surveillance operations. Congress had been reluctant to expand law enforcement surveillance activities with respect to the Internet, usually citing privacy concerns. After the September 11, 2001, terrorist attacks, congressional reluctance and public opposition, as measured by consumer polls, to expanded surveillance diminished.

The USA PATRIOT Act is a long and complex statute that made changes to over 15 different statutes, several of which directly impact Internet communications. Section 216 of the Act addresses pen/trap orders, which were initially defined under the Electronic Communications Privacy Act (ECPA) (18 U.S.C. 3121–3127) as including a device attached to a telephone line to trace and trap telephone numbers. Since about September 2000, the FBI had been routinely applying pen/trap devices to computer communications, and thus, under the USA PATRIOT Act, the pen/trap provisions apply to computer communications so that all e-mail addresses, Web addresses visited by a target, Internet protocol addresses, and other routing information can be obtained. The contents of the messages have never been retrievable under a pen/trap order, but while telephone numbers can be easily separated from a telephone message, e-mail addresses are not so easily separated from e-mail contents or e-mail subject headings in particular, and therefore pen/trap orders may be misused in the e-mail context.

The USA PATRIOT Act amends the Foreign Intelligence Surveillance Act (FISA) (50 U.S.C. 1841–1846), which historically had separated domestic criminal investigations from foreign investigations. Domestic surveillance was governed by Title III of the Omnibus Crime Control and Safe Streets Act of 1968 (18 U.S.C. 2510–22), which provided for adequate safeguards for basic constitutional rights, such as the Fourth Amendment probable cause requirements and judicial review. Foreign intelligence, which was governed by FISA, granted the attorney general the power to treat an alien as an agent of a foreign power, which meant that person was not entitled to constitutional rights. The boundaries between these two laws are blurred under the USA PATRIOT Act, which is most evident from the expanded definitions of "terrorist." Consequently, under the USA PATRIOT Act's pen/trap provisions, an ISP must respond to a court order as long as the law enforcement agent certifies that the surveillance "is relevant to an ongoing investigation." The Fourth Amendment requirements of probable cause when conducting wiretaps have been lowered. In addition, pursuant to the same lowered standard, any business may now be served with an order for the production of "any tangible thing," not just a business record. Clearly, ISPs, cable subscribers, and businesses in general ought to review their privacy policies and confidentiality agreements to ensure that they accurately reflect their new obligations under the USA PATRIOT Act.

There are significant differences in the regulations surrounding the protection of personal information in the United States and in Europe. These differences will likely impede global e-commerce and international agreements, and negotiations are needed to enable nation states' policies to develop in harmony. Reidenberg (2001) has suggested the promotion of negotiations for a General

Agreement on Information Privacy (GAIP) in connection with the World Trade Organization (WTO). He recommends this type of a treaty organization because it places data protection in a trade arena rather than a political arena, and by placing GAIP within the WTO, it would add social protection norms to a trade treaty. The GAIP would include many signatory countries and would focus on an institutional process of norm development to facilitate easily implemented standards for informational privacy. The WTO could define cost standards for data protection that could be incorporated into a multilateral trade agreement (Reidenberg, 2001).

ENCRYPTION AND ELECTRONIC SIGNATURES

The expansion of Internet technology has allowed people to use the Internet as a platform for worldwide business, and therefore, contracts are being formed among parties with no prior relationships. Parties want to rely exclusively on online communications; therefore, a reliable system of authentication is needed to ensure the success of e-commerce (Winn, 2001). The ability to transmit information secretly over distances has been accomplished for decades using cryptography. Digital signatures and encryption are the two key aspects of cryptography that have been recognized as essential tools for security and trust for e-commerce. This section begins with an explanation of the current global regulation regarding the control of exportation of encryption. The most popular form of encryption in use today is public key asymmetric cryptography (PKI), which is used in digital signatures. *Electronic Signatures-Technical Overview* distinguishes digital signatures from electronic signatures before offering an overview of the PKI technology. *Regulatory Models* explains the three legislative models that have been developed for the regulation of electronic signatures.

Cryptography is the art of using code to keep information secret, and encryption is the technique of encoding or scrambling communications. Most nations regulate the exportation of encryption technology because of the fear that abuse of the technology by terrorists or criminals would impede the ability of national security and law enforcement to do their jobs. Privacy advocates and free speech proponents in the United States agree that restrictions on exports of encryption infringe on individuals rights to informational privacy, Fourth Amendment and First Amendment rights. Members of the high-tech and software industries complain that such restrictions are anticompetitive vis-a-vis foreign nationals.

Encryption Exportation Regulation; United States Law and International Treaties

Under the Arms Export Control Act of 1978 102, 22 U.S.C.S. 2799aa-1 (2000; the Arms Export Control Act or the Export Administration Act [EAA]), the U.S. State Department decides whether an item is dual purpose, a category that includes such things as commercial products with military application. Control over the licensing and export of dual-purpose products is transferred to the Department of Commerce (DOC; Paik, 2000). The DOC under the EAA now regulates the export of all general-purpose encryption devices and software. The DOC Export Administration Regulations 15 C.F.R. 730–74 (2000) include source code and object code in the definition of software subject to regulation and exportation, which includes "downloading or causing the downloading of such software . . . or making such software available from electronic bulletin boards . . . " (Paik, 2000). During the Clinton Administration, there was significant discussion regarding the requirement that software companies create key recovery systems. One that was recommended by the government was known as the Clipper Chip. That requirement was not adopted, and current regulations allow U.S. citizens to ship any retail encryption product around the world to commercial concerns after a one-time technical review by an interagency panel (Paik, 2000).

The Wassenaar Agreement is an international agreement that addresses controls on encryption exports, and to which 33 countries, including the United States, subscribe. The Wassenaar Agreement, which was designed to promote cooperation among its members, was amended in 1998 to impose export control on export software for keys above 64 bits and to eliminate record keeping for low-level encryption. Several countries, including Israel, South Africa, India, and China, however, are not members. Thus, there is not harmonization in the global encryption export market (Paik, 2000).

Electronic Signatures—Technical Overview

An electronic signature is any method that logically associates an electronic representation of the identity of a person with the content of an electronic document or record. It implies acknowledged authorship or agreement, but there are times, with e-mail programs for instance, where an electronic signature can be applied automatically. Generally, with electronic records the goal is to protect the integrity of the content, and a generic electronic signature provides very low assurance that documents have not been altered. Digital signatures, a special subset of electronic signatures, can provide this assurance (Ballon, 2001).

Digital signatures serve three essential functions: authentication, integrity, and nonrepudiation. Authentication means that the party is who she says she is. Integrity is assurance that the communication has not been tampered with (i.e., it is in its original form), and nonrepudiation prevents the party from retreating from the transaction in the event a dispute arises. A digital signature denotes "an electronic imprint" that is created using PKI. PKI is also referred to as asymmetric encryption because two keys are used, a private key and a public key. The two keys are mathematically related, so when the message is encrypted with the private key and sent to the recipient, the recipient must use the sender's public key to decrypt it. Each user has a different public/private key pair (Berman, 2001).

A digital signature is not a digitized version of a person's handwritten signature; rather, it is a "message digest" of the document that is being encrypted, or sent. The message digest is created by processing the document through a unique computer-generated code known as a "hash."

Once the hash is created, the signer types in a personal identification number (PIN) that allows the private key to generate a long series of numbers and letters, which is the digital signature. The sender uses a one-way hash function to encrypt the message he wants to sign and then sends it. The computer-generated signature and the hash result are unique to the message. Every time the message passes through the hash function, the same message digest will be produced. To verify the signature of a digitally signed message, the receiver reverses the process with the public key (Berman & Biddle, 2001).

In order for a digital signature system to function, the parties must be assured that the public keys they obtain actually belong to the person it is purported to be from and not to a forger. The way to achieve this confidence is with a Certification Authority, which is an entity either public or private that attests to the integrity of the system. Certification Authorities issue certificates of authenticity as to the ownership of the keys. There have been several proposals as to the best solution to the problem of authentication (Zemmick, 2001). Currently, several banks offer this service.

The intention in all jurisdictions that have enacted electronic signature laws is to encourage the development of e-commerce, but there is a large disparity in their treatment of electronic signatures. According to Smedinghoff and Bro (1999), "predictability is a watch word for the growth of commerce and law can play a role in providing this valuable commodity," and yet the most striking feature of the various electronic signature laws enacted around the world is their lack of uniformity (Fischer, 2001). As we saw with the various privacy policies, failure of policy makers to remove national barriers will hinder successful global e-commerce and possibly widen the digital divide.

Regulatory Models

Electronic signature legislation can be seen as based on one of three models. The first of the three models is known as the mandatory or prescriptive model because it mandates a specific technology (Fischer, 2001). Alternatively, the minimalist legislative approach is technology neutral, and the third hybrid approach suggests a favored technology that affords presumptions under the law. Proponents of the prescriptive model, which include Germany, Italy, Malaysia, and Russia, mandate specific technology when authorizing electronic signatures as an alternative to pen and paper. Public key infrastructure, or a digital signature, which is currently the most sophisticated technology, is required. In addition, this mandatory or prescriptive approach often outlines specific criteria for the trusted third party, or Certification Authority. The rationale given for this approach is that the requisite security for e-commerce can only be obtained with these constraints. In addition, it is believed that these requirements will assure legal certainty, which is essential for public trust. Critics of this approach point out that this not only grants economic advantages for a particular existing technology, it is shortsighted because although the "best" technology in 2002 may be PKI, there is no guarantee that better, more sophisticated techniques will not become available.

In addition, this mandatory or prescriptive approach often outlines specific criteria for the trusted third party (CA) and usually overly limits the liability of the CA. Typical prescriptive digital signature laws place the burden for loss or theft of a private key on the consumer if there was a failure to exercise reasonable care. The consumer will bear the entire loss, thus insulating the CA from any liability. This structure of liability seems to unfairly burden the consumer in an effort to create a less risky role for the CA, whereas the CA could more efficiently protect itself.

The hybrid model, adopted under the EU Signature Directive of 1999 by the European Union, is often referred to as a two-tier model because it grants basic validity to all electronic signatures but provides special treatment to certain advanced signatures. Under this hybrid approach, an electronic signature cannot be denied legal effectiveness solely because it is electronic, but some technologies are given presumptions of authenticity if the signature meets certain requirements. Currently, the only technology that meets these heightened standards is PKI.

The rights and duties set out for the parties to an electronic transaction under the hybrid model reflect a market-driven philosophy. Unlike under the prescriptive approach, CAs will be found liable in damages for harm caused to someone who has reasonably relied on a certificate for the accuracy of the information, unless it can be proven that the CA was not negligent. The presumption is in favor of the consumer. Under the hybrid approach, CAs can limit their liability by contract, however, prior to entering into the transaction.

The minimalist, wholly technology-neutral approach provides that no electronic signature of whatever type may be denied legal effect, validity, or enforceability because it is in electronic form. The U.S. Electronic Signatures in Global and National Commerce Act of 2000 (E-SIGN), 15 U.S.C. 7001–06, 7021, 7031, endorses this approach. Australia and the United Kingdom have also enacted minimalist legislation, and New Zealand is considering a law similar to that of Australia (Fischer, 2001). The philosophical principles upon which it is based foster technological advancements by allowing the market to decide which technology is best. It also allows several systems to be developed simultaneously. E-SIGN was drafted and enacted swiftly in the United States, which was in part a response to the factious division among the several states within the U.S. in adopting various versions of electronic signature legislation.

The major criticism of E-SIGN is that it is too vague and that its lack of certainty will hamper electronic commercial growth. Critics of E-SIGN suggest that the nonrestrictive legislation might lead to parties being held liable for contracts they did not actually authorize, for example if a party somehow failed to protect the security of his signature device. Although the liability may be severe, the "liberty to contract" concept prevails, which leaves the parties free to be bound or reject the contract.

The electronic signature legislation that has been passed in countries around the world reflects the contrasting views of the minimalist and the prescriptive approach. The global initiative by the United Nations Commission on International Trade Law (UNCITRAL) working group on electronic commerce is not expected to have a significant impact on existing or proposed legislation, because business lost confidence in the proposal during negotiations as a result of the parties' insistence on a prescriptive approach mandating a public key infrastructure (Fischer, 2001). The future of commerce will not take place on paper, and it is important for the law to grow to facilitate the development of e-commerce. Attention needs to be given to harmonize the differences between nations that exist in the area of electronic signatures, possibly through a revising of the work previously produced by UNCITRAL.

INTELLECTUAL PROPERTY

Technology has had a profound impact on the perception of intellectual property and the appropriate distribution of the rights traditionally attached to trademark, patent, and copyright in the global context. Although technological advances have created unprecedented opportunities for economic prosperity in the area of intellectual property, legal systems have had to adapt to maintain firm standards while fostering financial growth. Consistency in legal paradigms across national borders, which is crucial to establishing the Internet as a reliable conduit for successful global commerce, has been relatively successful in the area of intellectual property. International intellectual property conventions have been consolidated under the auspices of Convention Establishing the World Intellectual Property Organization (WIPO). The Agreement on Trade Related Aspects of Intellectual Property (TRIPs) has formed a WTO–WIPO union by integrating much of WIPO's law into WTO's trade regime (Mort, 1997). Trends in practice also suggest a tendency toward uniformity. This section reviews substantive provisions of the WIPO Copyright Treaty, key terms of the Digital Millennium Copyright Act (DMCA), the European Copyright Directive, and the Electronic Commerce Directive, particularly as they reflect the guidelines for ISP copyright liability. Digital transmissions allow infringers to obtain and disseminate information quickly without being detected. The Internet presents difficult challenges for copyright owners, who want to identify and stop infringers. Locating a financially sound ISP to end the copying activity is the surest route for the copyright owner. This section further examines the discrepancies that exist between the United States and the EU's approach to database protection and business method software protection.

ISP Liability for Third-Party Copyright Infringement

Intellectual property conventions, which predated WIPO's creation in 1967, historically operated independently without institutional oversight (Mort, 1997). WIPO, which was designed as a specialized agency to administer major international conventions under the leadership of the United Nations Director General Secretariat, had difficulty enforcing rights and resolving conflicts. Serving as the sole international authority for over two decades, WIPO lacked the necessary enforcement powers to eliminate piracy of intellectual property. In 1986, as part of the Uruguay Round Negotiations, intellectual property protections were integrated into the General Agreement on Tariffs and Trade (GATT). In 1994, TRIPs was established and a symbiotic institutional relationship between the WTO and WIPO was formed (Mort, 1997). There was a simple integration of intellectual property protection into a trade-based sanction regime. In 1995, a cooperative agreement was signed between the two bodies to coordinate their efforts.

In 1996, WIPO concluded two treaties covering the protection of copyright and rights in digital environments, and the names of those treaties are WIPO Copyright Treaty and WIPO Performances and Phonograms Treaty. The WIPO Copyright Treaty established a distribution, rental, and communication right in creative works to the public. This distribution right under the WIPO Copyright Treaty may be accomplished through sale or other means of transferring ownership, but the right is limited to fixed tangible copies capable of circulation (Soma & Norman, 2000). There was no agreement between the delegates as to the scope of the doctrine of exhaustion for "first sale" rights, so that was left to be defined by each adopting nation. Another important provision of the WIPO Copyright Treaty related to the exhaustion of rights is that the right of public communication permits copyright holders to make their works available by wire or wireless means. Included in this right is the ability to make works available to the public so they can access them as they choose. In the event communication permits recipients to reproduce a tangible copy, national law must define liability for infringement.

The implications of this broad access right is that ISPs could be liable for direct and contributory copyright infringement causes of action that may be brought by the copyright owner against the ultimate recipient. The right created in the WIPO Copyright Treaty leaves details about liability for third-party copyright infringement to the contracting parties. This has caused much concern among telecommunications companies and ISPs, because such a broad interpretation of the treaty could lead to lawsuits from copyright owners (Mort, 1997).

Under case law in the United States, the issue of determining ISP liability for third-party copyright infringement starts with an examination of the type of service the ISP provides in relationship to the infringement claim. The categorization of the ISP is based upon the level of knowledge the ISP had of infringing activity, the control of the ISP, the length of time the material is stored on the ISP server, and any financial benefit received by the ISP. When defined according to function, one will analyze and apply liability to the ISP based upon the traditional common carrier versus publisher/distributor model (Soma & Norman 2000).

In the United States, the Digital Millennium Copyright Act limits ISP liability for third-party copyright infringement where the ISP complies with a detailed system of notice and removal. Where an ISP acts as a mere conduit for data, the DMCA will limit liability of the ISP relating to these transitory communications.

The recently passed EU Copyright Directive follows a logic similar to that of the DMCA in that it exempts ISPs from liability where they play a passive role as a mere conduit of information from third parties. An ISP cannot modify the work in any manner; must comply with industry standards for transmission and storage, and must remove infringing materials expeditiously in order to avoid copyright infringement liability. A third-party copyright infringement case would not be successful unless the ISP had been warned to remove it and did not do so (McDonald, 2001). Finally, the Electronic Commerce Directive, which is similar to the DMCA and the EU Copyright Directive, sets out guidelines for liability for ISPs where they play a passive role as a mere conduit of information from third parties. Similarly, the Electronic Commerce Directive limits ISPs from liability for other intermediary activities, such as storage of information or caching.

Databases

Database protection presents a controversial area of legislation in the global arena. A database, which is a compilation of information, is not protected under copyright law in the United States unless the arrangement or selection rises to a sufficiently high level of originality or uniqueness in its selection or arrangement. Database protection was limited by the United States Supreme Court in *Feist Publications Inc. v. Rural Tel. Serv. Co*., 499 U.S. 340 (1991), where the court held that a white pages directory, which consisted of preexisting factual material, lacked the requisite originality in selection coordination and arrangement of data to garner copyright protection. The simple listing of subscribers in alphabetical order by surname lacked originality, despite the excessive time, effort, and energy expended to organize it. Feist struck down the "sweat of the brow" doctrine, which some U.S. courts had used to find copyright protection. Opponents of greater legal protection for databases believe that the balance between control over information and allowing information into the public domain is best met with few legal restrictions imposed. "Information is meant to be free" is the ideology expressed in support of a reduced role of government. The European countries believe that there are sound economic justifications for affording protection to owners of databases based upon the "sweat of the brow" theory and in 1996 passed the EC Directive No. 96/9, on the Legal Protection of Databases (the Database Directive). The sui generis protection granted to databases is based under the Database Directive on a property concept, which bestows exclusive rights of ownership to the database compiler. The provisions to afford database protection was dropped from the WIPO Copyright Treaty over objections by members of the academic and scientific community, but in the United States, debate over this issue remains. Several legislative bills have been introduced in Congress based upon the belief that such protection would make the United States more competitive in foreign markets. None of these bills has passed as of the time of the this writing.

Software Patents

The topic of patenting software and, in particular, business methods has received a lot of attention from the pubic due to the economic success of several e-commerce entrepreneurs (Gladstone, 2002). The EU and American approach to protection of software has traditionally been reported as divergent, but on closer examination, the two policies appear to be converging, particularly in light of the decline of the heated electronic commerce boom. Historically, intellectual property and software was limited to copyright protection on the grounds that it was written in code, thus it was a literary work, and hence copyright protection was appropriate. In *Diamond v. Diehr,* 450 U.S. 175 (1981), however, the U.S. Supreme Court held that a process for monitoring the temperature inside a synthetic rubber mold using a computer and the Arrhenius equation for measuring cure time as a function of temperature and other variables was patentable subject matter. The court focused on the "postsolution activity" that resulted from the computer program (Cohen & Lemley, 2001). The lesson from the Diehr decision was to include a physical element or step in any future patent application that might recite a "mental process." This theory was adopted and clarified by the Federal Circuit in *In re: Alappat,* 33 F.3d 1526 (1994), which established that an "otherwise statutory process or apparatus requirement may be satisfied by drafting claims to include a general purpose computer or standard hardware or memory element that would be necessary for any useful application of the algorithm." The logic of these cases did not include claims reading on computer programs themselves, as opposed to programs implemented in a machine or system (McDonald, 2001). This obstacle was overcome in the *In re: Boureguard,* 53 F.3d 1583 (Fed Cir., 1995). While an appeal was pending, the U.S. Patent and Trade Office established that software stored in memory media is patentable as an article of manufacture.

Article 52 of the Convention on the Grant of European Patents, October 5,1973, 13 I.L.M. 270, the European Patent Convention (EPC) indicates that computer programs "as such" are not patentable and that programs for computers shall not be regarded as inventions. The European Patent Office (EPO) estimates, however, that they have issued over 20,000 patents on computer programs. In late 2000, there was a diplomatic conference to revise the EPC to change Article 52, bringing it in line with actual practice and with TRIPs, making it clear that patent protection would be available to technical inventions of all kinds (McDonald, 2001). This measure did not pass, however. The reversal and decision to hold onto the old position is interesting in that it reflects a rebellion against the rushed commercialization reflected in the e-commerce boom.

Critical to any discussion of intellectual property rights is finding the correct balance between creating incentives to encourage innovation by inventors and the public's right to access to information and knowledge. The trend in the United States to grant exclusive patent rights to software involving Internet technology continued unabated to the point where the U.S. Patent Office was granting patents on "how to get a business method patent." Beginning with the case of *State Street Bank & Trust vs. Signature Financial,* 149 F.3d 1368 (Fed Cir., 1998), where Signature Financial was granted a patent for a data processing system to implement an investment structure, the court endorsed business methods providing they comply with other requirements for a patent, thus laying to rest "the ill conceived exception to the law" that business methods were not patentable. U.S. patents have been issued on numerous technological processes; the "single click" patent covering a method and system for placing purchase orders via a commercial network that was granted to Amazon.com and contested by Barnesandnoble.com, Inc is one of the more controversial and publicly known cases to be litigated. Although Amazon.com initially was granted an injunction to prevent Barnesandnoble.com, Inc from using the process, the patent is still being challenged (Shulman, 2000).

The trend in the United States to grant patents on software for business methods informed the discussion within the Diplomatic Conference of the European Commission regarding their policy toward granting patents on software. It is likely that the concern over the inseparability of business methods from software patents in general may have encouraged no change in the EPC. Proponents of LINUX, the open source software, which encourages sharing of ideas to promote innovation, began campaigning against software patents in general in Europe in the late 1990s. This development may also have influenced the change in the EPC outcome (Mcdonald, 2000).

The economic downturn of the late 1990s that hit technology companies and Internet startups particularly harshly may have contributed to the slowdown in Internet-related business method patent filings, but the curtailment of the public's endorsement of companies whose sole or main asset was a business method patent was also a key factor. Empirical evidence, which demonstrates the withdrawal of funds from these "idea factories" (Shulman, 2000), suggests that the flurry of business method patents may not have been based on solid economic grounds. The multibillionaire entrepreneur Jay Walker, whose company, Walker Digital, claimed 70 business method patents, with 400 pending before the U.S. Patent Office, needed to lay off 80% of its workforce within a few short years of establishing itself (Shulman, 2000). The profusion of new software business method patents was exciting, but, in fact, they had a chilling effect on e-commerce, and when put to the Wall Street test, most of these companies did not fare very well. Although business method patents are still often being applied for, these are offensive or defensive acts taken to prevent others from gaining market share, rather than with an expectation of employing the patent. The recent decline in enthusiasm of business method software patents in the United States suggests that the EU position, which has been to proceed cautiously before modifying laws to broaden the individuals rights at the expense of the public's access to information, may be the better approach to follow (Gladstone, 2002).

CONCLUSION

The Internet reaches around the globe and it may be unrealistic to expect symmetry between nations, but policy makers must continue to strive for a common ground. There is sufficient legal and empirical evidence to support a nation asserting jurisdiction and enforcing its laws beyond its borders, but enforcement jurisdiction and questions of comity present additional difficulties, as we have seen recently with the internationally recognized Yahoo! case. In the areas of privacy and electronic signatures, discrepancies between countries remain apparent, and these variances are rooted in fundamental cultural, social, and philosophical differences. The Internet has become a medium for widespread commercial activity, but further growth will require agreement regarding mechanisms to regulate in all areas of human activity. Global consensus to limit ISP liability for copyright infringement, the reduction in the business method patent application surge, and the reluctance in the United States to pass a sui generis database protection law suggest a concerted global effort to not limit public access to information and to allow the Internet to serve as a conduit of knowledge dissemination. There is a clearly a trend toward common ground in several areas of intellectual property law in cyberspace. Harmonization in all areas of the law is the goal to strive for, because without seamless predictable systems, businesses and consumers will be reluctant to enter into transactions.

GLOSSARY

Cyberspace Where messages and Web pages are posted for everyone in the world to see. In Reno v. ACLU, 117 S.Ct. 2329 (1997), the first opinion about the Internet by the United States Supreme Court, it was stated that "[t]aken together, these tools constitute a unique medium—known to its users as 'cyberspace'—located in no particular geographical location but available to anyone, anywhere in the world, with access to the internet."

Cybersquatting Where a person or an entity registers a domain name with no intention of using the name for any purpose but to thwart the ability of the rightful owner or trademark holder to obtain the name.

Data Protection The right provided for under the Council Directive 95/45/EC of the European Parliament and of the Council of the European Union of Oct. 24 (1995); also know as the EU Privacy Directive.

Encryption The conversion of data into cipher text, a form of text not easily understood by unauthorized people.

Electronic Signatures in Global and National Commerce Act (E-SIGN) A U.S. law that was passed in 2000 that recognizes the legal effect of an electronic signature in whatever form it is made.

International Safe Harbor Principles The regulatory response of the U.S. government to the "adequacy standards" of the EU Privacy Directive agreed on by the European Commission.

Privacy The right to be left alone.

CROSS REFERENCES

See *Copyright Law; Cyberlaw: The Major Areas, Development, and Provisions; Digital Signatures and Electronic Signatures; Encryption; Patent Law; Privacy Law; Trademark Law.*

REFERENCES

Aaron, D. (2002). *EU Privacy Report. The Second Annual Privacy and Data Security Summit*. Retrieved August 14, 2002, from http://www.privacyassociation.org/docs/aaron-020201.pdf

ACLU v Reno, 521 U.S. 844 (1997).

Alexander, M. (2002). The first amendment and problems of political viabilty; The case of pornography. *Harvard Journal of Law and Public Policy, 25,* 977.

Arms Export Control Act of 1978 102, *22 U.S.C.S. 2799aa-1* (2000).

Ballon, C. I. (2001). Encryption and cryptography, in *E-Commerce and Internet Law: A Legal Treatise With Forms*. Little Falls, NJ: Glasser Books.

Berman, A. B. (2001). Note: International divergence: The "KEYS" to signing on the digital line—The cross-border recognition of electronic contracts and digital signatures. *Syracuse Journal of International Law and Commerce, 28,* 125.

Boam, C. P. (2001). The Internet, information and the culture of regulatory change: A modern renaissance. *CommLaw Conspectus, 9,* 175.

Brussels Convention (1998). *Convention on jurisdiction and the enforcement of judgments in civil and commercial matters, Sep. 30, 1998 OFFICIAL J.C027, 0001–0027*.

Brussels Regulation (2001). *Council regulation (EC) 44/2001 of 22 December 2000 on jurisdiction and the recognition and enforcement of judgments in civil and commercial matters, O. J.(L 12).1*.

Cable News Network, L. P. GoSMS.com, Inc, 2000 WL 1678039, Southern District of New York.

Calder v Jones, 465 U.S. 783, 1984.

Child Online Protection Act (COPA), Pub. L. No. 105–277, 112 Stat. 2681–736 (1998).

Children's Internet Protection Act (CIPA), P. L. No. 106–554, tit. Xii, 114 stat. 2763, 2763A–335 (2001).

Cohen, J. E., & Lemley, M. A. (2001). Patent scope and innovation in the software industry. *California Law Review, 89,* 1.

Common Decency Act (CDA), Pub. L. No. 104–104, tit. V, 110 Stat. 56, 133 (1996).

Convention Establishing the World Intellectual Property Organization (WIPO), July 14, 21 U.S.T. 1770, 828 U.N.T.S. 3 (WIPO) (1967).

Convention on the grant of European patents, October 5, 13 I.L.M. 270 (1973).

Diamond v. Diehr, 450 U.S. 175 (1981).

Digital Millennium Copyright Act, Pub. L. No. 105–304, 112 Stat. 2877 (1998).

Electronic Communications Privacy Act (18 U.S.C. 3121–3127) (1986).

Electronic Frontier Foundation (2001). *EFF analysis of the provisions of the USA PATRIOT Act*. Retrieved May 8, 2002, from http://www.eff.org/Privacy/Surveillance/Terrorism_militias/20011031_eff_usa_pat.riot_analysis.html

Electronic Privacy Information Center (2000). *Evolution of carnivore.* Retrieved May 8, 2002, from http://www.epic.org/privacy/carnivore/#documents

EU Privacy Directive (1995). Council directive no. 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, art. 2(c). *Official Journal of the European Community,* L281/31.

European Commission (2002). *Decision 2002/16 EC of 21 Dec 2001, contract clause decision*. Retrieved August 14, 2002, from http://europa.eu.int/comm/internal_market/en/dataprot/modelcontracts/02–16_en.pdf

European Commission (2002). *Explanatory memorandum to amended proposal for Council Brussels regulation, COM. 689 of 26 October, 2000*.

European Commission (1999). *Explanatory memorandum to the proposal for a council regulation on jurisdiction, COM 348 of 14 July 1999*. Retrieved August 14, 2002, from http://www.europa.eu.int/comm/justice_home/pdf /com1999-348-en.pdf

European Council (2000). *Council directive 00/31 of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the International Market, O.J.L. 178/1*.

European Council (2001). *Copyright directive, Council Directive 01/29, O.J.(L 167/10)*.

European Council (1996). *Directive no. 96/9, O.J.L. 77/20, on the legal protection of databases*.

European Parliament (2000). *Council directive 1999/93, 2000 O.J.(l 13).12 at Art 2 (20), E.U. Signature Directive.*

Financial Services Modernization Act of 1999, Gramm-Leach-Bliley Act, Pub. L. No. 106–102, 113 Stat 1338 (1999).

Federal Trade Commission Act, 15 U.S.C. 45 (1979).

Fischer, S. F. (2001). Saving Rosencrantz and Guildenstern in a virtual world? A comparative look at recent global electronic signature legislation. *Boston University Journal of Science and Technology Law, 7,* 229.

Foreign Intelligence Surveillance Act, 50 U.S.C. 1841–1846 (1978).

Freedom of Information Act, 5 U.S.C. 552 (1994; Supp. 1999).

Fromholz, J. M. (2000). *The European Union Data Privacy Directive*. *Berkeley Technology Law Journal, 15,* 461.

General Agreement on Tariffs and Trade, Oct. 30, 1947, 61 Stat. A-11, T.I.A.S. 1700, 55 U.N.T.S. 194, art. IX (1947).

Geist, M. (2001). *Is There a There There: Toward Greater Certainty for Internet Jurisdiction, 661 PLI/Pat 561*. Washington, DC: Government Printing Office.

Gladstone Alpert, J. (2000a) Survey of cyberspace law: An introduction: Keeping pace. *Business Lawyer, 56,* 1.

Gladstone Alpert, J. (2000b). The U.S. privacy balance and the European privacy directive: Reflections on the United States privacy policy. *Willamette Journal of International Law and Dispute Resolution, 7,* 10–31.

Gladstone Alpert, J. (2002). Why patenting information technology and business methods is not sound policy: Lessons from history and prophecies for the future. *Hamline Law Review, 25,* 2.

Godwin, M. (2001). *ALERT: Ask Congress to legislate to improve security not eliminate freedoms*. Retrieved August 14, 2002 from http://www.eff.org/effector/HTML/effect14.24.html#I

Gutnick v. Dow Jones, VSC 305, 2001.

Hanson v Denckla, 357 U.S. 235, 1958.

In re: Alappat, 33 F.3d 1526 (1994).

In re: Boureguard, 53 F.3d 1583 (Fed Cir., 1995).

International Safe Harbor Rules (1999). Retrieved May 8, 2002, from : http://www.ita.doc.gov/td/ecom/shprin.html

Johnson, D. J., & Post, D. (1996). Law and borders—The rise of law in cyberspace. *Stanford Law Review, 48,* 1367.

Lessig, L. (1991). *Code and other laws of cyberspace*. Washington, DC: Government Printing Office.

McDonald, B. (2001). *International intellectual property rights*. Retrieved May 8, 2002, from http://www.wrf.com

Mody, S. S. (2001). National cyberspace regulation: Unbundling the concept of jurisdiction. *Stanford Journal of International Law, 37,* 365.

Mort, S. A. (1997). The WTO, WIPO and the Internet: Confounding the borders of copyright and neighboring rights. *Fordham Intellectual Property, Media and Entertainment Law Journal, 8,* 173.

Organization for Economic Cooperation and Development. *Guidelines on the protection of personal data*. Retrieved from http://www.oecd.org/dsti/sti/it/secur/prod/PRIV-EN.HTM#4

Paik, J. L. (2000). NOTE: The encryption export tax: A proposed solution and remedy to the issues and costs associated with exporting encryption technology. *Cornell Journal of Law and Public Policy, 10,* 161.

Panavision Int'l. L.P v. Toeppen, 141 F.3d 1316 (9th Cir. 1998).

Pastore, M. (2002). *Economic downturn slows B2B commerce*. Retrieved May 8, 2002, from http://cyberatlas.internet.com/markets/b2b/article/0,1323,10091719571, 00.html

Piera, F. (2001). International electronic commerce: Legal framework at the beginning of the XXI century. *Currents: International Trade Law Journal, 10,* 8.

Reidenberg, J. R. (2000). Symposium: Cyberspace and privacy: A new legal paradigm? Resolving conflicting international data privacy rules in cyberspace. *Stanford Law Review, 52,* 1315.

Reidenberg, J. R. (2001). E-Commerce and Privacy Institute for Intellectual Property and Information Law Symposium: E-commerce and trans-Atlantic privacy. *Houston Law Review, 38,* 717.

Restatement (3rd) of the Foreign Relations Law of the U.S. Sections, 304, 401, 402, 421, 431 (1987).

Rice, D., & Gladstone, J. (2003). An assessment of the Effects Test in determining jurisdiction in cyberspace. *Business Lawyer 58*(2).

Shulman, S. (2000). *Software patents tangle the Web*. Retrieved May 8, 2002, from http://www.technologyreview.com/articles/shulman0300.asp

Soma, J. T., & Norman, N. A. (2000). International takedown policy: A proposal for the WTO and WIPO to establish international copyright procedural guidelines for Internet service providers. *Hastings Communications and Entertainment Law Journal, 22,* 391.

Smedinghoff, T., & Bro, R. (1999). Moving with change: Electronic signature legislation as a vehicle for advancing e-commerce. *John Marshall Computer and Information Law, 17,* 723.

S.S. Lotus, (Fr.v. Turk.) 1927 P.C.I.J. (Ser A) No. 10 (19227).

State Street Bank & Trust vs. Signature Financial, 149 F.3d 1368 (Fed Cir., 1998).

Title III of the Omnibus Crime Control and Safe Streets Act of 1968 (18 U.S.C. 2510–22) (1968).

TRIPs agreement: Agreement on trade-related aspects of intellectual property rights, Apr. 15, 1994, Marrakesh agreement establishing the World Trade Organization, Annex 1C, legal instruments—Results of the Uruguay round, vol. 31, 33 I.L.M. 81 (1994).

United States electronic signatures in global and national commerce act of 2000 (E-SIGN), 15 U.S.C. 7001–06, 7021, 7031 (2000).

Uniting and strengthening America by providing appropriate tools to intercept and obstruct Terrorism Act of 2001, Pub. L. No. 107–56, 115 Stat 272 (2001).

Van Bergen, J. (2002). *Repeal the USA PATRIOT Act*. Retrieved August 12, 2002, from http://truthout.com/docs_02/04.05D.JVB.Patriot.htm

Winn, J. K. (2001). The emperor's new clothes: The shocking truth about digital signatures and Internet commerce. *Idaho Law Review, 37,* 353.

What is the Wassenaar arrangement. Retrieved July 17, 2001, from http://www.wassenaar.org/docs/talkpts.html

WIPO Copyright Treaty, Geneva, December 20, 1996 (1996). Retrieved May 8, 2002, from http:// www.wipo.org/eng/iplex/index.htm

WIPO Performances and Phonograms Treaty, Geneva, December 20, 1996 (1996). Retrieved May 8, 2002, from http:// www.wipo.org/eng/iplex/index.htm

World-Wide Volkswagen Corp. v. Woodson, 444 U.S. 286, 1980.

Yahoo! Inc., v La Ligue Contre Le Racisme Et L'Antisemitisme, 169 F.Supp. 2d 1181, 2001.

United States v. Aluminum Company of America (ALCOA), 148 F.2d 416 (2d Cir.1945).

Zemnick, S. R. (2001). Student note: The E-Sign Act: The means to effectively facilitate the growth and development of e-commerce. *Chicago-Kent Law Review, 76,* 1965.

Zippo Manufacturing Co.v Zippo Dot Com, Inc., 952 F. Supp. 119, W.D. Penn., 1997.

International Supply Chain Management

Gary LaPoint, *Syracuse University*
Scott Webster, *Syracuse University*

INTRODUCTION

The term "supply chain management" is open to alternative interpretations. We view a supply chain as two or more parties linked by a flow of resources—typically material, information, and money. Supply chain management deals with managing flows of resources among departments within a firm and among independent enterprises.

Some believe there are fundamental differences between supply chain management and logistics management. We disagree. Writings on logistics management have long stressed the importance of a total systems perspective, although there has been a gradual shift of emphasis from a single enterprise focus to a broader multienterprise view today. For example, more than 25 years ago, Bowersox (1974, p. 1) defined logistics management as the process of managing all activities required to move raw materials, parts, and finished inventory strategically from vendors, between enterprise facilities, and to customers.

Approximately 10 years later, Houlihan (1985) stressed the strategic importance and opportunities of interenterprise logistics issues, and introduced the term supply chain management to emphasize this point. This broader view is now reflected in the Council of Logistics Management's (2002) definition of logistics:

> Logistics is that part of the supply chain process that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services, and related information between the point of origin and the point of consumption in order to meet customers' requirements.

In summary, the systems perspective of logistics management has become richer over time, in part because of advancing information technologies that reduce barriers to interenterprise supply chain planning and coordination. In the next section, we discuss how these technologies in combination with new trade agreements are driving rapid growth in international supply chains. There are many laws, processes, and service providers that are unique to international supply chains. The main lesson from this section is that a basic understanding of these laws, processes, and service providers can be essential for a firm's survival, let alone success. This provides the backdrop for a section that reviews laws, processes, and service providers specific to international supply chains. We provide sources for more detailed information and also discuss barriers that are limiting further growth in international supply chains. Finally, we conclude with a chapter summary.

TRENDS IN INTERNATIONAL SUPPLY CHAIN MANAGEMENT

International Trade

The volume of material, information, and money flowing through international supply chains is significant and expanding at a rapid pace. In the United States, for example, import and export trade has been growing faster than the U.S. economy, with this trend expected to continue over the foreseeable future. More than $2 billion worth of cargo is imported into the United States, every day (Toran, 2002). Imports to the United States are expected to grow at an average annual rate of 7.9% between 2000 and 2010 (Su, 2001). Similarly, U.S. exports are expected to grow at an average annual rate of 7.8% between 2000 and 2010 (Su, 2001). About $3 billion of goods and services are exported from the United States every day, with the dollar volume doubling about every 7 years between 1970 and 2000 (Testimony of Under Secretary of Commerce, 2001).

Outsourcing

A closely related trend is the growth in outsourcing in recent years. A 1999 survey found an increasing number of companies are contracting noncore business processes to an outside provider as a means to improve performance

(Global study finds, 1999). In another survey, 50% of executives viewed outsourcing of noncore functions to be critical to improving supply chain performance (Chabrow, 2002).

Outsourcing began to pick up pace in the early 1990s. The main motivation was to reduce cost, although this is less the case today when time-to-market, access to high-quality employees, and workforce flexibility can take precedence (Outsourcing Institute, 2002). Common early outsourcing targets included ancillary activities such as janitorial and cafeteria services. We have since seen a growth in third-party logistics companies as an increasing number of firms have outsourced transportation, warehousing, order fulfillment, and import–export functions. Similarly, a growing number of components and subassemblies that were once produced in house are now being sourced from low-cost regions such as Latin America, the Far East, and Eastern Europe. In 2000, for example, the average contribution of exports of goods and commercial services to the gross domestic products of developing nations was 26%, which represents a 53% increase since 1990 (World Trade Organization, 2002). Today, relatively low-value items requiring fairly simple production methods dominate most offshore sourcing. The capabilities to design and produce sophisticated products requiring skilled labor and complex processes are rapidly advancing in low-cost regions of the world, however. This trend is especially evident in the developing high-tech industrial sectors in the Far East. Original design manufacturers (ODMs) provide design and manufacturing services; in the consumer electronics industry, ODM revenues are expected to increase from $4.5 billion (in U.S. dollars) in 2001 to $77.8 billion in 2005 with most ODMs in Taiwan (Carbone, 2003).

What explains these trends of growing international trade and outsourcing, and what are the implications for management? Let us begin with the first question. Although there are arguably many contributing factors, the Internet and related information technologies in combination with the general lowering of trade restrictions in recent years stand out. The Internet is reducing the costs in both time and money of finding and initiating profitable business ventures, and in realizing efficient collaborative relationships among new supply chain partners. At the same time, the pool of viable business partners is continuing to expand as trade restrictions are being relaxed. We explore both of these factors in the following sections.

Internet and Related Information Technologies

Initially the Internet was a new source of communication; then it became a source of information. Today businesses use the Internet to advertise, take orders, manage inventories, provide information to customers, locate suppliers, and negotiate. The Internet has helped turn a planet of individual economies into essentially one global economy. Buyers and sellers from around the world come together in a virtual market place. Manufacturers are finding suppliers they never knew existed in locations never considered to be a possible source. Likewise, suppliers are finding buyers they never knew existed. A survey found that more than 70% of executives view the Internet as a key factor in fostering greater collaboration with business partners through increased supply chain visibility (Chabrow, 2002). Online purchases are estimated at more than $1 trillion in 2002 (Vigoroso, 2002) and predictions for global business-to-business (B2B) e-commerce range between $4.5 trillion and $8.5 trillion by 2005. International B2B e-commerce is predicted to grow annually at a rate of more than 70% between 2000 and 2005 with the fastest growth in Western Europe and Asia–Pacific (International B2B, 2001). For example, the European Union's online trade represented less than 1% of their total business trade in 2001 and is expected to grow to 22% by 2006. The increase in online trade spurs continued growth as adopters pressure other trading partners to accelerate their use of the Internet for commerce.

There is a wide range of Internet-based technologies and software applications that support the planning, buying and selling, and control associated with the movement of resources through supply chains (see http:// www.freightworld.com/software.html for a list of nearly 200 supply chain management software vendors). Many of these technologies and software applications are discussed elsewhere in this book (see chapters on related topics, such as B2B and B2C electronic commerce, collaborative commerce, customer relationship management, data mining, electronic data interchange, e-manufacturing, enterprise resource planning, e-procurement, extensible markup language, intelligent agents, supply chain management, online analytical processing, and online auctions). Our purpose in this section is to highlight software applications and related Internet-based technologies that are specific to international trade. As the following sections show, the movement of materials across international borders requires extensive and detailed documentation. There are also many (less than obvious) costs associated with steps in import–export processes. These two complexities unique to international trade have given rise to software applications that help streamline document preparation and exchange and that facilitate cost analysis of international sourcing alternatives. Examples of such software applications are briefly motivated and discussed here.

In the wake of September 11, 2001, security has been tightened, and international shipments are more heavily scrutinized. Customs agents in many countries are placing an even greater emphasis on compliance to import–export laws. Even the slightest of errors in paperwork can delay shipments at gateway ports. Bolero.net (Bolero.net, 2002) is an example of a software application that provides tools for preparing requisite international trade documents and for secure electronic document transfer over the Internet. By sending the complete documents ahead of the physical goods, customs clearance and ground handling can be prepared at the destination before arrival of the freight (Global Logistics Group, 2002). Electronic documentation also helps eliminate mistakes in paperwork that can result in significant delays. Other examples are Tradex2000 and ExportDoc Worldwide, which assist shippers in preparing export documentation. The software ensures that shippers enter required information and that all the necessary documentation is prepared. NAFTAssitant is a related software application with a more narrow focus on documents required under NAFTA and on being in compliance with NAFTA regulations (NAFTAssistant, 2002). NAFTAssitant maintains large databases of information on a firm's products such as the country of origin, harmonized tariff schedule numbers, descriptions, and part numbers.

As companies venture into the global marketplace, many lack even basic skills required to import and export. Obviously, making a decision to outsource a component from an international supplier based solely on the unit cost and not the total landed cost per unit would be foolish. But what costs should be included in the analysis? Determining all the costs associated with importing can be elusive, even for experienced importers. Costs may include value added tax, general service tax, duties with possible adjustments for duty drawback, ocean charges, inland transportation and terminal handling fees, to name just a few. Of equal concern is being in compliance with import and export laws. This is particularly difficult because of the myriad regulations and the continual updates to documents such as the Denied Parties List, Specially Designated Blocked Persons List, and the Commercial Control List. Xporta Solutions provides software that enables supply chain managers to compare all the costs associated with various sourcing alternatives. Their software also ensures that imports and exports comply with government regulations.

Extensible markup language (XML) is an example of an Internet-based technology that facilitates electronic document exchange. It is a technology that goes much beyond the realm of international trade and is discussed elsewhere in this book. There is a technology that builds on XML (known as ebXML), however, that is relevant for software applications of the type discussed here. ebXML is an evolving XML-based standard for global e-commerce. It defines computer-to-computer information exchange protocols in an XML framework enabling a standard to exchange business messages, conduct trading relationships, communicate data in common terms, and define and register business processes. ebXML is sponsored by United Nations Center for Trade Facilitation and Electronic Business (UN/CEFACT) and OASIS (ebXML: Enabling, 2002). Both of these organizations were formed to develop and establish e-commerce standards. Many of the software applications for international trade are designed to be ebXML compliant.

The technologies discussed here facilitate buying and selling over the Internet through efficient information sharing, transaction processing, and cost analysis. These technologies are advancing and becoming more widely accessible, reducing barriers to international trade and accelerating import–export processes.

Trade Agreements

New trading blocks have emerged throughout the world during the last decade. Trading blocks are countries that have established trade agreements between one another for the purpose of increasing trade. Trading blocks are typically regional. The primary means of increasing trade is to lower the tariffs on items traded between the countries. There are six trading blocks in the Western Hemisphere alone. Of these six, the North American Free Trade Agreement (NAFTA) is by far the largest in terms of trade. There are, however, discussions on the creation of a South American Free Trade Agreement (SAFTA). Other large trading blocks include the European Union (EU) and the Asia–Pacific Economic Cooperation (APEC). There are many smaller trade agreements throughout the world as well, all designed to ease tariffs and increase trade between selected trading partners (see, e.g., http://intl.econ.cuhk.edu.hk/rta/).

Implications for Supply Chain Managers

Trade barriers are lowering, supply chain management information technologies are advancing, and these technologies are becoming more widely available throughout the world. These factors are behind a long-term trend toward increased international trade and, in particular, supply chains that span an increasing number of countries. The challenge for a firm is to develop the management talent and harness the technologies to exploit this trend. The importance of this challenge is further heightened by the opportunities to penetrate emerging markets. Approximately 83% of the world's population resides in countries comprising 17% of the world's combined gross domestic product (GDP), but the economies in many of these countries are the fastest growing in the world (International gross domestic, 2000). The Far East, Eastern Europe, and Africa, in particular, are emerging as high-growth markets that will likely be a focus of marketers for at least the next two decades. The largest construction projects taking place in the world today are in China and other Far East countries. Companies that are adept at establishing and managing international supply chains will be well positioned to take advantage of these growth markets.

In summary, a firm's ability to identify, establish, and manage relationships effectively among multinational supply chain members has become a strategic core competency. Excellence in this area (a) provides cost, quality, and agility advantages through effective sourcing; (b) increases revenues as a firm is able to quickly move into high growth foreign markets; and (c) is not easily duplicated by others. In the next section, we outline the rudiments of what every manager should know about buying and selling across international borders. We cover the types of international service providers that are frequently used in international trade, and we cover the basic import–export laws and processes. Our coverage of laws and processes is from a U.S. perspective, but many of the features are common to other countries.

ELEMENTS OF INTERNATIONAL SUPPLY CHAINS

International Service Providers

Understanding the roles of key international service providers and establishing strong relationships with such firms is essential for effectively managing international supply chains. The service providers discussed in this section specialize in the movement of materials across international borders. International trading has many more variables than domestic trading, and these added variables increase the chances that orders or shipments can be delayed or assessed fines. Aside from the obvious complications such as distance, time, and language, other notable differences include government restrictions, different currencies, additional handling, different modes of transportation, additional paperwork, customs requirements, and trade terms. There are four commonly used international service providers: (a) customs brokers, (b) international freight forwarders, (c) international carriers, and (d) international banks.

Customs Brokers

Although not required by law, the services of a licensed customs broker are commonly used when importing. Thus, although it is true that individuals can clear their own import through customs, the complexity and periodic changes in requirements dictate a high level of specialized expertise and knowledge. Given the complexities of U.S. import regulations and potential ramifications of improperly importing merchandise, licensed customs brokers are essential players when importing. Being licensed affirms that the broker meets a certain level of competence as required by U.S. Customs to make entry of merchandise into the U.S. and permits a broker to clear freight on behalf of others.

Customs brokers operate on behalf of the importer of record. The importer of record is ultimately accountable for the import. U.S. customs brokers provide preentry services such as preadvance clearing, filing formal entries for shipments valued greater than $2,000, classifying material, paying duty, and performing postentry services (e.g., filing protests on behalf of the importer, filing for duty drawback). Brokers are responsible for maintaining their own records, but contrary to belief among some importers, brokers are not responsible for maintaining records for the importer of record. Federal law requires that the importer of record maintain all documentation used in the importation of material for 5 years. The law is specific as to how these records are to be maintained, including for example, how records are to be filed in a filing cabinet. Rules and regulations pertaining to customs brokers can be found at www.access.gpo.gov/nara/cfr/waisidx_02/19cfrv1_02.html, and general information can be found at the National Customs Brokers and Forwarders Association of America Web site (http://www.ncbfaa.org) and at http://www.customs.gov.

International Freight Forwarders

International freight forwarders act as a middleman between the exporter–importer and the carrier. Freight forwarders can work either on behalf of the shipper (the exporter) or the consignee (the importer). A freight forwarder's business is to move freight as quickly as possible at prices lower than what their client could generally receive on their own. Few exporters–importers ship enough volume to deal directly with an international carrier such as a steamship line or a commercial airline. Freight forwarders contract for space on these carriers by committing a certain level of business based on the total shipping volume of their customer base. By doing so, they can obtain lower rates than what the average shipper could obtain on their own. Although freight forwarders make international trade viable for the smaller to medium size company, even the largest of shippers use them. Having a presence at the ports of entry and contacts at these ports of entry is critical when moving freight internationally. When imported merchandise reaches a U.S. gateway, the slightest discrepancy or question places the merchandise on hold until the question is resolved. Many times the issue is easily resolved if someone is at the port of entry working on behalf of the importer. Otherwise, it can be difficult to make the proper contacts to get the problem resolved in the time frame in which the importer would like it resolved. Without a presence at the ports of entry, there is little urgency for resolution. Most companies cannot afford to maintain this presence with their own staffs.

Some international air freight forwarders are certified by the International Air Transport Association (IATA; http://www.iata.org). The air-freight industry is not regulated, and as such, air freight forwarders do not have to be certified. Air freight forwarders certified by IATA, however, are issued an IATA number. Many commercial airlines will ask for an IATA number before they will accept any freight from an air-freight forwarder.

In 1998 the Ocean Shipping Reform Act deregulated ocean shipping. The Act became effective on May 1, 1999. Part of the Act consolidated ocean freight forwarders with non-vessel-operating common carriers (NVOCCs). The consolidated group is now known as Ocean Transportation Intermediaries. Both ocean freight forwarders and NVOCC's are required to have a bond demonstrating financial stability, but only the ocean freight forwarder is required to be licensed by the Federal Maritime Association. In all practicality an NVOCC, also referred to as an NVO, acts as a carrier with the only difference being they do not operate a vessel. Although an NVO may seem like a freight forwarder, they are not the same. An ocean freight forwarder prepares all the documentation required for shipment and negotiates freight rates, among other activities. The difference is that the paperwork prepared by the forwarder is in the name of the forwarder, not the carrier, whereas, the paperwork prepared by the NVO is essentially in the name of the carrier. Freight forwarders are sometimes the largest customers of NVOs.

From a shipper's standpoint, freight forwarders act as a common carrier. A shipper must be careful to understand that international forwarders limit their liability, however. The prudent shipper will carefully review the liability limitations of an international freight forwarder prior to signing any contract or tendering any shipment.

Two organizations are involved in setting standards and practices in the freight forwarder industry. The first is the International Federation of Freight Forwarders Association (http://www.fiata.com), based in France, and the second is National Customs Brokers and Freight Forwarders of America Association (http://www.ncbfaa.org).

International Carriers

Truck, rail, air, and water are all used for international freight, and multiple carriers are often employed between origin and destination. This even occurs between contiguous countries due to restricted access to foreign carriers. In these cases, which are not uncommon, cargo must be transferred at the border to a carrier based in the country.

Carriers, whether domestic or international, may act as a common carrier or contract carrier. By law, common carriers must accept business from anyone who tenders them business (although there are ways carriers can discourage certain types of business and customers). A contract carrier, on the other hand, has a formal relationship with a shipper, the terms of which are spelled out in a negotiated contract. For situations in which there is a specific need (e.g., critical time frame, cargo requiring special handing) or significant ongoing volume, a shipper may be able to achieve better service or pricing (or both) through a contract carrier than a common carrier.

Shippers can work directly with international steamship lines and air carriers. Most international carriers presell their cargo space to freight forwarders, however, and with the exception of firms with significant volume on a continual basis, it is generally difficult for a shipper to book space. Most shippers do not deal directly with international carriers, other than perhaps UPS and FedEx, which operate their own aircraft. Even integrated carriers such as these will act as a forwarder for some of their international freight. Additional detail on issues relating to international steamship lines and air carriers can be found at the Federal Maritime Commission Web site (http://www.fmc.gov) and the International Air Transport Association Web site (http://www.iata.org).

International Banks

Financial risks are present in nearly every international transaction (e.g., buyer credit risk, political uncertainty, documentation errors, currency fluctuations, interest rate fluctuations). International banks may play a role in reducing financial risk by providing a letter of credit, more commonly known as an LC. An LC provides assurance of the buyer's ability to pay for a shipment, and a shipper may require an LC prior to shipment. This means that a shipper requires the buyer to deposit an amount equal to the value of the commercial invoice in the buyer's bank. When the material arrives, shipping documents are presented to the buyer's bank, which in turn transfers the funds to the shipper's bank to collect funds on behalf of the shipper. LCs can also be used to protect against fluctuating currency and interest rates by setting a value for each and holding those values for a specific period of time. As a word of caution, an exporter should always verify that an LC is in accordance with UCP 500. UCP 500, which stands for Uniform Customs and Practice, is the international convention that governs the actions and obligations of the bank involved in an LC transaction.

International Commercial Terms

International trade requires a thorough understanding of international commercial terms, or Incoterms. Incoterms are analogous to what is known domestically in the United States as free-on-board (FOB) terms. Ignorance of Incoterms can be costly; it is not uncommon for those unfamiliar with Incoterms to be taken advantage of by those who do.

There are 13 Incoterms. A listing and description of each Incoterm can be found at http://www.iccwbo.org/incoterms/wallchart/wallchart.pdf and http://www.ltdmgmt.com/incoterms.htm (see http://www.incoterms.org for more detailed information). Incoterms were created by the International Chamber of Commerce to standardize the terms used for international trade. Great care should be used in selecting the proper Incoterm. Although Incoterms may appear to be written in a manner that is specific and precise, their meaning is often subject to interpretation. Buyers and sellers must clearly define in their sales or purchase contract who has insurable interest and where the insurable interest is assumed. In particular, insurable interest should be clearly defined and detailed when shipments are consigned to a third party, such as a third-party logistics firm or a bank. Each party to a transaction must be well versed in the Incoterm specified in the purchase contract. A shipper or consignee may think the other party has insurable interest in a shipment only to find out after a loss that they did as well.

Another hazard of selecting the improper Incoterm is that a seller can unknowingly be made the importer of record on an order when exporting to another country. For example, suppose the buyer in a foreign country purchases material from the seller who is located in the United States. The U.S. seller exports the shipment to the buyer. One may think that the buyer would become the importer of record. The buyer, however, may have selected an Incoterm that essentially makes it appear to foreign customs that the seller, not the buyer, is the importer of record. The seller is now responsible for clearing the shipment through foreign customs and for paying any foreign duty and taxes that would have otherwise fallen on the buyer. In addition, the seller can be subject to other fees or duties as described by the law in the foreign country.

The 13 Incoterms fall into four groups that correspond to the first letter of the term (i.e., C, D, E, and F). E and F Incoterms minimize a seller's risk and cost and are least favorable to the buyer. C terms create a little more risk and cost for the seller and less for the buyer, and D terms create the most risk and cost for the seller. A seller, for example, should not accept a D Incoterm without a thorough understanding of its implications. The delivered duty paid (DDP) Incoterm, in particular, is the most dangerous from the seller's perspective. DDP can, depending on the country, make the exporter a nonresident importer of the country to which they are exporting. This means the exporter would be responsible for all transportation, duties, and taxes from the point of origin to the customer's dock, as well as any other costs that might be incurred along the way.

U.S. Export Laws and Processes

A shipment is considered an export if the destination of the shipment is outside the United States and its origin is within it. Specific documents are required to export material out of the United States. The particular documents required will depend to some degree on the nature of the product being shipped, the value of the shipment, the type of service required and the destination. A representative list is identified later in the chapter (general references for further information relating to exporting include http://www.tradeport.org/ts/trade_expert/infobase/basic/ and http://www.cia.gov/cia/publications/factbook/).

Transport Documentation

Every shipment must travel with some form of manifest. These are typically referred to as a bill of lading for ground and ocean shipments and an airway bill for air shipments. The bill of lading and the airway bill serve as a contract between the shipper and the carrier. Most international shipments use both a domestic and an ocean (or airway) bill of lading. These documents are typically nonnegotiable, which means that they do not convey title. As such, the carrier can deliver the order without the consignee presenting an original copy of the bill of lading. An order bill of lading, however, is negotiable and is often used in international transportation. If the shipment includes an order bill of lading, then the consignee must pay the value of the invoice to receive the original bill of lading from the shipper. The carrier will not tender delivery unless the consignee can present the original bill of lading to the carrier. Only then does title of the merchandise pass from seller to buyer. In essence, an order bill of lading is similar to a domestic cash-on-delivery (COD) shipment.

Shippers Export Declaration (SED)

Unless specifically identified as an exception in the export administration regulations, an SED will be required for every export shipment in which any individual line item is valued at $2,500 or more. For example, if an order consists of 10 line items and the value of each line item is less than $2,500, then no SED is required, even though the value of the entire shipment may exceed $2,500. A shipment is defined as all merchandise moving from one exporter to one consignee on one exporting carrier. Separate SEDs are required for each shipment.

The shipper or their designated agent prepares the SED. The person who signs the SED must be in the United States at the time of signing. The person who signs the SED, whether exporter or agent, is responsible for the truth, accuracy, and completeness of the SED, except insofar as that person can demonstrate that he or she reasonably relied on information furnished by others. As of November 1, 2001, SEDs must be filed electronically with U.S. Customs. If the shipping location is not capable of electronic filing, then a forwarder, agent, or the carrier must file the SED. The forwarder, agent, or carrier's authority to sign the SED must be executed by a formal power of attorney.

Commercial Invoice

A commercial invoice is required for every export shipment. Unlike SEDs, there are no exceptions. Foreign customs offices use the commercial invoice to determine the duty rates for imported merchandise. The commercial invoice contains the following information: (a) Country of origin for each item, (b) seller's name and address, (c) buyer's name and address, (d) description of the items being shipped, (e) schedule B number (this number, which is referred to as a harmonized tariff schedule number on imports, is a code that is used to determine the duty rate), (f) extended value of each individual line item, stated in U.S. dollars.

For customer orders, the value used on the commercial invoice must reflect the transaction value for that merchandise. If a U.S. company conducts business with its foreign subsidiaries on an arm's-length basis, prices charged to their foreign subsidiaries are to reflect the competitive conditions in the market and be comparable to prices charged to large unrelated customers.

Packing List

A packing list must accompany every export shipment. The packing list contains the (a) description of the material being shipped, (b) quantity, (c) shipper's reference number, and (d) customer purchase order number.

Certificate of Origin

Some export shipments may require a certificate of origin. This is not a requirement of U.S. export law but a request on the part of the buyer or a requirement of foreign customs. The certificate of origin is a document, usually signed by an officer in the company, attesting to the country of origin of merchandise described on the commercial invoice.

Export License

Export licenses are required if a shipment is consigned to anyone or any country identified on restricted trade lists maintained by the U.S. government. These lists are discussed in more detail later (see Other Considerations).

Miscellaneous Documents

In some circumstances, other export documents may be required. Some examples are listed in the following sections.

Fumigation Certificates or Declaration. Some countries require U.S. shippers to produce a fumigation certificate or a declaration stating that the packaging or crating material is made from nonconiferous wood. If the packaging or crating is made of coniferous wood, then the wood must either be fumigated or heat treated to kill any insects that may inhabit the wood. These requirements are country specific. If crating, packaging, or fumigation is performed by an outside service, the service will be required to complete and sign the fumigation certificate or declaration.

Inspection Certificates. Some customers may require inspection by a certified inspector prior to shipment. Upon inspection, the inspector will complete an inspection certificate that is typically faxed to the buyer before the shipment takes place. The certificate states that the shipment meets the requirements of the importer or the importer's country.

Hazmat Documents. In the event that a product is hazardous, a hazmat bill of lading or hazardous airway bill is required. Many air carriers will not accept international shipments of hazardous material, so any such shipments will require investigation to see which, if any, air carrier will accept the shipment.

Consularization. Some Middle Eastern and Latin American countries require export documents to be consularized, or stamped, by their foreign consulate prior to export. Consularization is a formality that certifies the shipment to be exported to that foreign country. A certificate of origin and a commercial invoice must be presented to the foreign consulate. Some consulates will stamp the documents with an official stamp, whereas others will issue a consular invoice. If a consular invoice is generated, then the original document must travel with the shipment and be presented to the foreign customs.

Destination Control Statement. A destination control statement (DCS) is required for all exports from the United States for items that are restricted for export by the U.S. government. Lists of restricted persons and countries are discussed later (see Other Considerations). The DCS is a statement to the consignee (receiver of material) that they cannot forward this shipment on to another party without the permission of the shipper.

ATA Carnet. One little-known tool that can be useful for exporting equipment or material that needs to move around while in a foreign country is the ATA carnet. (ATA stands for admissions temporary admissions.) The ATA carnet is like a passport for merchandise that allows temporary importation of products free from duties, taxes, and security deposits. It permits material to move freely within a foreign country for the duration of the carnet. Material traveling under an ATA carnet enters a foreign country duty free; when the material returns to the United States, it enters the U.S. duty free as well. The ATA carnet is particularly useful for trade show or exhibition materials. Carnets are not universal. They are currently accepted in 58 countries and 27 territories (Advincula, 2001). See http://www.uscib.org for a list of countries that accept carnets.

Other Considerations

During the 1990s, and particularly since September 11, 2001, the U.S. government has intensified its scrutiny of both import and export shipments. Specific individuals within the customer service organization process export orders in many medium size to large size companies. Prior to releasing an export order to the warehouse or the factory, these individuals must review several government alerts to ensure that the countries, individuals, or organizations receiving these materials either directly or indirectly are not restricted by the U.S. government. The following items must be reviewed prior to release of every export order to the warehouse or the shipping floor to ensure that the order is not in conflict with established regulations.

Denied Persons List (DPL). The DPL is maintained and updated by the Bureau of Industry and Security, formerly known as the Bureau of Export Administration (BXA). U.S. law prohibits exports to individuals and businesses on the DPL. The DPL is available at http://www.bxa.doc.gov/dpl.

Commerce Control List and Economic Control Commodity Number (CCL/ECCN). The CCL is maintained and updated by the Bureau of Industry and Security. The list identifies all commodities that are restricted for export by the U.S. government. Every item on the CCL is assigned an economic control commodity number (ECCN). If an item is listed on the CCL, then proper authorization must be received from the U.S. Department of Commerce prior to export. Authorization is typically in the form of an export license. The CCL is available at www.access.gpo.gov/bis/ear/ear_data.html.

Embargoed Nations List. The U.S. government has identified nations with which it strictly forbids conducting business, either directly or indirectly. This list is maintained and updated by the U.S. Treasury Department, Office of Foreign Asset and Control. The embargoed nations list is available at http://www.treas.gov/ofac.

Special Designated Nationals and Blocked Persons (SDNB). The SDNB is maintained and updated by the U.S. Treasury, Office of Foreign Asset and Control (OFAC). The SDNB contains the names of individuals, organizations, and business concerns whose assets have been frozen by the U.S. government. Conducting business with those who appear on the SDNB is prohibited unless approval in the form of a license is received from OFAC. The SDNB list is available at http://www.treas.gov/ofac.

In addition to reviewing these issues, the export manager must also be familiar with the International Trade on Arms Regulations (ITAR) and the Antiboycott Regulations. ITAR defines the regulations associated with the exportation of products that can be used as a weapon or in conjunction with a weapon such as a component or as part of a weapon system. Items of concern are identified on the U.S. Munitions List in part 121.1 of ITAR (see http://www.pmdtc.org/reference.htm#itar).

Antiboycott laws prohibit U.S. citizens and U.S. firms and their subsidiaries from participating in foreign embargoes that the U.S. government does not sanction. The effect is to prevent U.S. firms from being used to implement foreign policies of other nations that run counter to U.S. policy. For example, it is in violation of antiboycott regulations for a customer in a foreign country to require that a particular airline of a specific foreign country not be used to ship their order. Antiboycott regulations are maintained by the Bureau of Industry and Security (BIS) and can found at http://www.bxa.doc.gov/antiboycottcompliance.

Foreign subsidiaries of U.S.-based companies are required to comply with the same U.S. export laws. The penalties for violating these restrictions can be harsh. Criminal and civil penalties can be assessed against violating companies, including revoking a company's export privileges. Additionally, individuals who were associated with the violation can be held personally responsible and can be subject to personal fines or imprisonment depending on the severity of the offense.

U.S. Import Laws and Processes

A shipment is considered an import if the destination of the shipment is the United States and its origin is outside of the United States. The party responsible for the import is referred to as the importer of record. The importer of record is responsible for U.S. Customs clearance, and duties and taxes associated with the import. Generally the importer of record is the buyer of foreign merchandise. With the exception of the shippers export declaration, the documents required for importing are the same as for exporting. Mandatory items include a manifest, commercial invoice, packing list, and either a bill of lading or airway bill. The same miscellaneous documents used in exporting can also be found in importing.

For the most part, an importer requires a licensed U.S. Customs broker to clear the shipment through U.S. Customs and an import bond. The bond ensures that U.S. Customs gets paid the duty it is owed. Bonds can be purchased individually as imports arrive or a company can purchase a continuous bond, which is also known as a surety bond. A continuous bond permits a large importer from having to purchase a bond each time it receives an import.

Another form of bond is the temporary import bond, commonly referred to as the TIB. The TIB allows an item to enter the U.S. duty free on a temporary basis with the requirement that the item will be exported out of the U.S. within 1 year. During that year, the item cannot be sold or transferred. If at the end of a year the material is not ready to be exported, the importer can apply for a TIB extension. If the item fails to be exported during the term of the TIB or if the item is sold, then U.S. Customs can, and generally will, levy fines and penalties on the importer.

There are two key areas for U.S. import compliance. The first is known as the entry. When an import arrives at a U.S. gateway, it must be cleared to enter the United States. This is where the customs broker comes in. The customs broker will prepare an "entry" that will permit U.S. Customs to process the shipment and allow it to enter the country. The entry form contains four main pieces of information pertaining to the shipment. First, the entry form identifies the origin of the shipment to ensure the material is not originating from an embargoed nation or a specially designated national and blocked person. Second, the entry form identifies the harmonized tariff schedule (HTS) classification number used to describe the item(s). The HTS classification determines the duty rate of an item. Also, certain items are subject to countervailing or antidumping duties, so proper classification of an item is critical. Third, the entry form identifies the value of the shipment so a duty can be calculated. Fourth, the entry form identifies whether or not the two parties involved in the transaction are related. It is unlawful for an importer to take advantage of a related party by having it assign a value to a product that is unreasonably low for the purposes of lowering the duty. Because duty is based on the value of a good, a lower value would result in a lower duty assessment. It is the broker's responsibility to ensure that the information on the entry is accurate.

The second key area of U.S. import compliance is record keeping. In 1993, the 103rd congress passed the Customs Modernization Act also referred to as the Mod Act. The Mod Act shifts the legal burden to exercise "reasonable care" in conducting international trade onto the trade community (Linet, 2002). Prior to the Mod Act, U.S. Customs had assumed the responsibility of ensuring that imports were in compliance. U.S. Customs requires all importers to maintain records of their import activities. "Any owner, importer of record, consignee or their agent who imports or knowingly causes goods to be imported into the customs territory of the United States shall make and keep such records in accordance with U.S. Customs regulations" (Tariff Act, 2002). Many U.S. companies unknowingly think that because they use a customs broker that the broker is responsible for maintaining the importer's records. This is a serious misjudgment on the part of the importer. The broker is only required to maintain its own records, not those of the importer.

U.S. Customs may audit the records of any U.S. company at any time. The North American Free Trade Agreement (NAFTA) also permits the Canadian government to audit the records of a U.S. company that has shipped goods from the U.S. to Canada under NAFTA. Importers should conduct a self-assessment of their import procedures to determine if they are in compliance with U.S. Customs regulations (see http://www.customs.gov/xp/cgov/import/regulatory_audit_program/ for self-assessment guidelines). In addition, a firm should keep a record of all inquiries made to U.S. Customs. Inquiries can be in the form of requests for a ruling or requests for assistance in classifying an item. Finally, internal import and export procedures should be documented.

Assists

An assist is when a U.S. firm provides assistance to a foreign supplier to produce or provide merchandise that will eventually be sold and exported back to the United States firm. An assist can be in many forms. Common forms of assists are financial assistance, manufacturing equipment, tools, dies, molds, technical drawings, intellectual property, and some forms of technical advice.

The U.S. firm, which in this case is also the importer of record, is accountable for the duty on the value of the assistance given to the foreign supplier at the duty rate applicable to the finished product being imported. For example, suppose a U.S. firm provides $100,000 in startup capital to a foreign concern to make a product that the U.S. firm will in turn purchase and import back into the United States. Additionally, the U.S. firm also sends the foreign supplier tools and dies totaling $250,000 in value, and the finished product, when imported into the United States will be subject to a 5% duty rate. This means that the U.S. firm, in addition to the normal duty on the finished product, will have to pay $17,500 ($350,000 × 5%) on the value of the assist.

Often a U.S. company will purchase foreign merchandise through a purchasing agent, even material they have arranged to have made for them as just described. This removes the responsibility of arranging transportation and customs clearance from the U.S. firm. It does not, however, remove the obligation of paying for the assist. The U.S. firm is responsible for notifying the purchasing agent of any assists so that the purchasing agent can pay the assist when the merchandise is imported. The purchasing agent would in turn charge the U.S. firm for the amount paid to U.S. Customs.

Foreign Trade Zones

Foreign trade zones (FTZs), also known as free trade zones and enterprise zones, are specific areas within a country that have been designated as a duty free zone. For customs purposes these zones are considered to be outside the country, and thus, no duties are paid when merchandise enters an FTZ. Merchandise leaving an FTZ for domestic use must go through customs and duties must be paid. FTZs offer a number of advantages for importers and exporters (Ballou, 1999), including the following:

- Deferral of duties until the point in time when product is needed within the host country.
- Imported goods may be temporarily stored or modified and assembled in an FTZ, then shipped to another country with no duties or customs processing.
- Errors in documentation can be corrected in an FTZ prior to going through customs, thereby avoiding fines.
- Product subject to spoilage, damage, and loss do not incur duties on the amount lost.

The main disadvantage of an FTZ is reduced revenue from import duties by the host country.

In the United States, FTZs fall under the supervision of the U.S. Customs Service and the Foreign Trade Zone Board. There are two types of FTZs in the United States: (a) general purpose and (b) special purpose subzones. A general purpose FTZ is typically licensed to a municipality, public agency, port authority, or industrial park to operate on behalf of the general public. A subzone, on the other hand, is a site authorized by the Foreign Trade Zone Board to operate as an FTZ but is generally not physically located in an FTZ. Subzones are usually licensed to specific firms, frequently as a manufacturing site or distribution center. They are designed to permit companies to receive the benefits of a FTZ without being physically located in the FTZ. At present there are 256 general purpose FTZs and 438 single company subzones operating within the United States. Web sites maintained by the Foreign Trade Zone Board (http://www.ia.ita.doc.gov/ftzpage), the National Association of Foreign Trade Zones (http://www.naftz.org/body_what.htm), and U.S. Customs (http://www.customs.gov) contain additional detail on FTZs.

Barriers to Growth

There remain several barriers limiting faster growth in international trade. First, although progress is being made (e.g., ebXML), there are no widely accepted and used standards for global e-commerce. Different countries and companies operate using different protocols. The United Nations has created UN/CEFACT, the United Nations Centre for Trade Facilitation and Electronic Business, to help develop a global standard in conjunction with other standards organizations.

Another barrier is currency. At present there are hundreds of currencies each with their own valuations. Large and unpredictable variations in currency value make it difficult to trade in some regions of the world. Much discussion is beginning to take place on the merits of common currencies. To increase international trade, there will invariably have to be a reduction in the number of currencies used in trade, either by combining currencies, such as in the case of the Euro, or through the use of currency unions or currency boards. Barter is still used in some instances, but that does little to improve local economies and can only be used on a limited basis. In the not too distant future, there may only be a few key trading currencies. This is emerging today. The Euro currently includes 12 European countries and will surely spread to others. The U.S. dollar is already used as the primary currency in many countries, and many others have linked their currencies to the U.S. dollar. It is conceivable that as trading blocks merge to form super trading blocks, each trading block will develop a common currency. Frankel and Rose (2002) found that common currencies increase bilateral trade by a factor of three. As we have seen with the Euro, however, creating common currencies is difficult.

Security measures for U.S. imports are tighter today than before September 11. In January 2002, U.S. Customs initiated the Container Security Initiative (CSI) as a means to ensure the integrity of container shipments entering the United States. The initiative requires inspection of ocean container contents, either through direct physical examination or via noninvasive methods, at the foreign port of embarkation. In addition, all ocean carriers and NVOCCs are now required to send an electronic manifest 24 hours prior to the loading of the vessel at a foreign seaport. Depending on where a shipment is originating, it has been estimated that this new requirement will delay a shipment from 2 to 4 business days (TDC Trade.com Business Alert, 2002). As a result, shippers need to have material at the seaport at least 1 week earlier, resulting in approximately 20% more inventory in the pipeline. Increased security by U.S. Customs at U.S. ports of entry is also causing delays. Companies can reduce delays by participating in the Customs Trade Partnership Against Terrorism (C-PAT), however. By participating in C-PAT, a company agrees to conduct a comprehensive self-examination of its supply chain security using guidelines established by U.S. Customs and by the trade community (U.S. Customs, 2002). The primary benefit to companies who participate in C-PAT is less scrutiny by U.S. Customs on inbound shipments, which translates to lower processing times at ports of entry and border crossings.

CONCLUSION

We argued that the Internet and related information technologies in combination with a lowering of trade restrictions are driving growth and opportunity in international supply chains, and that these conditions are likely to extend into the foreseeable future. The primary implication being that a firm's ability to effectively identify, establish, and manage relationships among multinational supply chain members is becoming increasingly attractive as an area for top management attention and investment. It is not a simple endeavor to develop this ability into a strategic core competency because of the breadth and complexity of international trade. If attained, however, such competency is not easily duplicated and as such presents significant opportunities with respect to both supplier and customer markets. Strength in this area exposes opportunities for cost, quality, and agility improvements through sourcing and for revenue growth because a firm is in position to move quickly into high-growth foreign markets.

GLOSSARY

Assist A relationship in which a U.S. company provides assistance to a foreign supplier to produce or provide merchandise that will eventually be sold and exported to the U.S. company in the United States. Duties are paid on the value of the assist.

Carrier A term denoting the firm providing transportation services to a shipper.

Commercial invoice A document required for every export and import shipment. It contains such information as country of origin, buyer and seller names and addresses, descriptions of products being shipped, harmonized tariff schedule codes, and product values.

Consignee The receiver of a shipment.

Duty drawback The recovery of duty paid on an import when the item is later exported out of the United States.

Foreign trade zones (FTZs) Specific areas within a country that have been designated as a duty-free zone.

Free-on-board (FOB) terms U.S. domestic shipping terms that determine when title and risk passes between the seller and buyer.

Freight forwarders Entities that provide domestic and international shipping services. Most do not operate their own vehicles. They obtain lower rates by acting as a consolidator and contracting with common and contract carriers.

International commercial terms (Incoterms) Trade terms used in international commerce (e.g., international equivalent of FOB terms). They define the costs, risks, and obligations of buyers and sellers in international transactions.

International freight forwarders Entities that act as middlemen between the exporter and importer and the carrier.

Harmonized tariff schedule (HTS) Used by governments to classify material and to determine duties. The schedule contains classification numbers for products. For U.S. imports these numbers are called HTS numbers, for U.S. exports these numbers are called schedule B numbers. The first six digits of the HTS number are universal. The remaining digits are country specific.

Import bond A certificate insuring that U.S. Customs gets paid the duty it is owed. Import bonds are required for imports to the United States. An import bond is also known as a surety bond.

Importer of record Generally the buyer of foreign merchandise that has the merchandise imported into its country.

Non-vessel-operating common carriers (NVOCCs) Found in ocean shipping only, these are considered common carriers, but they do not own or operate a vessel. NVOCCs and freight forwarders are similar in function and process except a freight forwarder cannot sign paperwork on behalf of the carrier whereas an NVOCC can.

Packing list A document that must accompany every export—import shipment. It identifies what is being shipped, the quantity, the shipper's reference number, and the customer purchase order number.

Shipper A term used in transport documentation to denote the seller of the merchandise.

Third party logistics firms Entities that provide comprehensive logistics services (e.g., transportation, ware housing, order processing), thus representing a significant third party complementing the parties of buyer and seller.

Trading block A group of countries that have established trading agreements for the purpose of increasing trade between the participating countries. A trading block is also known as a regional trading agreement.

U.S. customs brokers Brokers licensed to clear imports into the United States that ensure all the necessary paperwork for entry is in order.

CROSS REFERENCES

See *Managing the Flow of Materials Across the Supply Chain; Supply Chain Management; Supply Chain Management and the Internet; Supply Chain Management Technologies; Value Chain Analysis.*

REFERENCES

Advincula, R. (2001, March). *Temporary imports, ATA carnet*. Trade Information Center Web site. Retrieved November 26, 2002, from http://www.ita.doc.gov/exportamerica

Ballou, R. H. (1999). *Business logistics management*. Englewood Cliffs, NJ: Prentice Hall.

Bolero.net: Accelerating global trade. (2002). Retrieved December 10, 2002, from http://www.bolero.net/

Bowersox, D. J. (1974). *Logistical management*. New York: Macmillan.

Carbone, J. (2003, January 16). ODM's offer design expertise, quicker time to market. *Purchasing.com*. Retrieved January 26, 2003, from http://www.manufacturing.net/pur/

Chabrow, E. (2002, February 6). Survey: Internet key to collaboration. *Information Week*. Retrieved December 10, 2002, from http://www.informationweek.com/story/iwk20020206s0007

Council of Logistics Management. (2002). *Purpose and policies*. Retrieved December 10, 2002, from http://www.clm1.org/aboutUs/aboutUs_policy.asp

ebXML: Enabling a global electronic market. (2002). Retrieved December 10, 2002, from http://www.ebxml.org/

Frankel, J., & Rose, A. (2002, February 12). An estimate of the effect of common currencies on trade and income. Retrieved December 4, 2002, from http://www.wcfia.harvard.edu/rsrchpapsum.asp?id=99

Global logistics group introduces paperless trading. (2002, October 28). Retrieved December 10, 2002, from http://www.bolero.net/

Global study finds business process outsourcing is driving organizational change at the nation's largest companies. (1999, April 1). PricewaterhouseCoopers

Web site. Retrieved November 24, 2002, from http://www.pwcglobal.com

Houlihan, J. B. (1985). International supply chain management. *International Journal of Physical Distribution & Materials Management, 15,* 22–38.

International B2B e-commerce predicted to grow a surprising 76% year to year from 2000–2005. (2001). Retrieved February 3, 2003, from http://www.etranslate.com/en/press/

International gross domestic product, population, and general conversion factors information. (2000). Retrieved December 16, 2002, from http://www.eia.doe.gov/emeu/international/other.html

Linet, P. E. (2002). Customs and international trade law. Retrieved November 27, 2002, from http://www.linetlaw.com

NAFTAssistant: Even free trade has a price. (2002). Retrieved December 12, 2002, from http://psisoftware.com

Outsourcing Institute. (2002). Retrieved November 27, 2002, from http://www.outsource.com

Su, B. W. (2001, November). The U.S. economy. *Monthly Labor Review*. Retrieved November 27, 2002, from http://www.bls.gov/opub/mlr/2001/11/art1full.pdf

Tariff Act of 1930 as amended, section 508 and section 509 located in 19 USC 1508. (2002). Retrieved November 27, 2002, from http://www.gpo.gov

TDC Trade.com Business Alert. (2002, October 12). Retrieved December 4, 2002, from http://www.tdctrade.com

Testimony of Undersecretary of Commerce for international trade to the U.S. House of Representatives. (2001, June 21). Committee on International Relations, June 21. Retrieved December 6, 2002, from http://www.export.gov

Toran, M. W. (2002). Special report: Marine/aviation, maritime security takes top priority. *Risk and Insurance*. Retrieved December 2, 2002, from http://www.riskandinsurance.com

U.S. Customs. (2002). Retrieved December 4, 2002, from http://www.customs.gov

U.S. Customs Fact Sheet (2002). May. Retrieved from http://usinfo.state.gov/ December 8, 2002).

Vigoroso, M. (2002, April 2). The world map of e-commerce. *E-Commerce Times*. Retrieved November 27, 2002, from http://ecommercetimes.com

World Trade Organization. (2002). Ration of exports of goods and commercial services to GDP of least developed countries, 1990–2000 (Table III.83). Retrieved December 2, 2002, from http://www.wto.org

FURTHER READING

Hof, R. (1999, March 22). What every CEO needs to know about electronic business. Retrieved November 27, 2002, from http://www.businessweek.com

Office of Trade and Economic Analysis, Department of Commerce, U.S. Census Bureau. (n.d.) Retrieved November 29, 2002, from http://ita.doc.gov

Internet Architecture

Graham Knight, *University College London, United Kingdom*

INTRODUCTION

The Internet is a rather loose assemblage of individual networks; there is little in the way of overall administration. The individual networks are owned by a huge number of independent operators. Some of these are major corporations with large, high-capacity networks; others are private individuals operating tiny networks of two or three computers in their homes. Between them these networks employ just about every networking technology yet invented. The great strength of the Internet is that it allows these diverse networks to act together to provide a single global network service.

The interactions between a network and its neighbors are, in essence, both simple and robust. This makes for easy extendibility and fuelled the early growth of the Internet. New participants needed only to come to an agreement with an existing operator and set up some fairly simple equipment to become full players. This was in great contrast to the situation within the world of telephone networks, where operators were mostly large and bureaucratic and where adding new interconnections required complex negotiation and configuration and, possibly, international treaties.

This chapter is organized into four main sections. I begin with a discussion of network interconnection and how a new service, the IP service, enables it to be achieved. In the second section, I look at how useful applications can be built on the IP service. Next I take a more detailed look at how the IP service is provided in practice. Finally I take a brief look at how the Internet is developing into a "multiservice" network capable of carrying many types of traffic.

NETWORK INTERCONNECTION—THE IP SERVICE

The Internet story began in the early 1960s with the invention of the idea of sending data through computer networks in discrete lumps (called "packets") rather than as continuous streams of bits. In general, large objects such as computer files would need to be chopped up into several packets and reassembled at the destination. A major early experiment in packet-switching was the ARPANET instigated by Leonard Kleinrock, Lawrence G. Roberts, and others and sponsored by the U.S. Department of Defense. By the end of 1969, the ARPANET was up and running, with four computers connected. The network grew over the next few years and began to support applications such as file transfer and electronic mail. In 1972, Bob Kahn, who had been influential in ARPANET design, began to turn his attention to network interconnection or "internetting." One of Kahn's key insights was that there would, in the future, be many independent networks, each based on a different technology. He recognized that it would be impossible to require all these networks to be compatible with the ARPANET way of doing things. Instead he designed an open "Internetworking Architecture." Essentially, networks could do what they wanted internally, but if they wanted to interwork with other networks they would need to implement the interworking architecture as well. Kahn, in collaboration with Vint Cerf, went on to flesh out these basic ideas. They devised a set of rules called the "transmission control protocol" (TCP), which specified how two computers should cooperate to provide reliable delivery of data across the internetwork. Shortly after, it was recognized that not all applications would want the same levels of reliability guarantees, so TCP was split in two: one part provided reliability and retained the name TCP; the other part dealt with the delivery of packets and was called the "Internet protocol" (IP). TCP and IP have survived, unaltered in most important respects, until today—a tribute to their designers in the early 1970s.

The ideas of Kahn and Cerf were soon put into practice in a three network "internet" consisting of the ARPANET itself, a packet radio network and a packet satellite network. The latter gave the internetwork, an international reach with a node at University College London in the United Kingdom run by Peter Kirstein. Originally it was thought that one day there might be perhaps 100 participating networks. The invention of local area networks (LANs) in the 1980s changed all that and brought an explosive growth in the number of participating networks which, today, are numbered in hundreds of thousands. In 1986 the U.S. National Science Foundation funded a backbone network linking five supercomputing centers. This network (NSFNET) grew into the first real Internet backbone with connections to many countries around the world. By 1995 it was clear that the Internet, for many years a research and academic network, had a major commercial future, and NSFNET's backbone role was handed on to a set of linked commercial network providers.

Network Technologies and the Services They Provide

The fundamental components of the Internet are "packet networks" (called simply "networks" for the remainder of this chapter). This term indicates any technology that is capable of transferring data packets from one computer to another; Figure 1 shows an abstract view of such a network. Usually many computers are attached to the network, which consequently provides them with "addresses" that can be used to specify the source and destination of packets. Most of the computers connected to networks are classified as "hosts." These are computers than run programs; desktop PCs and servers come into this category. A few computers provide functions related solely to network operation. For example, "gateways," "routers," or "bridges" relay packets between networks.

There is a huge variety of technologies that fit into this general model. For example, LANs such as Ethernet (Metcalfe & Boggs, 1976) use so-called medium access control (MAC) addresses. These are 48-bit addresses allocated by the manufacturers of the interface cards that are plugged into the hosts. The "network service" provided by a LAN is simple: a computer constructs a packet including a destination MAC address and asks the network to deliver it—exactly like posting a letter. This is termed a "connectionless" (CL) service. A rather different style of service is provided by, for example, asynchronous transfer mode (ATM) (Vetter, 1995) networks. In these networks, there is an initial negotiation phase during which the computer tells the network what kind of data it wants to send and the address of the destination. This gives the computer the opportunity to specify the quality of service (QoS) it requires (bounds on throughput, transit delay, etc.). The network must decide whether it can meet the request. If it can, it tells the computer that a "connection" has been established to the destination and that data transfer can begin. This is similar to making a telephone call in which one dials a number and establishes a call before speech can begin and is termed a "connection oriented" (CO) service. Addresses in CO networks are used only during the connection set up phase.

Interconnection via "Convergence"

Virtually all networking technologies provide services similar to those just described, even though the mechanisms they use to implement the service may be radically different. Within these broad categories, however, there are many awkward differences: Each technology imposes its own limit to the size of packet that may be sent, each has its own idiosyncratic addressing scheme, some technologies allow subtle control of QoS others do not, and so on. These differences pose difficulties for network interconnection. We cannot simply take a packet from one network and forward it onto the next; all sorts of subtle transformations would be needed. The solution

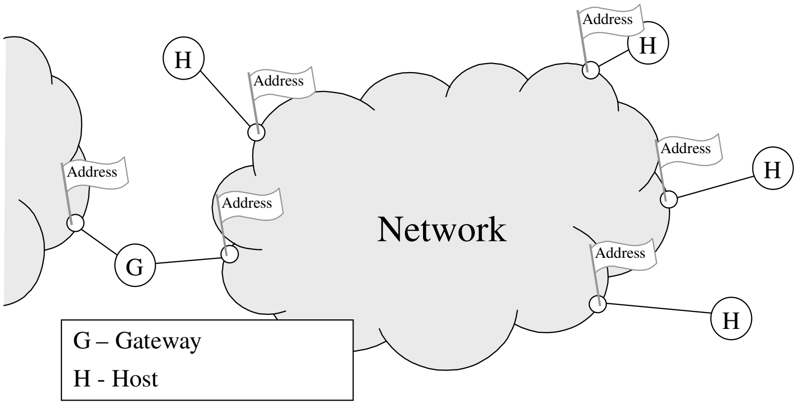


Figure 1: An abstract view of a network. Computers, termed "hosts" have physical interfaces to the network. These interfaces have addresses. The network is capable of delivering packets to hosts identified by their addresses.

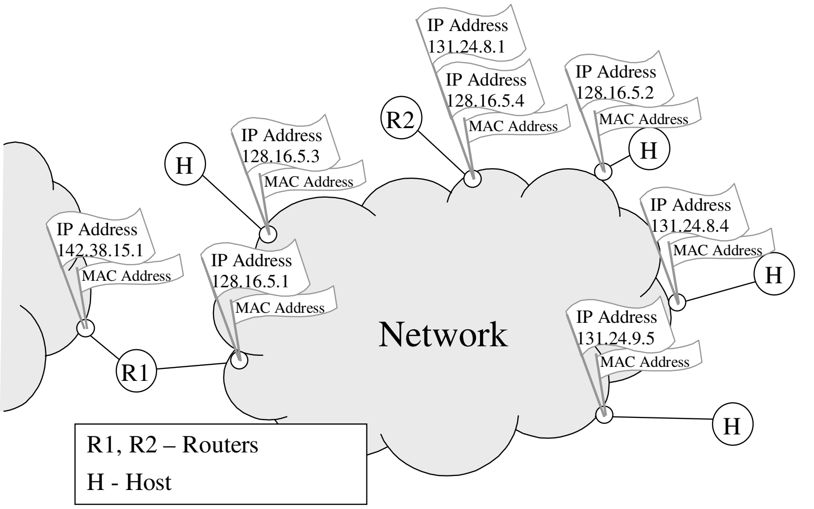


Figure 2: Internet protocol addresses allocated to network interfaces. In this case, the first 16 bits of the IP address identify the "IP Network." Thus there are three IP networks: 142.38, 128.16, and 131.24.

is "convergence." We invent an abstract network service and, for each networking technology, we provide "convergence functions" to implement the abstract service. Through these means we make all networks look the same and the interconnection problem becomes manageable.

The Internet Protocol—A Convergence Protocol

The IP provides a convergence function. Version 4 of the protocol (still the most widely deployed) and the abstract network service it implements are specified in RFC 791 (Postel, 1981a) published in 1981. (Many Internet RFCs, or Requests for Comment, are cited in this article. All are available online on the Web site of the Internet Engineering Task Force: http://www.ietf.org/rfc.html.) The service is a simple CL one that allows single packets (called IP datagrams) to be sent to destinations identified by "IP addresses." An IP address is a 32-bit quantity that is divided into two portions. The first portion, the "network prefix," identifies an "IP network" (sometimes now called an "IP subnetwork"). The second portion identifies an interface on that network. The 32 bits are normally written as four decimal bytes separated by dots (e.g., 128.16.5.2).

Hosts, IP Networks, and Routers

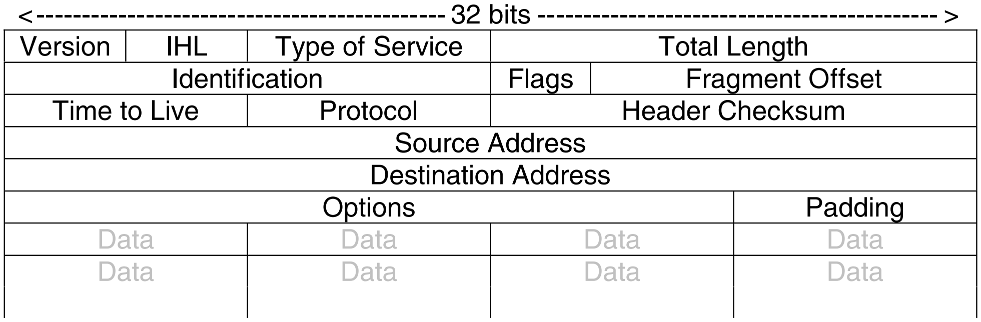
Figure 2 shows the network from Figure 1 set up to implement the IP service. Each interface has now acquired an IP address in addition to its original "technology-based" address. To make things a little more concrete, we will assume the networks are LANs, in which case the "technology-based" addresses will be MAC addresses. An important point is illustrated here: An IP network is an abstract thing and does not necessarily correspond one-to-one with a physical network. In Figure 2 we see one physical network partitioned into two IP networks (128.16 and 131.24). The reverse is also possible with several physical networks operating as a single IP network. Traffic is relayed between one IP network and the next by devices called "routers" (the term "gateway" was used in early Internet documents and survives to some extent).

The basic algorithm for delivering an IP datagram to its destination is broadly as follows:

```
If (destination network prefix = = "my"
  network prefix),
  send datagram direct to its destination
  across the local physical network
else
  choose a suitable router
  send datagram to router across the local
  physical network
```

This algorithm is executed by the source host and, if necessary, by a succession of routers along the path to the destination. Thus a datagram sent from host 128.16.5.2 in Figure 2 to host 128.16.5.3 would be sent directly while one to host 150.23.4.1 would probably be sent to router interface 128.16.5.1. Note that this router interface belongs to the same IP network as the source host, a requirement of classical IP operation. To make the configuration in Figure 2 complete we have had to add a second router (R2) so that the hosts in the 131.24 IP network can communicate with the rest of the world. This router is mandated solely by the Internet routing architecture; it provides no additional physical connectivity.

The step in the algorithm "choose a suitable router" is easily stated but less easily implemented. Sometimes there will be several possible routers, and hosts and routers must maintain "routing tables" indicating which router is the best one to use as the "next hop" on the path to a particular IP network. Because there are hundreds of thousands of IP networks in the world, maintaining such tables can be a problem. In practice, most tables contain



<-- 32 bits -->

<table>
<tr><td colspan="1">Version</td><td colspan="1">IHL</td><td colspan="2">Type of Service</td><td colspan="4">Total Length</td></tr>
<tr><td colspan="4">Identification</td><td colspan="1">Flags</td><td colspan="3">Fragment Offset</td></tr>
<tr><td colspan="2">Time to Live</td><td colspan="2">Protocol</td><td colspan="4">Header Checksum</td></tr>
<tr><td colspan="8">Source Address</td></tr>
<tr><td colspan="8">Destination Address</td></tr>
<tr><td colspan="6">Options</td><td colspan="2">Padding</td></tr>
<tr><td colspan="2">Data</td><td colspan="2">Data</td><td colspan="2">Data</td><td colspan="2">Data</td></tr>
<tr><td colspan="2">Data</td><td colspan="2">Data</td><td colspan="2">Data</td><td colspan="2">Data</td></tr>
</table>

Figure 3: Internet protocol (IP) datagram header format. Note the 32-bit source and destination IP addresses. The remaining fields are discussed briefly in the text.

entries for just a few "local" IP networks. Datagrams for all other networks are sent to a single "default" router which, it is hoped, will know what to do with them.

IP Addresses

We have assumed that the network prefix part of the IP address is 16 bits. This means we can have 16 bits to identify the individual interfaces on each network a total of $2^{16}(= 65{,}536)$, which is a large number of interfaces to have on one network. Therefore, the IP designers decided there should be three classes of IP address to suit the requirements of small, medium, and large networks. For small networks, the network prefix is 24 bits and the host part 8 bits, allowing 256 hosts (in fact, it is 254 because one is used for "this host" and another for local broadcast). Obviously routers need to know how many bits there are in the network prefix because these are the bits they look up in their routing tables. This information is encoded in the first few bits of the address according to the following scheme, where N indicates the network prefix and H the host part:

Class A	0NNNNNNN	HHHHHHHH	HHHHHHHH	HHHHHHHH
Class B	10NNNNNN	NNNNNNNN	HHHHHHHH	HHHHHHHH
Class C	110NNNNN	NNNNNNNN	NNNNNNNN	HHHHHHHH

The Internet Assigned Numbers Authority (IANA) is the top-level authority for allocating IP addresses. IANA delegates authority over blocks of addresses to regional authorities who, in turn, allocate blocks to Internet service providers (ISP). Finally ISPs assign IP addresses to end users.

The IP Datagram and the Service It Offers

The basic IP service is simple; a host constructs a packet of data and gives this to the IP service along with the IP address of the destination. The IP service then makes its "best effort" to deliver the datagram. The service offers no guarantees of success, however. It may lose the datagram without informing either the source or the intended destination, delivery may be subject to arbitrary delay, two datagrams dispatched to the same destination may be misordered on arrival—it may even deliver two copies of one datagram. If an application running in a host finds this rather cavalier attitude toward service unacceptable, it is up to the application to do something about it. For example, many applications exploit the TCP (Postel, 1981b), which keeps track of wayward datagrams, retransmits missing ones, and ensures sequenced delivery. (More details on TCP are provided later.)

In fact, simplicity is the IP service's greatest strength and one of the main reasons for its considerable success. A simple convergence protocol is easy to implement, which encourages its deployment on as wide a range of technologies as possible. Having only minimal features in the network service maximizes flexibility for applications running in hosts. For example, some applications may happily tolerate the occasional lost datagram, especially if this avoids problematic costs and delays imposed by retransmission. If the network service itself took on the responsibility for retransmission these applications would suffer. Thus the design exemplifies the end-to-end argument as propounded in Saltzer, Reed, and Clark (1984).

IP datagrams themselves have a simple structure consisting of a header (typically 20 bytes) plus a "payload," the data the datagram is carrying across the Internet. The header format is shown in Figure 3. The fields are interpreted as shown in Table 1.

Implementing the IP Service

The IP datagram delivery algorithm given earlier includes the phrase "send datagram direct to its destination across the local physical network." We must specify how this is to be done for each network technology. The three main issues involved in this are discussed next.

Encapsulation in "Native" Formats

Each network technology specifies its own packet format. Usually this consists of a header, which often contains addresses, followed by the data payload, perhaps followed by a checksum used to detect errors. In essence, encapsulation means we construct our IP datagram and make this the payload of the packet. Figure 4 shows this in the case of Ethernet (note that it is common to refer to Ethernet "frames" rather than "packets"). Encapsulation would be similar for the other LAN technologies that have been standardized by the IEEE. CRC (cyclic redundancy check) is a checksum intended to discover transmission errors.

It is important to remember the strict division of labor here. The Ethernet technology pays no regard to

Table 1 Meanings of the IP Header Fields

IP Header	Meaning
Internet Protocol (IP) Version	4 for IPv4, 6 for IPv6
IHL	The length of the IP header (normally 20)
Type of Service	This is discussed in the section on quality of service
Total Length	Total length in bytes (including the header)
Identification, Fragment offset	Used to implement fragmentation
Flags	Bits used to mark special status of a datagram
Time to Live	A counter that is reduced by one by each router through which the datagram passes; when it reaches zero, the datagram is discarded, ensuring misrouted datagrams do not circulate forever
Protocol	A number to identify the protocol that this datagram is carrying (transmission control protocol, user datagram protocol, etc.)
Header checksum	Used to pick up transmission errors that might corrupt header fields.
Source/Destination Addresses	Discussed earlier
Options	Rarely used (space precludes discussion here)

the payload; it is interested only in the Ethernet header and the CRC checksum. Likewise, Internet technology—routers and so on—is interested only in the IP datagram and its header.

In practice things may be slightly more complex than Figure 4 suggests. An Ethernet may be carrying other protocols in addition to IP datagrams. Systems processing incoming Ethernet frames need to be able to determine what protocol the frame is carrying. This entails the insertion of another header between the MAC and IP headers. This header will include a "protocol identifier," a "well-known" number indicating "this is an IP datagram."

Some technologies require special treatment. ATM, for example, uses small packets, called cells, each of which carries 48 bytes. An associated standard, ATM Adaptation Layer 5 (AAL5), explains how a larger packet, an IP datagram, for example, can be fragmented into cells and reassembled at the destination.

Address Resolution

Each network technology has its own addresses; in this chapter, these are termed "physical addresses." Obviously a technology understands its own physical addresses; no networking technology understands IP addresses, however. These are processed exclusively by the convergence function. Once the IP delivery algorithm has determined whether an IP datagram can be delivered direct or whether it must be sent to a router, we are left with the problem of determining the correct physical address to use. The business of mapping from an IP address to the corresponding physical address is called "address resolution." Two techniques are used to achieve this:

1. Most LANs have a "broadcast service." All systems attached to the LAN will receive and process a frame sent to the MAC broadcast address. For address resolution a frame is sent to the broadcast address containing an address resolution protocol (ARP) (Plummer, 1982) request for the desired IP address. The host that recognizes the specified IP address as its own replies with an ARP response containing its own MAC (physical) address.
2. For nonbroadcast network technologies it is normal to deploy an ARP server that maintains a table of mappings. Attached systems register their mappings with the ARP server at boot time. Systems that need to resolve a IP addresses send ARP requests to the server and receive ARP responses.

In both cases, mappings are cached by the requesting system for future use.

Fragmentation

Each network technology imposes its own limit on the number of data bytes that may be carried in a frame,

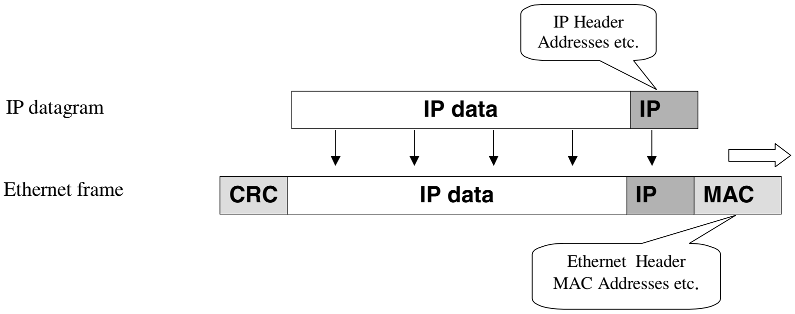


Figure 4: Simple encapsulation of an Internet protocol datagram in an Ethernet frame; the datagram simply becomes the payload of the frame.

referred to as the maximum transfer unit (MTU). These limits pose problems for IP because the path to a destination may traverse dozens of networks and the system that originates the IP datagram knows only the first hop of the route. Consequently, it cannot know the MTU of the complete path (the "path MTU"). In IP version 4 (IPv4), this is solved by allowing hosts and routers to fragment datagrams. An IP datagram that is too large will be split into two or more smaller IP datagrams (each complete with its IP header). The destination host will reassemble the original IP datagram from the various fragments.

Unfortunately fragmentation is not an efficient procedure. Suppose the originating system generates a series of IP datagrams of length 1,600 bytes, and somewhere along the path there is a network with an MTU of 1500 bytes. Every IP datagram will be split into two fragments doubling the work for the remaining routers along the path. If the originating host knew the path MTU in advance, it could have divided its data into 1,500-byte pieces, avoiding fragmentation and causing less work for the routers. A second problem arises if one fragment of a datagram is lost. Reassembly will not now be possible so all the work done by the networks and routers in delivering the other fragments will have been wasted.

The modern approach is to avoid fragmentation in routers by probing the network to discover the path MTU. This can be done by sending IP datagrams with a flag set that means "do not fragment." A router that is unable to forward such an IP datagram because it exceeds the MTU for the next hop network must discard it. When it does so, it sends an Internet control message protocol (ICMP) message (discussed in a later section) back to the originating host explaining what has happened. The originator can then try again with a somewhat smaller IP datagram. This is the approach favored in the latest version of IP, IP version 6 (Deering & Hindon, 1998), which actually prohibits fragmentation in routers.

BUILDING ON THE IP SERVICE

It is clear that the service provided by IP is inadequate for some applications. Several IP datagrams may be needed to carry one Web page. We care that all these datagrams are delivered exactly once and in the right order. There are many applications that have similar requirements. Rather than each of them inventing its own mechanisms to meet them a common protocol—TCP—is used.

TCP—The Transmission Control Protocol

TCP is an "automatic repeat request" (ARQ) protocol; such protocols send packets of data that each include a sequence number. It is up to the receiver to check the sequence numbers, ensure that all packets have arrived, and reorder them if necessary. The receiver sends "acknowledgment" packets back to the transmitter. The transmitter retransmits packets that are not acknowledged sufficiently promptly. In the TCP case, the packets are termed "segments," each of which is carried in an IP datagram. Each segment consists of a TCP "header" (containing the sequence and acknowledgment numbers) that may be followed by a block of data (see Figure 5). At any instant there are likely to be data bytes that have been sent but not yet acknowledged. A limit, termed a "window," is set on the number of these bytes. This window is specified by the receiver in the acknowledgments it sends. By reducing this window the receiver may slow down or even stop the flow of data. This flow control mechanism is essential in preventing a slow receiver from being overwhelmed by a fast transmitter.

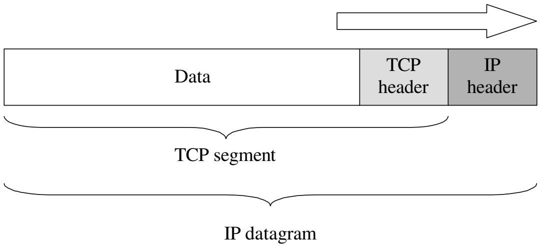


Figure 5: A transmission control protocol (TCP) segment carried in an Internet protocol (IP) datagram. Coincidentally, both the TCP and IP headers are normally 20 bytes long. The data part may be any length subject to a limit imposed by the network technology in use. See also Figure 4.

The TCP protocol, although it uses the IP service, is independent of it. The components that implement the IP service, including all the routers, are interested only in the IP header—the fact that a particular IP datagram is carrying a TCP segment is of no importance whatsoever to these components. The components that are interested in TCP are the pieces of software that implement the logic of the TCP protocol. These exist in the end systems or "hosts"; typically such software is supplied as part of an operating system such as Linux or Windows.

This independence between the TCP "layer" and the IP "layer" is illustrated in Figure 6. TCP is just one possible "transport protocol" and will be used for most applications that require reliable, sequenced delivery. Logically we can think of the two TCP modules in the hosts as exchanging segments and acknowledgments directly with each other (represented by the broken, horizontal arrow). Physically, however, a TCP module must pass its segments down to the local IP module (also a piece of software within the operating system) that will wrap it in an IP datagram and send it on its way. The IP module at the receiver will unwrap the IP datagram, retrieve the TCP segment, and pass this up to the local TCP module. This is represented by the vertical arrows.

TCP's Role in Congestion Control

A weakness in the Internet architecture is the limited control over congestion it affords. Just as in road transport networks, congestion occurs when too much traffic attempts to pass through a particular point. Were the IP layer CO, there would be an opportunity to refuse new connections if these seemed likely to overload some router along the path. The Internet's basic CL service has to accept all the traffic that is thrown at it, however, and do the best it can. TCP can help alleviate this problem. As part of its normal operation a TCP module becomes aware of dropped or delayed IP datagrams because these manifest

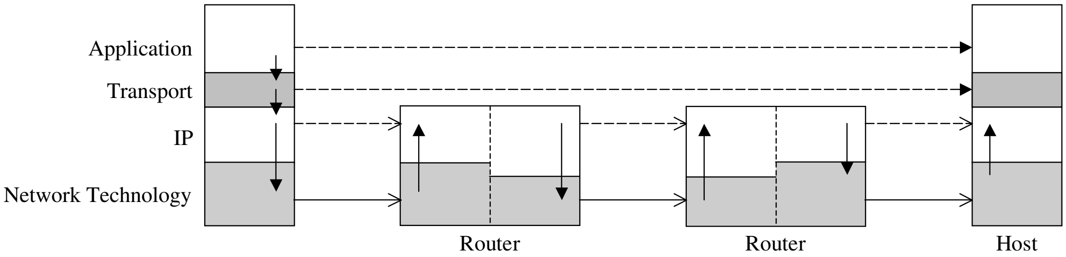


Figure 6: The fundamentals of the Internet architecture. The Internet Protocol (IP) layer provides convergence, allowing many network technologies to take part. It also provides the relaying operation from router to router toward the destination host. Above IP comes the "transport" layer," which is where TCP fits. This layer is entirely the responsibility of the hosts.

themselves as missing or delayed TCP acknowledgments. When these occur, TCP concludes (usually correctly) that congestion is occurring and employs a "congestion control" algorithm attributable to Jacobsen (1988). This algorithm results in TCP temporarily slowing its data transmission rate. If all the TCP connections passing through the congested router behave in this way, the congestion should be relieved.

This mechanism has been observed to work well in practice. Unfortunately there has been a recent growth in real-time applications which, typically, use user datagram protocol (UDP; Postel, 1980) rather than TCP. UDP is almost a null protocol and lacks all of TCP's reliability and congestion control mechanisms, leaving such issues to the applications themselves. Efforts are being made to endow such applications with "TCP-like" behavior, but it seems unlikely that this alone will suffice to deal with the congestion problem. Some kind of resource management inside the Internet seems essential; approaches to this are outlined in the final section of this chapter.

Ports—Identifying Applications and Processes

Hosts are identified by IP addresses. Assuming we know the IP address of our intended recipient, the IP layer will (usually) deliver our IP datagram successfully. We need also to identify a particular application or process present on the host, however. This is the role of transport layer ports. Here one can draw a close analogy with telephone extensions; the telephone number (IP address) identifies the building, the extension number (port) identifies an individual within the building. All Internet transport protocols include source and destination ports in their headers; these are simply 16-bit integers.

A "client-server" model is assumed; the client is the piece of software that initiates requests, the server software waits and responds to client requests as these arrive. Public services usually use "well-known" ports (see Table 2). Thus, when a browser (client) requests a page from a Web server, the TCP segments it sends will have the destination port set to 80 in the TCP header. The operating system on the server host will have a table that tells it that port 80 is the WWW server. It will ensure that the WWW server software receives the incoming data and processes it. The client system chooses a source port. It should choose a value that is not currently in use for any other purpose. This source port will be used as the destination port in any replies from the server.

Other Internet Transport Protocols

Some applications do not need the reliable sequenced delivery provided by TCP, especially because TCP buys this reliability at the cost of delays when retransmissions occur. A great strength of the Internet architecture is that such applications can select a different transport protocol. The simplest is the UDP, which adds nothing to the basic IP service; however, its simple header does include the ports necessary for application selection. UDP is a popular choice for delay-sensitive applications and also for applications that just need to send one message and get one reply. It is sometimes favored by applications that work mostly within a LAN; for example, the SUN Network File System (NFS) can use UDP instead of TCP.

Several other protocols are in widespread use, including the real time protocol (RTP; Schulzrinne, Casner, Frederick, & Jacobson, 1996), which provides sequence numbering and time stamping for applications such as streamed audio and video. Sometimes two transport protocols are used in tandem: RTP is often carried inside UDP datagrams; Transport Layer Security (TLS; Dierks & Allen, 1999), which provides authentication and confidentiality, may be used in conjunction with TCP.

Table 2 Some "Well-Known" Transport Ports and Associated Protocols.

SERVICE	PROTOCOL	"WELL-KNOWN" PORT
HTTP (Web server)	TCP	80
FTP (FTP server)	TCP	21
Telnet (Remote rlogin)	TCP	23
SMTP (e-mail)	TCP	25
DNS (Name and address mappings)	UDP	53

Note. A client wishing to contact one of these services will use these as the "destination port" in the messages it sends. DNS = domain name service; FTP = file transfer protocol; HTTP = hypertext transfer protocol; SMTP = simple mail transfer protocol.

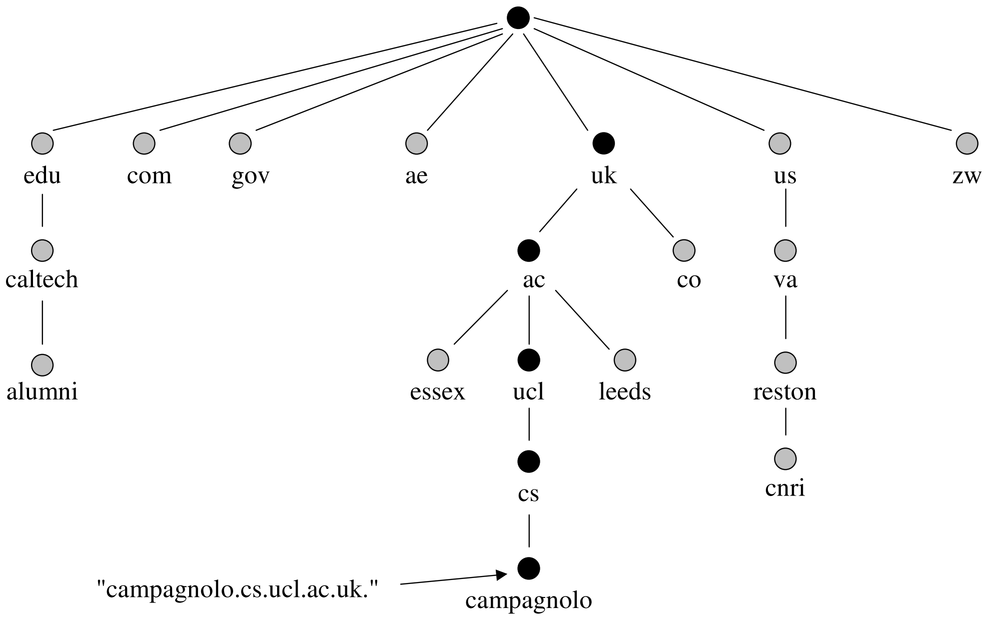


Figure 7: A portion of the domain name tree.

Naming Internet Objects

We have seen that IP addresses provide globally unique labels for the interfaces between computers and the Internet. Such numeric labels, like telephone numbers, are efficient for computers but not especially memorable for humans. They are also liable to change according to technological diktat—for example, when a computer is moved to a new IP network. Furthermore, the Internet makes many other sorts of object accessible—networks, e-mail accounts, files, and so on—and we need globally unique labels for these. Internet domain names (DN; Mockapetris, 1987a) solve these problems by providing a systematic way of generating globally unique and reasonably user-friendly names. Like other schemes with similar objectives the DN scheme relies on a recursive devolution of authority over the namespace.

Figure 7 shows a portion of the domain namespace. Each node (leaf or nonleaf) represents a domain that consists of itself and all the nodes in the subtree beneath it. A fully qualified domain name is a concatenation of the labels on the path from a node all the way back to the root (e.g., *campagnolo.cs.ucl.ac.uk*). The authority for a zone is responsible for ensuring that no two sibling nodes have the same name. The top-level domains include the worldwide generic domains (*com*, *org*, *net*, *edu*, etc.), two U.S.-only generic domains (*mil* and *gov*), and country code domains (e.g., *us* for the United States, *uk* for the United Kingdom, etc.). Each authority has its own rules for name allocation and each may delegate authority over subdomains. For example, in the United Kingdom, the Department of Trade and Industry, which administers the "uk" domain, has allocated an "ac" subdomain and has delegated authority over it to a joint-board of UK universities. This board has allocated domains, and delegated authority, to individual universities.

Although the principal use of DNs is the naming of computers and interfaces, they can be used for any purpose. Furthermore, there are no inherent relationships between domains and IP networks. It happens that the computer named "*campagnolo*" with address 128.16.23.3 belongs to the Computer Science department of University College London, an academic institution in the UK and its DN, *campagnolo.cs.ucl.ac.uk*, reflects this closely. There is nothing to stop the department from allocating the name to a computer on some other network in some other country, however; the name is theirs to use as they like.

The Domain Name System

The domain name system (DNS; Mockapetris, 1987b) is a distributed directory service that supports, among other things, the automatic mapping of DNs to IP addresses. Generally speaking, each domain must provide a computer on which runs a DNS "nameserver" that maintains a table of mappings for the names in its domain. This is termed the primary nameserver for the domain. The DNS protocol allows hosts to query these mappings and so to map names onto addresses. Local DNS servers hold mappings for local machines and answer queries for these directly. Nonlocal names are passed on to other DNS servers. Figure 8 illustrates this when a local nameserver cannot resolve the DN *www.ibm.com*. Steps 2 through 5 show the iterative mode; the query is passed to the *root* nameserver for the *com* domain which refers its questioner to a different nameserver. The local nameserver will cache the result of the query so that, next time around, we will see, at most, steps 1 and 6. As with all caching strategies, there is a danger that the cached information will become out of date, so a timeout is used. In addition to this, nameserver responses always contain a flag to indicate whether they are *authoritative,*

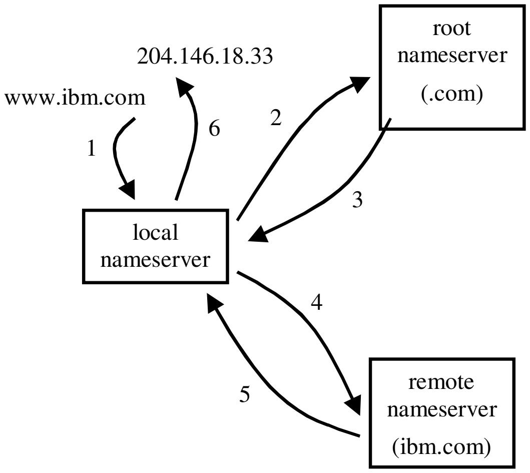


Figure 8: Distributed operations in the domain name system. The name cannot be resolved by the local nameserver, which passes the query to the nameserver for the top-level (com) domain and so on.

that is, whether they come direct from a primary nameserver.

The DNS can answer a number of other queries. Especially important are "MX" mappings. These map a DN used in an e-mail address to the name of a host willing to accept mail for the domain. Thus an e-mail client wishing to deliver a message to *bloggs@cs.ucl.ac.uk* can discover the mapping *cs.ucl.ac.uk* ⇒ *smtp.cs.ucl.ac.uk*.

Universal Resource Locators

Universal resource locators (Berners-Lee, Fielding, & Masinter, 1998) are the familiar URLs used on the World Wide Web. The URL format is flexible; however, most URLs that users encounter have the following format:

```
<protocol>://<domain name>/<local name>
```

The `domain name` is a DN as described in the previous section. It identifies a host on which runs a Web server program that has access to the resource (a file for example) identified by `local name`. Although it is common for hosts used for this purpose to be given a DN starting "WWW," there is no particular need for this, and any name will do. The `protocol` indicates the application protocol to be used when contacting the WWW server. These days this is usually the hypertext transfer protocol (HTTP; Fielding et al., 1999), described in a later section It is still quite common, however, to see FTP, the original Internet "file transfer protocol" (Postel, 1985).

Internet Applications

Like transport protocol modules, applications are implemented by software residing in hosts. Application software exchanges messages across the Internet. In the most widely used applications, these messages are formatted using structured ASCII text. For example, a browser application (Internet Explorer, Netscape etc.) uses HTTP to retrieve information from a Web server. An example of an HTTP "GET" request message is the following:

```
GET http://www.cs.ucl.ac.uk/research/index.html HTTP/1.0
Authorization: Basic QWxhZGRpbjpvcGVuIHNlc2FtZQ==
```

HTTP employs TCP in the transport layer. Therefore HTTP messages like the one above (encoded into binary using ASCII codes) will form the data part of a TCP segment. If all goes well, the HTTP response will include the file "index.html." Typically this is likely to be too big to fit into a single TCP segment, so several will need to be sent. Other familiar applications (e.g., FTP for file transfer and simple mail transfer protocol [SMTP] for e-mail; Klensin, 2001) use similar protocols. Designing such protocols is a matter of deciding what messages need to be sent, under what circumstances each needs to be sent, and specifying the format to be used.

Some application protocols provide generic services on which specific applications can be built. Among this group are the Remote Procedure Call and Remote Invocation protocols such as CORBA and Java RMI. These adopt an object oriented model with one part of the application invoking operations on objects on a remote host. Each operation requires parameters to be sent and results to be returned. CORBA and Java RMI provide general mechanisms for handling these aspects.

In most operating system environments applications exist in user (rather than kernel) space. To operate they need to interact with the transport layer modules that usually reside in the kernel. The operating system will provide a set of system calls to support basic operations such as opening a TCP connection, sending or receiving a block of data, and so on. The so-called socket system calls, which originated with Berkeley Unix, are a widely adopted standard.

MAKING THE IP LAYER WORK

The previous sections established the basic components of the Internet architecture:

- An Internet layer, which allows IP datagrams to be sent to destinations identified by IP addresses;
- A transport layer, which enhances the Internet layer service to provide a service satisfying an application's requirements; A variety of transport protocols exist, each of which resides exclusively in hosts and operates end to end
- An application layer, which includes multiple applications, each of which employs a suitable transport layer

This section considers the operation of the Internet layer in more detail.

IP Routers

I have already described the basic algorithm for IP delivery. Every IP router implements this algorithm and applies it to each IP datagram it receives. The router must make such a "forwarding decision" for every IP datagram it receives. The router must look at the destination IP address, extract the network prefix, and look this up in its routing table. As a result of this look-up, the router will know the IP address of the "best" next-hop router (or host). Once this has been accomplished, all necessary fragmentation, encapsulation, and address resolution can be performed and the IP datagram dispatched. This style of operation is called "incremental routing" because each router concerns itself only with the next hop of the route and not with the complete route.

In this section I take a second look at network prefixes and consider how routing tables can be maintained.

IP Networks, Subnets, and Supernets

Recall that an IP network is a logical entity—a collection of interfaces in which all IP addresses share a common network prefix. The three IP address classes described earlier have proved insufficiently flexible. Class C IP networks (254 hosts) are too small for many organizations, whereas class B IP networks (65534 hosts) are too big. The problem is exacerbated by the fact that there can only be 2^{14} (= 16384) class B IP networks, and most of these have already been allocated. Consequently, most allocations today are for class C IP networks, and large organizations end up with multiple class C IP networks. This causes problems with the "classical" Internet delivery algorithm presented earlier because it insists that an IP datagram sent between distinct IP networks must travel via at least one router. On the face of it, this will force the introduction of many routers simply to satisfy the model assumed by the delivery algorithm. At the other extreme, an organization lucky enough to have a class B network may nevertheless want to install internal routers. These may be needed to allow interconnection between two dissimilar network technologies such as Ethernet and ATM. Another motivation is the desire to partition a large LAN into several smaller LANs; large LANs can suffer from many "collisions" as multiple hosts compete to send traffic. The classical delivery algorithm requires that distinct network prefixes must be used on either side of such routers,–so it is necessary to ask for additional IP networks.

The solution to these problems is to abandon the class-based address scheme and to employ subnets and supernets.

Subnets

Suppose an organization possesses the class B IP network 128.16 and wishes to operate internally as several small networks linked by routers. Using subnetting the organization may decide that, internally, 20 (for example) rather than 16 bits should be used for the network prefix. This allows 16 subnets with network prefixes (in binary) as follows:

10000000.00010000.0000
10000000.00010000.0001
...
10000000.00010000.1111

A new notation is needed to specify an IP subnetwork address under this scheme. Using this notation the second address above would be written 128.16.16.0/20 whereas the original class B address would be written 128.16.0.0/16. An alternative (and obsolescent) way of expressing the same thing is to give a "subnet mask." This is an IP address with 1s in place of all the network prefix bits, 0s otherwise. The mask for the scheme above would be 255.255.240.0 (11111111.11111111.11110000.00000000 in binary).

Table 3 16 Class C IP Network Addresses Sharing the Same First 20 bits.

11010010	00110010	10100000	210.50.160
11010010	00110010	10100001	210.50.161
11010010	00110010	10100010	210.50.162
11010010	00110010	10100011	210.50.163
...	...	...	...
11010010	00110010	10101111	210.50.175

Of course none of this will be effective unless all the routers and hosts involved know that 20-bit network prefixes are in use. They must be told this explicitly—an extra piece of configuration for each participating system.

Note that subnets are an internal matter. To the outside world the network just described still appears as a single class B IP network. This is valuable because it means that external routers need only have one routing table entry for the whole IP network.

Supernets

If care is taken in how one allocates multiple class C IP networks to an organization, they can be made to look like a single, larger IP network. Table 3 shows 16 class C IP network addresses chosen so that they share the same first 20 bits.

This can be treated as a single "supernet" with address 210.50.160.0/20—an example of so-called classless interdomain routing (CIDR; Fuller, Li, Yu, & Varadhan, 1993). Unlike subnets, supernets generally are visible externally. It is important that they are visible because this means that an external router need have just one routing table entry for the supernet rather than 16 entries for the individual class C IP networks.

Routing Tables

Figure 9 shows a configuration of networks and subnets that we can use to illustrate the use of routing tables. (In reality, there would be many more hosts; most have been omitted for clarity). We assume this configuration is operated by a single administration that has decided to partition its class B IP network into several subnets. The main partition is into three 20-bit subnets, two on the main site and one for a set of remote machines which are accessible via ISDN (integrated services digital network). (Note that ISDN is not an IP network, it merely provides links between a router on the main site and those at the remote sites.) The "remote" subnet is further partitioned. This kind of sub-subnetting assists with routing.

Consider the transmission of an IP datagram from host *dharu* on the main site to host *mavic* on remote site B. The routing table on *dharu* can be very simple (Table 4). Our datagram for *mavic* (128.16.23.2) matches the default

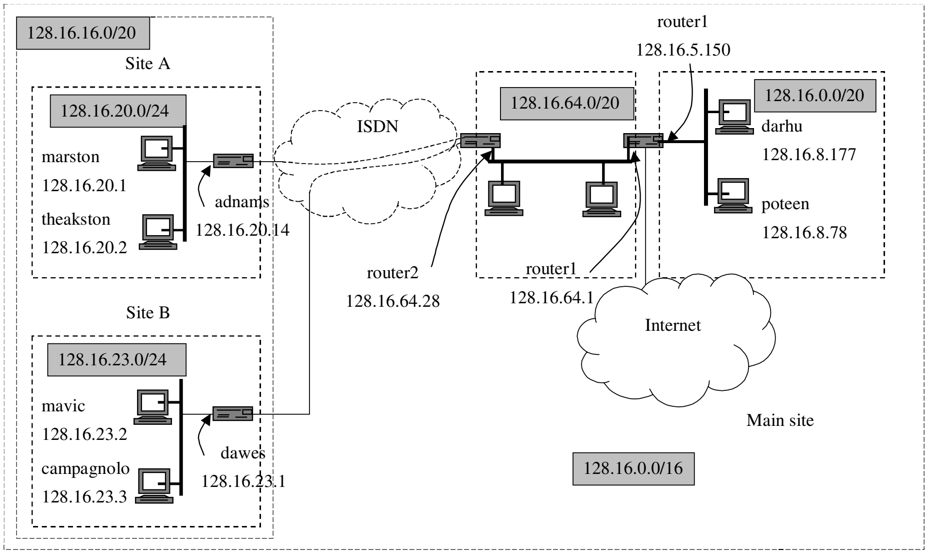


Figure 9: A class B Internet protocol network (128.16.0.0/16) partitioned into two subnets for the main site (128.16.0.0/20 and 128.16.64.0/20) and one for remote sites (128.16.16.0/20). The remote site subnet is further partitioned into two 24-bit subnets.

entry and so will be forwarded to *router1*. *Router1* is the main router giving access to the rest of the Internet. Its routing table could be large but must include at least the entries shown in Table 5. These include entries for the two local subnets and the entire remote subnet; notice that there is no need to have separate entries for each individual remote site. Our datagram matches the first entry and so will be sent to *router2*. *Router2* does need to distinguish between the various remote sites (Table 6). Here ISDN1 and ISDN2 represent the two ISDN calls. These are dial-up calls and would be managed so that they were only set up when data needed to be transferred. Our datagram matches the third entry and so will be sent to the router called *dawes*.

Suppose now that *mavic* is important so that we wish to have a permanent connection to it rather than relying on ISDN; we might lease a line between *mavic* and *router2*. We would now need to modify *router2*'s table (Table 7):

In Table 7 we now have a special entry for *mavic* (note that there is no need to put anything in the "next hop" column because there is only one possible destination at the other end of the leased line). Our datagram now matches two entries; there is a 24-bit match for the third entry and a 32-bit match for the fourth entry. This is a common situation and the tie-break rule is "choose the longest prefix match".

In some cases another factor is included in the forwarding decision a "metric" signifying the cost of a particular route. This can be used to choose between two otherwise equivalent routes to a destination. Many metrics can be used. The simplest is a hop count, an estimate of the number of routers along the route to the destination. Other possibilities are estimates of delay, error rate, or throughput along a particular route. Some systems maintain multiple metrics and choose which one to use according to the QoS to be given to the datagram (this is discussed in more detail in a later section). If we are using a dynamic routing protocol (see next section), the metrics may vary with time to reflect changing network conditions.

It is clear from this discussion that the forwarding decision can be complex. Because this decision must be made for every datagram passing through a router, its efficiency is crucial to router performance. Router manufacturers expend a great deal of effort in devising clever ways of speeding up the forwarding decision and so increasing the throughput of their systems.

Routing Information

It is crucial to Internet operation that information in routing tables is both timely and accurate. If routers have inconsistent information, it is possible for a datagram to circulate among a group of routers without ever reaching its destination—a "routing loop." The problem is compounded by the fact that the Internet is controlled by thousands

Table 4 Internet Protocol Routing Table for Host *dharu*

NETWORK	INTERFACE	NEXT HOP
128.16.0.0/20	Ethernet	Direct
Default	Ethernet	128.16.5.150

Table 5 Internet Protocol Routing Table for Router 1

NETWORK	INTERFACE	NEXT HOP
128.16.0.0/20	Ethernet2	128.16.64.28
128.16.0.0/20	Ethernet1	Direct
128.16.64.0/20	Ethernet2	Direct

Table 6 Internet Protocol Routing Table for Router 2

NETWORK	INTERFACE	NEXT HOP
128.16.64.0/20	Ethernet	Direct
128.16.20.0/24	ISDN1	128.16.20.14
128.16.23.0/24	ISDN2	128.16.23.1
Default	Ethernet	128.16.64.1
. . .	. . .	. . .

of independent and autonomous administrations. A section of the Internet run by a single administration is called an autonomous system (AS). An AS is free to decide for itself how to manage routing internally. It may employ "static routing" whereby routes are worked out by hand and the corresponding routing tables are installed in all the administration's routers. Alternatively, the administration may decide to adopt a "dynamic routing protocol" that enables routers to exchange information on the current state of the network and so derive routing tables according to some algorithm. Most routers within an AS have detailed knowledge only of networks and subnets belonging to the home AS. Traffic for a network in "foreign" AS will be sent, by default, to a router that acts as an "exterior gateway," that is, one that has a link to an exterior gateway belonging to a neighboring AS. These exterior gateways exchange information on which networks they can reach. These exchanges enable exterior gateways to learn plausible routes for reaching any network on the entire Internet.

Dynamic Routing and Routing Protocols

The original Internet architects were keen to ensure that, as far as possible, the Internet would be self-managing. For example, if some Internet component failed or its performance was degraded, the network should automatically reconfigure so as to direct traffic away from the problematic component. Dynamic routing protocols can help to achieve this. Early work on the ARPANET adopted a routing protocol of a class called "distance vector." These work as follows.

Periodically each router receives reports from its neighbors containing estimates of the neighbors' "distances" from other networks. The information in these reports is combined with information about local links so as to derive a new routing table. For example, if a router measures its current distance to neighbor X as 10 and X reports its distance as 100 from network Y, this means that Y is reachable via X at a distance of 110. Various measures of distance have been tried. The original ARPANET algorithm used queue length—the number of datagrams awaiting transmission on a link. The routing information protocol (RIP; Malkin, 1998), which is still in widespread use, uses a simple hop count. Changes occurring in network state are reflected in local measurements made by routers. These, in turn, are reflected in the information sent to neighbors in the next periodic routing update. Unfortunately, this means that many periodic updates must take place before all routers become aware of a change in network state. Until this occurs routers will have an inconsistent view of network state and routing loops can occur. This and other problems with distance vector algorithms has led to widespread adoption of an alternative class of algorithm called "link state." The "open shortest path first" (OSPF; Moy, 1998) routing protocol is of this type.

Table 7 Modified Internet Protocol Routing Table for Router 2

NETWORK	INTERFACE	NEXT HOP
128.16.64.0/20	Ethernet	Direct
128.16.20.0/24	ISDN1	128.16.20.14
128.16.23.0/24	ISDN2	128.16.23.1
128.16.23.2/32	Leased	—
Default	Ethernet	128.16.64.1
. . .	. . .	. . .

In a link state algorithm each router measures the state of its local links; again, various measures of "distance" could be used. This information is then flooded to all other routers in the network. The result is that each router receives link state information from all other routers and from this information a routing table is calculated. Because all routers receive the same information, routing loops should not occur. Routing table calculation usually employs an efficient algorithm attributable to Dijkstra (1959).

Routing protocols such as RIP and OSPF are used within an AS and are referred to generically as "interior gateway protocols." Another class of protocol, "exterior gateway protocols," serves to exchange routing information between ASs (e.g., border gateway protocol [BGP]; Rekhter & Li, 1995). These protocols carry "reachability" information—lists of networks known to an AS and for whom it is prepared to carry traffic. Many issues, including commercial and security issues, have an impact on what is included in such lists.

The Physical Structure of the Internet

The Internet has a loose structure. Its topology is mainly a consequence of perceived commercial need being satisfied by a network service. The result is a broadly hierarchical structure, however, with some operators concentrating on providing "core" networks, others concentrating on connecting private users, and still others providing transit between users and the core. Figure 10 illustrates a possible configuration.

The core networks (called national service providers [NSP] in the United States) compete with each other to carry Internet traffic. They include internetMCI, SprintLink, PSINet, and UUNet Technologies. They have high-capacity links between their routers—capable of operating at gigabits per second and more. NSPs connect with each other in two ways:

- Through independent "network access points" (NAP). These are commercial operations and provide high bandwidth switching between networks. They are generally confined to a small geographical area (e.g.,

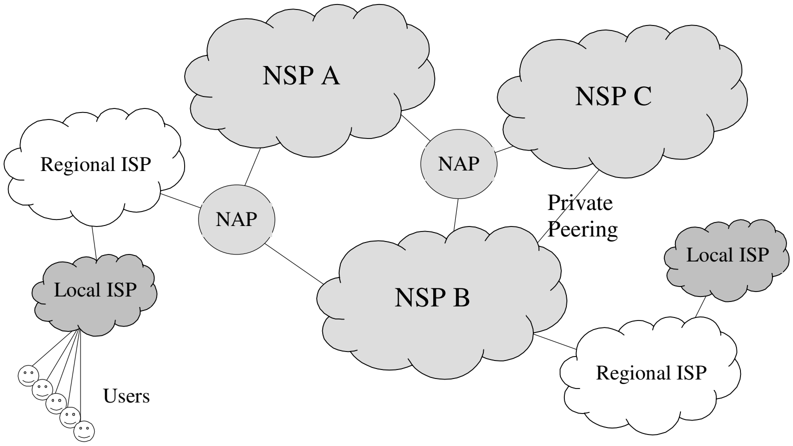


Figure 10: High-level Internet structure. The networks shown may have complex structure internally. Each is run by a separate administration and is an "autonomous system" in Internet terminology.

a single building) and consist of high-speed networks such as gigabit Ethernet that host routers belonging to the NSPs. NAPs are sometimes called metropolitan area exchanges (MAEs).

- Through private "peering" arrangements. These are direct connections between routers belonging to the NSPs.

Regional Internet service providers (ISPs) are next in the hierarchy. As the name suggests, they cover a smaller geographical area than an NSP. To provide a service, they must connect to an NSP. This may be either direct or through a NAP. Local ISPs connect end users to the Internet. This may be via dial-up (for many private users) or leased line (for corporate users).

Naturally all this must be paid for, and the ultimate source of revenue are the customers of the ISPs. It is not easy to discover precisely how money flows between the providers because such information is commercially sensitive. In principle anyone could construct a NSP or ISP and negotiate with existing operators to obtain interconnections. Existing operators may or may not welcome the newcomer—it is all a matter or commercial policy.

Peering arrangements vary. Two operators may simply agree to carry each other's traffic on a quid pro quo basis, or there may be payments in one direction or the other. Possibly an operator may agree to carry only certain traffic, based on source or destination for example. DIFFSERV (discussed in a subsequent section) will make all this more complex because it allows packets to be marked for special treatment. Operators will need to establish service-level agreements and accounting procedures to deal with this.

Continual improvements have been made in access technology that carries traffic between the customers' premises and the ISP. Much of this exploits the existing telecommunications infrastructure. The public switched telephone network (PSTN), in conjunction with MODEMs and ISDN, provide the ubiquitous dial-up service. This service provides ISPs with a convenient charging model based on "connect time" but takes no account of the volume of datagrams sent and received (and hence the real load imposed on the network). The point to point protocol (PPP; Simpson, 1994) is used over these connections; it provides access control and delimits datagrams on the wire. Faster access can be provided by asymmetric digital subscriber loop (ADSL) and cable modems that exploit telephone wiring and television cable, respectively. With these technologies, the full path from the customer's equipment to the ISP often involves other technologies such as Ethernet and ATM. Standards exist that specify how PPP may be used over these technologies, and this is evidently desirable because of PPP's access control and, not least, its familiarity (this does result in somewhat baroque architectures, however).

Variations on the Theme—Firewalls, Tunnels, Virtual Private Networks, Network Address Translation, and Proxies

Internet architecture developed in an environment in which maximum openness and wide connectivity were principal objectives. Furthermore, the network was small, and there seemed to be few architectural limits to expansion. The situation has now changed; wide connectivity is frequently seen as a security threat, and the recent vast expansion has seen the depletion of the IPv4 address space. In this section, I discuss some adaptations (some would say distortions) of the architecture in response to this changing situation.

It is common to deploy a firewall router, which selectively discards datagrams. Such routers are placed so as to control access between a site and the wider Internet, usually with the aim of improving security. Control can vary in complexity. A simple control is to specify which ports can and cannot be used. For example, allowing only destination port 80 on outgoing datagrams effectively restricts outgoing traffic to World Wide Web requests. Controls may also operate on IP addresses and protocols. Firewalls

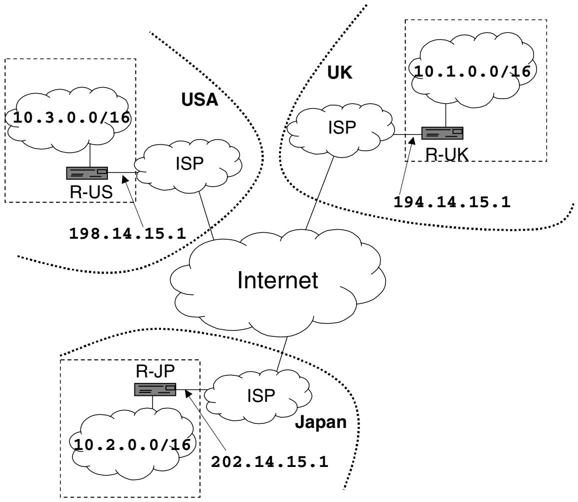


Figure 11: A virtual private network built on the Internet. An organization with three sites finds the Internet to be a cost-effective way of providing connectivity between them. The arrangement allows traffic to flow between the sites but excludes traffic to or from the wider Internet.

may be operated by ISPs as well as end users. The aim then is either to protect end users from attack or to promote the ISP's commercial interests by restricting the activities available to some classes of user.

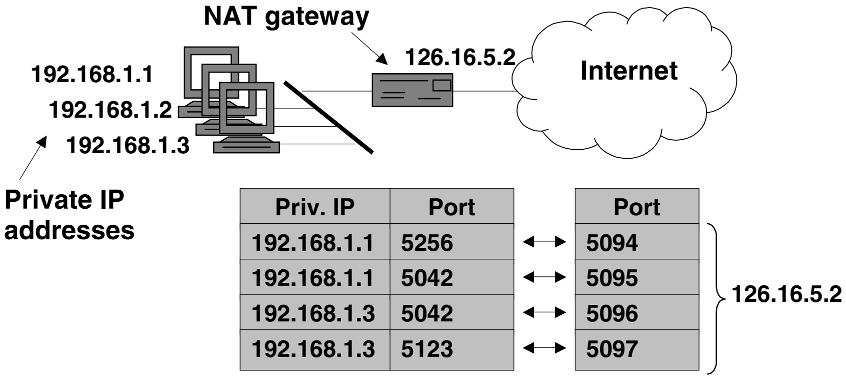
A way of exploiting the Internet while preserving independence from it is to build a virtual private network" (VPN) from IP-in-IP tunnels. Figure 11 shows such an arrangement in which an organization has three widely spaced sites each with its own network of computers and a connection via a router to a local ISP. Hosts on the organization's networks are allocated addresses from a network specifically designated to be private (such addresses should never appear on the "real" Internet). A datagram sent by a host on the U.S. site to one in the United Kingdom will be routed to the router *R-US*. This router, like the other two, has a real IP address allocated by the local ISP. Its routing table tells it that the datagram must be sent to the router *R-UK*. To achieve this the datagram, complete with its header, becomes the data part of a new datagram with destination address 194.14.15.1. When the datagram reaches *R-UK* it is "unwrapped," and the original datagram is forwarded on to the United Kingdom network. If desired, the three external routers can be configured to disallow all traffic to or from the rest of the Internet. A further refinement is to use "secure tunnels" by exploiting the IPSec standards (Atkinson & Kent, 1998; discussed later) and encrypting the tunneled datagrams.

The private address space can also help a site which has insufficient addresses for local hosts; for example, many consumer ISP contracts allocate just one IP address per customer even though the customer may have more than one host requiring an IP address. This problem can be solved in part by network address translation (NAT).

Figure 12 shows one way of using NAT as specified in RFC1631 (Egevang & Francis, 1994). It shows a single "real" IP address being shared by several hosts on a local network each of which has a "private" IP address. As packets leave the site, the NAT gateway changes their source addresses in the IP header to the real IP address that has been assigned. The reverse transformation must be applied to the destination addresses of incoming packets; the problem is to know which private IP address should be selected. This is achieved by including the transport protocol ports in the mapping. The mapping ensures that all source ports currently in use are distinct. These source ports will be the destination ports of incoming IP datagrams and can be used to select the correct mapping. Thus an incoming datagram with destination port 5095 will have its destination IP address and port set to 192.168.1.1 and 5042, respectively.

NAT is a rather ugly fix and breaks an architectural rule that IP addresses should identify the end points of a communication. Some applications, notably FTP, make use of this rule, and NAT gateways have to treat FTP specially as a consequence. The number of such special cases seems likely to increase in the future.

The NAT operation described is fine provided the applications act as clients; but suppose a user wanted to run a Web server on the machine with address 198.168.1.1. The NAT gateway would map this address so that the



Priv. IP	Port		Port
192.168.1.1	5256	↔	5094
192.168.1.1	5042	↔	5095
192.168.1.3	5042	↔	5096
192.168.1.3	5123	↔	5097

Figure 12: Network address translation. The private IP addresses and ports used locally are mapped to/from real IP addresses and ports as they pass through the NAT router.

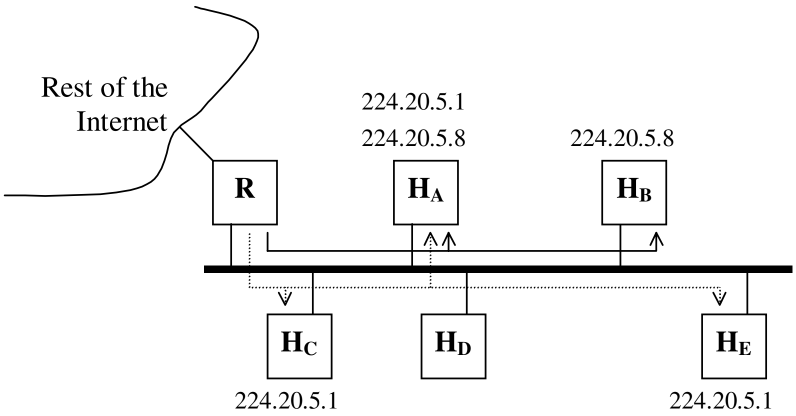


Figure 13: Multicast delivery to local area network hosts.

outside world sees the server running at address 126.16.5.2, which is fine. The user, however, would like the server to appear to listen on the "standard" WWW port, port 80. This means a permanent mapping in the NAT gateway is necessary so that 198.168.1.1 port 80 always maps to 126.16.5.2 port 80. This demonstrates that NAT gateways have a natural firewall capability; if the NAT gateway does not have a mapping for port 80, no incoming Web request can succeed.

It is in no one's interest to generate unnecessary Internet traffic. World Wide Web proxy servers situated on a customer or ISP network aim to confine a sizable portion of Web traffic to the local area. Proxy servers maintain a cache of popular, recently accessed Web pages. Browsers are configured so that Web requests are sent not to the true destination server but to the proxy. If the requested page has been cached, the proxy will respond; otherwise it will forward the request to the true destination. Mechanisms exist to ensure that any cached copy is adequately up-to-date, but problems still arise. A variation on this theme is the so-called transparent proxy. Here Web requests are intercepted and redirected to a proxy server even though the user has not requested it and the browser has not been configured to permit it. This unwarranted interference with traffic has been justified on grounds of efficiency but is a clear breach of the Internet end-to-end principle. Problems, when they arise are likely to be mysterious and intractable.

Multicast IP

Several modern applications require data to be delivered to multiple destinations, including multiway conferencing and video distribution. A network service that does this efficiently is called a multicast service. Many of the networks on the edges of the Internet are LANs, and these have inherent support for multicast. The aim of multicast IP is to route multicast datagrams across the Internet in a way which ensures they reach all the LANs where there are destination hosts. Efficient multicast should never send more than one copy of an IP datagram between two routers. It should be the job of routers to replicate an IP datagram and forward copies down multiple paths as necessary. The mechanisms for achieving this were developed mainly by Deering (1998). From a user's point of view, the model resembles that of a radio transmission where, in order to receive, one must "tune in" to a frequency. The Internet analog for a frequency is a multicast group address. Hosts wishing to receive multicasts to a particular multicast group address have to inform the network that they wish to do so. This "user to network signaling" is handled by the Internet group management protocol (IGMP; Fenner, 1997). Multicast addresses look like ordinary IP addresses, which means there is no problem with carrying them in the source and destination fields of IP datagrams. They are "class D" addresses with the following format:

Class D   1110XXXX   XXXXXXXX   XXXXXXXX   XXXXXXXX

If a router on a LAN receives an incoming datagram with a multicast address (e.g., 224.20.5.1), it has to decide whether to multicast the datagram on the LAN. If there is at least one member of the 224.20.5.1 group on the LAN, the datagram must be multicast; otherwise, it should be dropped. For example, Figure 13 shows hosts on a LAN belonging to two multicast groups (224.20.5.1 and 224.20.5.8). Any datagram to either of these addresses must be multicast on the LAN. The router discovers group memberships on the LAN through IGMP.

Multicast Address Resolution

For transmission on a LAN to take place, the IP multicast address must be resolved to a LAN multicast address. In the case of Ethernet, for example, an algorithmic procedure is used that forms a MAC address as

<00000001000000000010111100$_2$> + < low-order
23 bits of the IP multicast address>

This is not a one-to-one mapping, and it is possible to end up with two IP multicast groups on a LAN mapped to the same MAC multicast address. This is unfortunate but not disastrous. It means that a host that has joined the group with address 224.20.5.1 will also receive datagrams intended for (e.g.) 224.148.5.1 and will have to filter these out with software.

Multicast Routing

A routing mechanism is necessary to ensure that all transmissions to a multicast address reach the correct set of routers and hence the correct set of LANs. In common with the normal Internet strategy, this mechanism has to be robust in the face of network failures. This can be achieved by minimizing the state information retained by routers and allowing this information to be continually refreshed, which requires an efficient dynamic multicast routing protocol This turns out to be a difficult problem to crack and is still the subject of much research.

Several multicast routing protocols make use of information held by routers as a result of their normal unicast (host-to-host) operation. For example, a router employing a link-state routing protocol has enough information to be able to calculate a shortest path route from any source to any destination. By calculating shortest paths from a multicast source to all known destinations, a reasonably efficient multicast tree can be constructed. All routers involved should calculate the same tree so that no problems with loops result from inconsistencies. As it stands, this technique will deliver multicast datagrams to all destinations, including those that have no group members. This is solved by allowing routers to send "prune" messages saying, "do not forward datagrams for this address to me." The multicast extensions to open shortest path first protocol (MOSPF; Moy, 1994) is of this type. The distance vector multicast routing protocol (DVMRP; Waitzman, Partridge, & Deering, 1998) uses similar ideas but is based on a distance vector unicast routing protocol. Another protocol, protocol independent multicast, (PIM; (Estrin, 1997) uses similar techniques when many group members are concentrated in an area (dense mode) but a different technique when members are widely distributed (sparse mode).

IP Security

Many applications, especially in e-commerce, require security. In particular, they require strong authentication of the participants in a transaction and, possibly, traffic confidentiality. (It is worth noting that the source address in the IP datagram header is easily faked and should not be relied on for authentication purposes). Such security can be provided in the applications themselves. Providing security in the IP layer is attractive, however, because potentiality, it can satisfy the needs of most applications.

A series of standards collectively known as IPsec specify the rules for IP security. There is an initial handshaking phase between source and destination during which a security agreement (SA) is negotiated. This agreement includes details of the precise security mechanisms and options to be used.

IP datagrams carrying IPsec traffic have an extra header immediately following the IP header. Two types of header are possible:

1. An authentication header (AH) is used when just authentication is required. This header includes a digital signature for the source of the datagram.
2. An encryption security payload (ESP) header is used when authentication and confidentiality is required. It includes parameters relevant to the encryption algorithm. The header is followed by the encrypted payload and an authentication trailer that has a function similar to that of the AH.

Implementation of IPsec is fairly limited but growing. It finds particular application in conjunction with VPNs built from IP-in-IP tunnels as described earlier.

Mobile IP

There is a difference between "wireless" and "mobile." A host on a wireless LAN may move physically around a building, but, as long as it remains on the same LAN, it has not moved as far as the Internet is concerned—it can still be reached by the same route. For the Internet, mobility occurs when a host moves to a location served by a new router. When this happens, the mobile host's IP address is wrong because its network prefix remains that of its "home" IP network. Unless this is tackled, the host will not receive IP datagrams because these will be routed to the home network. The problem is solved by means of two agent processes (see Perkins, 1996). The newly arrived host registers itself with the "foreign agent" (FA). If it is willing to accept the newcomer, the FA contacts the "home agent" (HA), which resides on the home network. From this point forward, all incoming packets are intercepted by the HA and forwarded to the FA; this uses IP-in-IP tunneling similar to that used in a VPN. The resulting route—source ⇒ HA ⇒ FA ⇒ mobile host—may be inefficient. The scheme is effective, however, because it requires no changes to be made to fixed hosts and routers.

Management in the IP Layer

ICMP

ICMP is the Internet control message protocol (Postel, 1981c). ICMP messages are used mainly to convey information about errors or conditions that require attention. There are many message types including the following:

- Echo request and response, used to probe hosts and routers to check whether they are alive and responding. The "ping" program, available on most systems, employs these messages.
- Time exceeded, used to indicate that an IP datagram has been discarded by a router because its TTL field has reached zero.
- Destination unreachable and host unreachable, used by routers to indicate discard for a variety of reasons.

The weakness of the ICMP mechanism is that its messages suffer the same problems as do other IP datagrams, thus one can never be sure they will be delivered. They can be regarded as providing useful hints about the nature of problems but cannot be relied upon.

Host Configuration (DHCP)

To operate, an IP host needs several pieces of information, including its IP address, the address mask for the local network, the IP address of the local DNS server, and the IP address of at least one router. Configuring all this information by hand on a network with thousands of hosts

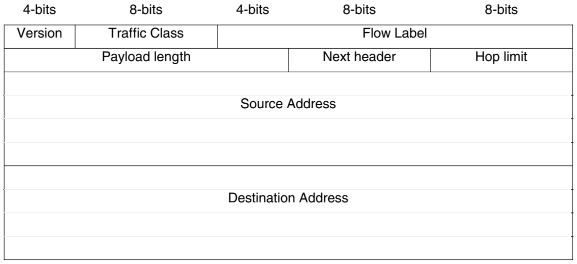


Figure 14: Internet Protocol version 6 header.

would be an administrative nightmare. The dynamic host configuration protocol (DHCP; Droms, 1997) provides a solution. When a host starts up, it broadcasts an IP datagram requesting a DHCP server. Assuming such a server is present on the local network, it will reply with the relevant information. In many cases, the IP address allocation will be dynamic. This allows a pool of IP addresses to be shared among a larger number of hosts. Only those currently connecting to the Internet need be allocated addresses, which goes some way to tackling the shortfall of addresses in IPv4.

The IP v6

A major motivation for a new version of IP was the shortage of IPv4 addresses. IPv6 (Deering & Hinden, 1998) addresses are four times as long, which should give enough addresses for the foreseeable future. (If one IPV6 address is allocated each microsecond, the addresses will last for 10^{25} years). Other major driving factors were mobility, which requires multiple address allocations as a host moves through the Internet; ease of configuration so that simple devices can be plugged and work without human intervention; security; and better support for different qualities of service.

The IPv6 header is shown in Figure 14. Despite the fourfold increase in address length, the header is just 40 bytes. This is something of an illusion, however, because some functions (e.g., security) require an additional "extension header."

The "traffic class" and "flow label" fields relate to QoS issues. Traffic class is used within the DIFFSERV scheme to indicate one of several predefined traffic classes. IPv6 routers should understand these classes and prioritize datagrams appropriately. The flow label relates to the finer grained INTSERV scheme. It is used by a host to mark a stream of related IP datagrams, which should be accorded a previously negotiated priority. These schemes are explained in the next section.

IPv6 wraps up several add-ons to IPv4 and makes these formally part of the protocol. These include ICMP, IGMP, DHCP, and IPsec. It has a much richer address structure that allows much greater scope for address aggregation; whereas IPv4 simply allows a collection of hosts to be aggregated into an IP network, IPv6 allows a hierarchy of aggregations. For example, one scheme has a "top-level" identifier followed by a "next level" identifier (probably denoting a particular site) a site identifier (like an IPv4 subnet) and an interface identifier (like an IPv4 host part).

The service provided by IPv6 is essentially unchanged, and the overall architecture remains as before; IPv6 provides simple datagram delivery, and the end-to-end layers (transport and application) build on this. To date IPv6 deployment has not been widespread. The two protocol versions can and do coexist, however, and pockets of IPv6 are beginning to emerge. It seems possible that the emergence of more powerful wireless devices such an "third-generation" mobile phones may see a ballooning of demand for IP addresses and a consequential rapid deployment of IPv6.

TRAFFIC MANAGEMENT IN THE INTERNET

For some time it has become clear that the basic "best efforts" IP service was not suitable for all applications. This is particularly so for applications with tight constraints on delay. These applications require some stronger guarantees on the maximum end-to-end delivery times for their IP datagrams (i.e. they require specialized QoS). This translates into a requirement for priority treatment by the routers through which their traffic passes. Several problems need to be addressed:

How can applications tell the routers their requirements? We need some form of "signaling" between applications and the routers.

How can routers identify which IP datagrams require which treatments?

Can we build routers that can give different priorities to different applications so that their QoS requirements are met?

INTSERV

The Integrated Services (INTSERV) initiative (Braden, 1994) was the first attempt to address these problems. It introduced the concept of a "flow"–a sequence of IP datagrams all headed to the same destination(s) and all requiring the same QoS. An application which requires special QoS for a flow engages in an explicit negotiation with the network routers it will use. The resource reservation protocol (RSVP) (Braden, Clark, & Shenkar, 1997) is employed for this signaling. The source sends an RSVP message describing the flow in terms of its peak and mean rates. The receiver then calculates the capacity that routers should reserve to meet the delay requirements. It then sends a reservation message asking all the routers along the path to reserve the required capacity. Of course the capacity might not be available, in which case an error message will be sent back to the receiver. In a CO network, this negotiation would be done only once, when the connection is set up. In the CL Internet, it must be repeated at intervals to cope with changing traffic patterns, changing routes, and so on.

During negotiation the source indicates how the flow is to be identified. This can be a combination of source and destination IP addresses, protocol identifiers and ports. This means that routers have to look at multiple header fields (and know something about transport protocols) to determine to which flow an IP datagram belongs, which will certainly slow things down. IPv6 includes a 20-bit "flow label" field into which a (hopefully) unique value can be placed. This may enable routers to classify incoming IP datagrams by examining only one field.

DIFFSERV

It is now recognized that the INTSERV approach, used alone, is appropriate only for small networks. The problem is that hundreds of thousands of flows are likely to pass through routers on core networks, and managing all these is impossible. What is needed is something much more coarse grained that allows routers to implement a few "behaviors" suited to broad classes of traffic. This is the approach used in the differentiated services (DIFFSERV; Blake et al, 1998) scheme. The "type of service" field in the IP header, now renamed the DIFFSERV field, is used to indicate the class to which an IP datagram belongs. A router has only to look at this field to know which behavior should be applied.

Effectively DIFFSERV allows the Internet to offer several classes of service to its users. The basic class offers the familiar "best efforts" service and would be relatively cheap. Traffic with tight delay constraints would use a class guaranteeing very high priority—and would pay accordingly. For this to work, ISPs must have mechanisms in place to count packets belonging to each class so that appropriate charges can be made. They must also negotiate service level agreements with neighboring networks to ensure there is an onward commitment to treat high-priority traffic appropriately. Furthermore, the guarantees offered with respect to a particular traffic class cannot be completely open ended; sources will have to agree to keep the rate at which traffic is generated within prenegotiated bounds. Traffic will have to be "policed" to ensure that these bounds are respected. Despite these complications, there now seems to be momentum behind DIFFSERV that should lead to wide deployment.

Although DIFFSERV was originally seen as an alternative and more scalable approach to INTSERV and RSVP, the two can be seen as complementary. For example, RSVP may be used to request reservation of resources for a DIFFSERV service class. RSVP also has a role in traffic engineering in conjunction with MPLS.

Performance Issues

Traffic volumes continue to increase as do the capacity of networks and links; new applications require tighter bounds on delay. This has a number of implications for the engineering of Internet components.

- Host operating systems need better real-time performance. Some popular operating systems are poor in this regard and contribute a major proportion of end-to-end delay.
- Routers, especially core routers, must handle increased traffic volumes. At the same time, they are expected to classify IP datagrams on the basis of the DIFFSERV field and process accordingly. Coping with this increased rate and volume of per-datagram processing is challenging for router manufacturers. One approach, multiprotocol label switching (MPLS; Rosen, Viswanathan, & Callon, 2001), attempts to exploit the fact that a whole stream of IP datagrams is likely to require the same processing. MPLS labels such streams through the use of an additional field. Once a stream has been labeled, routers have only to look at the label to determine the route, the priority, and so forth.
- Generally, transmission rates increase while end-to-end delays remain constant. Thus ARQ protocols such as TCP must operate in an environment in which many more segments may be sent before an acknowledgment can be expected. This can lead to inefficient operation some enhancements have been made to TCP to combat these.

CONCLUSION

The Internet is built through the implementation of a common service—the IP service—on its constituent networks. This is a minimalist service providing little more than basic connectivity. Keeping the IP service simple leaves maximum flexibility to the applications that may choose transport protocols suited to their needs. Many applications require reliable, sequenced delivery of data. These applications may employ the TCP protocol that satisfies these requirements and, additionally, assists with the management of congestion.

The IP service itself is based on the use of routers that relay traffic between networks. The routes followed are not normally predetermined and may vary with network load. IP addresses identify hosts and the IP networks to which these belong. Correct assignment of IP addresses is essential to Internet operation. Internet expansion has put pressure on the original addressing scheme. This has led to "work-arounds" such as NAT and, ultimately, to a much more flexible scheme within IPv6.

The management of congestion remains an issue and has an impact on newer, delay-sensitive applications. Initiatives such as DIFFSERV address the needs of such applications and are intended to lead to an Internet capable of carrying "multiservice" traffic having a variety of QoS requirements.

The early history of the Internet was characterized by the following philosophy: Keep the network simple, put the complexity in the hosts. More recently, commercial pressures have encouraged providers to add complexity to the network so as to offer "enhanced" services to their customers; firewalls, NAT gateways, tunnels, and proxies all reflect this strategy (and all bring awkward problems in their wake). At the same time, it is proving difficult to persuade providers and customers to support ostensibly more elegant enhancements such as IPv6, IP multicast, and IPSec. Thus we see a tension between Internet architects who wish to defend and build on the elegance of the original design and commercial interests who have customers to serve and need solutions that can be deployed straight away. One must hope that this tension will continue to be creative.

GLOSSARY

Autonomous system (AS) A collection of hosts, networks, and routers operated by a single administration. Routing within an AS is the sole responsibility of the AS administration. Routing between ASs uses a standardized Exterior Gateway Protocol such as BGP.

Border gateway protocol (BGP) A protocol used to exchange information between autonomous systems (AS). The information indicates which ASs are "reachable" from the sending AS.

Dynamic host configuration protocol (DHCP) A protocol that allows a host to obtain Internet configuration information from a server. DHCP is especially applicable when a host is first switched on.

DIFFSERV An initiative aimed at providing quality of service control on the Internet by classifying traffic into a few broad categories. Each category is then given appropriate treatment by routers.

Domain name (DN) A unique name that is used to name an object in the context of the Internet, for example, www.bluffco.com.

Domain name system (DNS) A system of servers that provides a directory service for the Internet. Each server manages part of the domain name space and communicates with other servers to look up names in "foreign" domains.

File transfer protocol (FTP) The standard protocol on the Internet for transferring files between hosts. Its role has, to some extent, been usurped by HTTP.

Gateway In the context of the Internet, a gateway is an alternative name for a router.

Host A computer attached to a network on which applications run.

Hypertext transfer protocol (HTTP) The protocol used by a Web browser to retrieve information from a Web server.

Internet control message protocol (ICMP) A protocol that provides simple management capabilities alongside the Internet Protocol. ICMP is used mainly for reporting problems and errors.

Internet group management protocol (IGMP) The protocol used by hosts to manage their membership of IP multicast groups.

INTSERV An initiative aimed at providing quality of service control on the Internet through explicit reservation of resources.

Internet protocol (IP) The protocol that governs basic data delivery on the Internet. It defines the format of the IP datagram complete with its source and destination addresses. IP will attempt to deliver the datagram to the intended destination but will not attempt to recover any errors that may occur.

IP datagram A packet formatted according to the rules of the Internet Protocol.

IPsec A protocol that can add confidentiality, integrity, and authentication to the Internet Protocol.

IPv6 Version 6 of the Internet protocol; a new standard that incorporates greatly expanded addressing and solutions to mobility, multicast, security, management, and configuration problems.

Network address translation (NAT) A scheme that allows several hosts to share a single "real" IP address.

Packet A combination of data plus control information that is treated as a unit by a data communication network.

Port A number that identifies a particular application or service available on a host.

Router A device that relays IP datagrams between networks. It determines where to send the datagram by referring its destination address to a routing table.

Resource reservation protocol (RSVP) A signaling protocol that allows hosts to reserve Internet resources to obtain a desired level of quality of service.

Transmission control protocol (TCP) A protocol that provides reliable, sequenced, flow-controlled data exchange between hosts. Common applications such as file transfer and the Web use TCP to provide the reliability they need.

User datagram protocol (UDP) A simple protocol that allows an IP datagram to be sent to a particular port on a particular host. UDP does not enhance the IP service in any way.

CROSS REFERENCES

See *Internet Security Standards; Standards and Protocols in Data Communications; TCP/IP Suite.*

REFERENCES

Note: The Requests for Comment (RFCs) cited in this chapter are from the Internet Engineering Task Force and are available online on its Web site, http://www.ietf.org/rfc.html

Atkinson, R., & Kent, S. (1998). Security architecture for the Internet Protocol. *Internet RFC 2401.*

Berners-Lee, T., Fielding, R., & Masinter, L. (1998). Uniform resource identifiers (URI): Generic syntax. *Internet RFC 2396.*

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998). An architecture for differentiated service. *Internet RFC 2475.*

Braden, R., Clark, D., & Shenker, S. (1994). Integrated services in the Internet architecture: An overview. *Internet RFC 1633.*

Braden, R., Zhang, L., Berson, S., Herzog, S., Jamin, S. (1997). Resource reservation protocol (RSVP)–Version 1 functional specification. *Internet RFC 2205.*

Deering, S. E. (1998). Multicast routing in internetworks and extended LANs. *Proceedings of SIGCOMM ACM Special Interest Group on Data Communication '88,* 55–64. New York: ACM Press.

Deering, S. E., & Hinden, R. (1998). Internet protocol, Version 6 (IPv6) specification. *Internet RFC 2460.*

Dierks, T., & Allen, C. (1999). The TLS protocol, version 1.0. *Internet RFC 2246.*

Dijkstra, E. W. (1959). A note on two problems in connection with graphs. *Numerische Mathematik, 1,* 269–271.

Droms, R. (1997). Dynamic host configuration protocol. *Internet RFC 2131.*

Egevang, K., & Francis, P. (1994). The IP network address translator (NAT). *Internet RFC 1631.*

Estrin, D., Farinacci, D., Helmy, A., Thaler, D., Deering, S., Handley, M., Jacobson, V., Liu, C., Sharma, P., & Wei, L. (1997). Protocol independent multicast-sparse mode (PIM-SM): Protocol specification. *Internet RFC 2117.*

Fenner, W. (1997). Internet group management protocol, version 2. *Internet RFC 2236*.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., & Berners-Lee, T. (1999). Hypertext transfer protocol—HTTP/1.1. *Internet RFC 2616*.

Fuller, V., Li, T., Yu, J., & Varadhan, K. (1993). Classless inter-domain routing (CIDR): An address assignment and aggregation strategy. *Internet RFC 1519*.

Jacobsen, V. (1988). Congestion avoidance and control. *Proceedings of Association of Computing Machinery SIGCOM ACM Special Interest Group on Data Communication '88* (pp. 314–329). New York: ACM Press.

Klensin, J. (Ed). (2001). Simple mail transfer protocol. *Internet RFC 2821*.

Malkin, G. (1998). RIP version 2. *Internet RFC 2453*.

Metcalfe, R. M., & Boggs, D. R. (1976). Ethernet: Distributed packet switching for local computer networks. *Communications of the ACM, 19,* 395–404.

Mockapetris, P. (1987a). Domain names—concepts and facilities. *Internet RFC 1034*.

Mockapetris, P. (1987b). Domain names—Implementation and specification. *Internet RFC 1035*.

Moy, J. (1994). Multicast extensions to OSPF. *Internet RFC 1584*.

Moy, J. (1998). OSPF version 2. *Internet RFC 2328*.

Perkins, C. (Ed.). (1996). IP mobility support. *Internet RFC 2002*.

Plummer, D. C. (1982). An Ethernet address resolution protocol. *Internet RFC 826*.

Postel, J. (Ed.). (1980). User datagram protocol. *Internet RFC 768*.

Postel, J. (Ed.). (1981a). Internet protocol—DARPA Internet program protocol specification. *Internet RFC 791*.

Postel, J. (Ed.). (1981b). Transmission control protocol—DARPA Internet program protocol specification. *Internet RFC 793*.

Postel, J. (1981c). Internet control message protocol—DARPA Internet program protocol specification. *Internet RFC 792*.

Postel, J., & Reynolds, J.K. (1985). File transfer protocol (FTP). *Internet RFC 959*.

Rekhter, Y., & Li, T. (1995). A border gateway protocol 4 (BGP-4). *Internet RFC 1771*.

Rosen, E., Viswanathan, A., & Callon R. (2001). Multiprotocol label switching architecture. *Internet RFC 3031*.

Saltzer, J. H., Reed, D. P., & Clark, D. D. (1984). End-to-end arguments in system design. *ACM Transactions on Computer Systems, 2,* 277–288.

Schulzrinne, H., Casner S., Frederick R., & Jacobson V. (1996). RTP: A transport protocol for real-time applications, *Internet RFC 1889*.

Simpson, W. (Ed). (1994). The point-to-point protocol (PPP). *Internet RFC 1661*.

Vetter, R. J. (1995). ATM concepts, architectures, and protocols. *Communications of the ACM, 38,* 30–38

Waitzman, D., Partridge, C., & Deering, S. (1998). Distance vector multicast routing protocol. *Internet RFC 1075*.

Internet Censorship

Julie Hersberger, *University of North Carolina at Greensboro*

INTRODUCTION

Internet censorship has become a major social issue in various countries as Internet access becomes more available to large numbers of users globally. Currently, there are more questions than answers in regards to essential Internet censorship issues. Censorship is dependent on the following critical points: Who controls Internet access, both legally and ethically, in not only the United States but also in the international arena determines Internet governance? Who then decides what content should or should not be accessible constitutes content control? Who owns Internet content and how ownership is determined are the realm of intellectual property? As Internet connectivity increases worldwide, legal and ethical discussions and arguments become increasingly complex. Others might perceive what some consider to be an ethical or legal denial of access to the Internet as censorship. Censorship is technically a legal action although the term is not always used as such in everyday usage. Even the attempt to define the term "censorship" in order to begin useful discussion proves difficult. In addition to the issues of governance, content discussions, and intellectual property questions, new advances in technology also affect Internet access and raise censorship concerns.

The definition of key terms involved in such a discussion need to be established first in order form an analytical framework to be established. Definitions of key terms are important but dilemmas arise as to whose definitions prevail. The United States is one of the countries at the forefront of Internet use and has thus been at the forefront of the Internet access controversy. Historically, there has been a long ongoing discussion of information access, censorship, obscenity, and pornography based on First Amendment rights of free speech and assembly in the United States. Freedom of speech, however, is not a legal absolute and does not always protect certain areas such as libel (defamation), incitement (including where hate speech overreaches and advocates violent acts), obscene materials and child pornography, or Internet spam. Issues of security, privacy (especially as this applies to the endangerment of children), and intellectual property are also Internet categories where censorship comes into play.

As the novelty of the Internet diminishes and as more and more users become more Internet savvy, the controversy over access and censorship of the Internet becomes more manifest. The Internet allows for the full range of speech media as compared to single-transmission media such as books, television, or movies. Confusion exists in determining legal governance of the Internet and in subsequent development and application of the definition of the term "censorship." In the strictest legal interpretation of a definition of censorship based on First Amendment rights, only the U.S. government can technically censor speech, represented as text or symbolic speech, which includes visual and aural media.

In lay terms and common usage, however, censorship is simply defined as an attempt by anyone to limit access to text or visual media, or the term is applied to a person or a group who attempts to act as a censor. This bifurcation of legal and lay perspectives becomes even more confounding due to the confusion over the legal standards for what constitutes obscene, indecent, or inappropriate material. Obscenity is a court-determined standard (*Miller v. California,* 1973) and establishes that a work is judged obscene if it fails a three-prong test. The three prongs consist of (a) whether the average person applying contemporary community standards would find the work, taken as a whole, appeals to prurient interests; (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work taken as a whole lacks serious literary, artistic, political, or scientific value. "Obscene" is a designation that only applies to materials in terms of sexual content. Individuals often refer to hate speech or violent images as being obscene but they are not legally so. Indecent material includes the category of pornography and continues to be a very problematic and very subjective area, as is true also of the domain of inappropriate material. Indecent and inappropriate materials are still legally protected for dissemination to adults; however, courts have decreed that it is illegal to distribute pornography to minors under the age of 17 (*Ginsberg v. New York,* 1968). Child pornography is not legally accessible in the United States in any format. Persons accessing child pornography are open to legal prosecution regardless of the format accessed. Distinguishing what constitutes pornographic material from indecent or inappropriate works becomes very difficult due to the lack of useful definitions.

Such designations are of critical importance in terms of promulgating laws, policies, and procedures. The main ongoing discussion is how such laws are transferable to the Internet. Due to the emotional nature of the discussion of Internet censorship, confusion over governance issues, and a lack of understanding of the legal basis for what constitutes censorship, access to the Internet promises to remain a prominent legal and social concern for years to come. What is legal in the traditional sense of censorship and access may not be easily transposed to cyberspace. As these issues are examined, important decisions will define, refine, and clarify who will have access to what on the Internet locally, nationally, and internationally. Censorship needs to be considered on a legal level, but there are ethical implications. Text and images are legally protected under First Amendment measures or international equivalents, but this does not imply that all such material ought to be available. For an excellent analysis of ethics and Internet access see Spinello's *Cyberethics* (2003).

GOVERNANCE ISSUES

Internet governance is, in many ways, an oxymoron as the original purpose of connecting networks was to facilitate the transfer of information easily and without oversight. Censoring information exchange was not a concern during the early developmental phase of the current Internet architecture. The historical evolution of the Internet, beginning with the U.S. Department of Defense's ARPANET by the Advanced Research Projects Agency in the late 1960s, intentionally established a network whose main function was to facilitate rapid digital communication in times of national emergencies. An analogy would be similar to that of the interstate highway network that was established to facilitate rapid transportation efforts during national emergencies. Few people were part of this network composed mainly of computer scientists within the U.S. government and a few universities. Later, in the 1980s, the National Science Foundation network (NSFNET) was developed and ARPANET's role became greatly diminished. In addition to government networks, NSFNET included several university networks and thus the term Internet came to be. The primary purpose of the early Internet was, again, information sharing, so the network was developed to easily transmit all types of textual and graphic data. This attention to detail in promoting ease of transmission was to become a major governance issue later. Adding to the governance confusion is that even though the U.S. government supported the early Internet efforts, it was developed for government and academic use, not commercial use. In the 1990s, as the Internet became much more of an open network, with commercial enterprises flourishing and users outside of the government and academia gaining access, growth increased at a rapid rate. The Internet architecture, which was developed to facilitate sharing, made it extremely difficult to impose restrictions on information sharing. Commercial needs, though, often require restrictions due to intellectual property issues. As a result the governance of the Internet, and concomitantly who can legally censor access or content material is complex and confusing. Even if some form of national or international consensus could be reached regarding Internet governance it is unclear what actions could be taken to limit information exchange. Internet architecture still supports ease of file sharing, leading to difficulty with the ability for music exchange. Music piracy versus the legal exchange of music files has been addressed recently in courts. Free downloads of music files in some instances may boost music sales. Downloading entire music CDs rather than paying for the official product amounts to piracy. Copyright holders may legally limit access to music files, which does not constitute censorship in the traditional sense of intellectual property rights. Music fans view the law as an abridgment to their right of access to information and from their perspective, such limitation amounts to censorship. Case law will be used to clarify and refine legal governance issues.

Regulatory agencies attempting to govern Internet access and content may begin to investigate control at the architecture and infrastructure. New systems may utilize changes in code that would shift control from users to developers and providers. Such a change would not be popular with long-term Internet users who are used to the anarchic atmosphere of trading information at will. One of the most attractive qualities of the Internet is that decentralization aids in the speed and functionality of information transfer. Future governance issues concerning new architecture and infrastructure approaches will be decided in national and international courts. The evolution of e-commerce adds another layer of governance issues, especially in the global market. Censoring of commercial information again lands in the legal sector, but whose laws ought to restrict global commerce remains a critical question.

CENSORSHIP AND INTERNET CONTENT

Internet censorship affects individuals, groups, nations, etc. Two major issues must be addressed concerning governance of the Internet and censorship. Who controls access to the Internet is as critical a question as denying a right of access to the Internet. The second issue concerns Internet content, or what is actually available on the Internet. Both issues comprise legal and ethical elements. A teenager could level a charge of censorship against a parent who installs filtering or tracking software, but this cannot be considered legal censorship since the parents' action could be interpreted as ethical parenting. A further example of possible censorship with regards to denying a right of access is where public libraries or school libraries install filtering software to block predetermined categories of Web sites from use. Whether this is governmental censorship is still being addressed by the U.S. Supreme Court. Schools may also be able to argue that they can legally limit what content is actually published on their Web site. *Hazelwood v. Kuhlmeier* (1988) could be cited as a precedent in limiting student speech on the Internet. In *Hazelwood,* the courts ruled that schools could legally control the content of school newspapers and such a decision could be argued to transfer to the Internet. Whether many may interpret the outcome of the case as legal, or ethical, the decision remains very open to further discussion in terms of censoring access to and the content of the Internet.

Internet service providers (ISPs) also have the ability to impose restrictions on Web content without legally abridging First Amendment rights. Free markets allow for consumers to choose from a variety of Internet providers and any restrictions or limitations of accessing content and posting content should be clearly stated in any agreements arranged between consumer and provider. An example of this is the Prodigy case in the early 1990s (Stratton Oakmont, Inc. v. Prodigy Services Co, 1995) where an Internet service provider imposed restrictions on what could and could not be posted to its electronic bulletin board service. Prodigy claimed that they were not a common carrier and thus not bound to transmit all messages. Furthermore, they claimed other electronic outlets that allowed for free expression existed so that the content that they were prohibiting did not amount to a restriction of access to information. The inclusion of First Amendment rights, FCC regulations, and antitrust laws all within the Prodigy case show the impact of Internet censorship in the commercial arena but show that there are differing limitations from those of the government limiting free speech. In the global legal arena, efforts to determine Internet regulation regarding censorship are still in the process of governance clarification and discussion. Cultural norms vary greatly worldwide, and what is legally accessible in some countries may be illegal in others. Whose rights will then prevail in international legal cases is still being determined in courts worldwide.

LEGAL ISSUES AND INTERNET CONTENT

Laws, policies, rules, and regulations have all been enacted in order to limit both access to the Internet and content posted. Historically, as laws and rules are enacted, individuals and groups have a right to legally challenge such actions where they are perceived to abridge access rights. A robust body of case law exists in the United States relative to censorship topics including sexual content, pornography, obscenity, hate speech, libel, and unwanted speech. At issue currently is whether these historical rules and refinements that have been established to apply to past forms of communication media apply to the Internet. Current legal cases in the United States provide insight into the judicial, legislative, and presidential perspectives regarding Internet content control with such efforts aimed at the protection of children from inappropriate Internet content. The Communications Decency Act of 1996 (CDA) and the Child Pornography Prevention Act of 1996 (CPPA) were both congressional statues, signed into law by then President Bill Clinton. Even though the statutes embodied congressional and presidential support, legal challenges were filed to challenge the constitutionality of the measures. Subsequent congressional attempts to limit the access of minors to Internet materials include the Children's Online Protection Act (COPA) and the Child Internet Protection Act (CIPA). As these have all been government actions to restrict access to the Internet, the classic definition of censorship would apply. Examining the court challenges to the enactment and enforcement of the CDA, CPPA, COPA, and CIPA can identify critical issues.

Internationally, a growing body of statutes and challenge cases is beginning to clarify and refine Internet censorship issues in other countries. These new cases are beginning to challenge jurisdictional claims regarding content control on an international basis.

United States

Legally, the body of case law affecting Internet access and censorship has been growing since the mid-1980s. No attempt is made here to produce a comprehensive review of case law, but a review of selected important cases that affect access to content on the Internet provides insight into the critical issues concerning Internet censorship. Most cases address obscene or pornographic materials and their access by children. Other cases address interstate commerce and jurisdiction issues. The first Internet obscenity case (*U.S. v. Thomas,* 1994) also addressed the issue of geographical jurisdiction of cyberspace. In this case, a Tennessee court found a California couple guilty of transporting obscene materials across state lines via their computer over interstate phone lines to a Memphis postal inspector. The couple had established a private, adults-only electronic bulletin board system (BBS) where subscribers could order pornographic materials online in addition to chatting and exchanging files. The Miller test's (*Miller v. California,* 1973) reliance on community standards to dictate what is obscene is particularly difficult to adjudicate in a global medium. The Thomases requested a venue change from Tennessee to California where presumably community obscenity standards for indecent materials would be more liberal than those of Tennessee. Lawyers argued that the new technology required a rethinking of the traditional geographic location of cases. In obscenity cases, trials may be held in the district from where the materials were transmitted or in the district in which they were received. Another jurisdictional issue for Internet obscenity cases may also add where the material was published as another possible venue for prosecution of cases if it differs from the location where materials were transmitted or received. The court did not embrace the argument advanced by the Thomases, that they had no control over where their Internet subscribers are located geographically and that a chilling effect would occur if BBS owners had to screen who would receive their materials. According to the court decision, members had to submit applications with home addresses and phone numbers so the couple had the necessary information to screen accepting applications from users in states with stricter community standards than California. In effect, the court ruling requires BBS operators to become experts not only on community standards in the 50 U.S. states, but also become experts on international law. The critical issue of community standards in a global medium is also addressed in the Supreme Court's consideration of CDA and COPA.

The Communications Decency Act was overturned by the Supreme Court in *Reno v. ACLU* (1997). The CDA prohibited the transmission of obscene or indecent materials to minors using community standards. In its attempt to protect minors, the court found that the CDA overly restricted adult access to Internet information. Three main considerations led a court majority to overturn the CDA. First, no existing technology existed to effectively bar minors from gaining access and thus the act was considered to be overbroad by the court. Secondly, the act was deemed also overbroad for failing to define "indecent" or "patently offensive," and thirdly, the affirmative defenses offered were not narrowly tailored enough (*Reno v. ACLU,* 1997). Overall, the court found that the extension of the CDA's breadth was unprecedented and struck down the act.

In its next iteration, the CDA morphed into CDA II known also as the Children's Online Protection Act. Once the act was signed into law, the American Civil Liberties Union filed suit. In *Ashcroft v. ACLU* (2002) the Supreme Court heard the argument that COPA too was overbroad and overly dependent on community standards. COPA differed from the CDA in three important ways. First, where the CDA addressed the Internet as a whole, COPA focused on the World Wide Web (WWW) only. Second, COPA only addressed commercial speech, not all speech. Thirdly, whereas the CDA attacked "indecent" and "patently offensive" speech, COPA only restricted speech that was harmful to minors. The majority of the justices supported sending the case back to the Third Circuit court to address three key issues. Was COPA overbroad for reasons other than its use of community standards? Was the statute constitutionally vague? Finally, would the statute survive strict scrutiny analysis (*Ashcroft v. ACLU,* 2002)? Unlike the decision in *U.S. v. Thomas* (1995) where consumers applied for membership to a BBS by sending their home information, COPA regulates World Wide Web access and no technology currently exists to establish users' geographic addresses. In effect, a community that wishes to exist without access to certain materials not only rids itself of that material, but all Internet users would be prohibited from accessing that information. The court reenforced that implementation of COPA regulations is still enjoined until the lower court re-addresses these critical areas. Whether community standards are nationally or geographically bound will be clarified in future discussions of this and other legal challenges. Which community standards are used to determine access to or content on the Internet will continue to be debated and refined over time.

Another ruling by the U.S. Supreme Court (*Ashcroft v. Free Speech Coalition,* 2002) addressed the Child Pornography Prevention Act of 1996 and its prohibition of graphic depictions of explicit sex involving real or computer simulations of minors. The justices all agreed that pornography depicting real children in sexually explicit situations remains obscene and beyond First Amendment protections. The issue directly addressed simulations of child pornography using computer-generated images. In *New York v. Ferber* (1982) the court determined that the process of producing child pornography, in addition to the distribution and sale of such material, was harmful to actual minors and should be prohibited despite whether the resulting images could be adjudicated to be legally obscene. Justice Kennedy's opinion points out that the CPPA reaches beyond the standards for obscenity as established in either *Miller* or *Ferber.* The CPPA prohibited all formats of images including photographs, film, video, picture, and computer or computer-generated image or picture. The Court majority agreed that such a prohibition would apply to works of classical art, to films of *Romeo and Juliet,* and to the recent Academy Award-winning film, *American Beauty,* rendering these works vulnerable to censorship under the Act. The court majority also agreed that the Act was overbroad in attempting to prohibit depictions of minors in explicitly sexual situations using adult actors. Justice O' Connor agreed with the majority on this point, but disagreed that prohibiting computer-generated simulations was overbroad. The severe penalties levied CPPA offenders were also a concern of the court majority. A first-time offender could be sentenced to a maximum of 15 years in prison with a repeat offender receiving not less than 5 years or more than 30 years maximum sentencing. The court majority agreed that such a severe penalty threat would have a chilling affect on legitimate artistic efforts. Finally, the court majority did not agree with the CPPA that the fact that virtual child pornography might be used to seduce minor children and thus harmful to minors was reasonable. Nor did the argument that virtual pornography might increase a pedophile's appetite for real child pornography and encourage illegal acts persuade the justices to limit access to virtual child pornography. Other sections of the CPPA are still being discussed in the legal arena, and it is likely that subsequent statues will continue to attempt to legally block access to questionable Internet content concerning child pornography.

In addition to COPA and the CPPA, Congress also enacted the Children's Internet Protection Act of 2000. In the CIPA, Congress tied the receipt of e-rate (education rate) funding to the installation of blocking or filtering software in public schools and public libraries. The e-rate funding was established by Congress (Universal Service Fund for Schools and Library Program) as telecommunications discounts provided to libraries in low-income communities in order to support Internet access for educational purposes. Libraries also were required to enact a policy that addressed inappropriate Internet access, with what constitutes "inappropriate" materials being determined by a local school board or library. The policy also needed to prohibit unauthorized access by minors, protect personal information of minors, and to restrict minors access to materials harmful to adults. Libraries were required to hold a public hearing on the proposed Internet safety policy. The American Library Association (ALA) filed suit in the U.S. District Court, Eastern District of Pennsylvania. The suit is not contesting the provision that public schools were required to install blocking or filtering software on all public access computers; however, the suit contends that blocking or filtering access to public library Internet terminals is a First Amendment violation. The statute authors claim that tying the receipt of federal funding to acquiring and monitoring use of blocking or filtering software is not a First Amendment issue. The ALA suit notes several weaknesses to the statute, pointing out that public libraries were established to provide educational and recreational materials to a wide range of users. Libraries are not legally required to provide Internet access, but if they choose to do so, they are governed under the laws of traditional public fora where First Amendment rights apply. The ALA suit claims that the government witnesses overstated the efficacy of blocking and filtering software used in those public libraries. Further noted was that the government witnesses admitted that their blocking and filtering software standards applied equally to both adults and minors, thus reducing access for adults to the level of what is appropriate for minors only. A critical point is that the suit does not claim an absolute right to access speech, but that right exists for adult library patrons not to have to petition the government for the right to access disfavored speech to which they are legally entitled (*ALA v. U.S.,* 2002). The suit contends that the statute is overly dependent on the proposition that e-rate funding is restricted only to educational pursuits, a proposition that has never been legally established.

The Third District Circuit Court of Appeals issued a unanimous finding that the law did abridge the First Amendment concerns of prior restraint, vagueness, and overbreadth. In the effort to protect children from accessing indecent or inappropriate materials, copious amounts of legally protected speech would be prohibited for access by adults. The judges mainly based their opinion on the premise that Internet filtering is faulty, blocking text but not images, failing to prohibit much that is illegal while blocking a substantial amount of speech that is legally protected. The case is being appealed to the U.S. Supreme Court as of the writing of this chapter.

A pattern has emerged where Congress enacts legislation that attempts to limit adult access to the Internet in order to protect children and then consequently various free speech groups challenge these statues in court. So far, the Supreme Court has been reluctant to limit access to legally protected speech for adults at that level which does not cause harm to minors. It is expected, however, that this pattern will continue as the legislative and judicial branches of government attempt to establish any legal limits to Internet access.

International

Other countries are also addressing issues of Internet censorship in the legal arena. However, gaining access to government actions that constitute censorship internationally can be difficult. Many countries restrict access to government records and the media does not always publicize such government actions. Thus it can be difficult to gauge the extent of Internet censorship occurring internationally. A few cases are beginning to be made public and some countries are making public their legislative actions. Australia's government has introduced legislation making it a criminal act for users and ISPs to make materials unsuitable to minors available, even if these materials are limited to use only by adults. Singapore, another country whose government limits access to Internet speech, has done so by establishing posting laws regarding political campaign posting regulations. The European Union Council (EU) and the European Parliament have also addressed the issue of the production of, processing of, possession of, and distribution of child pornography. One final example is that of China where the government shut down cybercafés that did not block or filter access to pornographic sites. These are but a few examples of actions to limit access to the Internet in just a few countries worldwide.

A current legal case of interest concerns a Web developer's posting of information prohibited in another country and their being sued to block that information in order to conform to the laws of that other country. In the case of UEJF & LICRA v. Yahoo (2000) a French court found that Yahoo!, Inc. based in California had violated French penal code by offering Nazi and Third Reich memorabilia for sale on their auction site. Two nonprofit organizations sued to force Yahoo to block transmission to France. The International League against Racism (Licra) and the Union of French Jewish Students (UEJF) initiated the suit. The Paris court found that Yahoo had violated the plaintiffs' rights under French law. If Yahoo did not cease disseminating the auction information to France a fine of 100,000 francs per day (U.S.$13,000) would be levied. The fact that Yahoo's policy prohibits the offering of items connected to hate speech played an important role in the court's decision. Yahoo then counter-sued in a California district court in claiming that French court orders were neither recognizable nor enforceable in the United States. The California court upheld Yahoo's claim. In a very recent decision, the Ninth Circuit of Appeals in San Francisco overturned the lower court ruling (*Yahoo! v. Licra,* 2002). The judges ruled that Yahoo could not sue persons who had no presence in that country nor have committed a crime in that country. The court also noted that the defendants ought to be able to exercise their rights in their home forum. The Appellate Court stated that the lower court had been "caught up in the novelty of the case, the French disregard for the First Amendment, and Yahoo's need for an American forum to protect its rights." The court also strongly stated that U.S. courts ought not issue advisory opinions regarding foreign court decisions when no attempt has been made to enforce the orders or judgments in the United States. Finally, the court stated that U.S. district courts ought not provide the means for "forum-shopping plaintiffs" to circumvent legitimate foreign court decisions in order to create "dueling courts" as an exercise in public relations or for other purposes (*Yahoo! v. Licra,* 2002, pp. 46–47). Some legal scholars are worried that this decision could limit access internationally to that speech legally acceptable in the most restrictive of countries worldwide. Others argue that other countries have the right to determine what is lawful within their sovereign borders and that such legal cases are needed to sort out jurisdictional issues.

Efforts to censor access to information on the Internet will continue to be proposed in the government arenas both nationally and internationally. Hopefully, opposing groups will be able to question such legislation in the national and international courts to discuss important Internet access issues regarding Internet content.

CENSORSHIP AND INTELLECTUAL PROPERTY ON THE INTERNET

Ownership of Internet information is another area where charges of censorship versus legitimate access and content limitations must be considered. As is true with content-based censorship questions, there is a long history of intellectual property (IP) legal challenges in both the United States and in other countries worldwide regarding traditional communication media such as print, radio, and television. Whether these traditional laws and decisions are transferable to Internet IP issues continues to be challenged, tested, reviewed, and refined. Interpretations of the legal status of property and the subcategory of intellectual property vary widely in the global market. Intellectual property laws attempt to balance the rights of product creators, owners, and users through copyright, patent and trademark statutes, and case law decisions.

Alleged copyright and trademark violations are the two areas where Internet censorship accusations are most frequently leveled. The Copyright Act of 1976 was amended in 1980 to include the protection of software development at a level similar to that of traditional protected works. The ease of accessing and sharing Internet information raises many copyright infringement accusations. The "fair use" section of the Copyright Act has been interpreted and reinterpreted in courts in the United States. Tensions arise when claims of fair use come into conflict with claims of ownership and denial of right of access. Usually, the decision is not that users do not have the right to access information, but that they do not have the right to copy and reuse the material traditionally granted under the doctrines of fair use and the right of first sale.

The Digital Millennium Copyright Act (DMCA) was enacted by Congress in 1998. The DMCA bans the use of technological devices that provide the means to circumvent protection systems and other access measures in an attempt to limit file sharing. The DMCA is aimed at protecting the rights of copyright creators and holders while striving to make certain file sharing illegal when software is developed to circumvent antipiracy encryptions coded into software for commercial sale or to develop software that enables the illegal copying of software. Critics of the DMCA argue that the act overly favors creators and rights holders and results in censorship when it abridges individual rights of traditional copyright law such as first-sale rights, fair use, and thus the right of free speech.

A case of interest is that of Jan Johansen, who is being criminally prosecuted in his home country of Norway for developing DeCSS software that allows DVDs to potentially be copied or played on unlicensed equipment. Johansen is only a teenager but managed to develop code that circumvents DVD protection systems. Johansen then shared the code via his Web site for international access. The Motion Picture Association of America (MPAA) filed suit against the use of DeCSS under the terms of the DMCA that supports the rights of copyright holders of creative works to limit access. Encyrption codes are not protected as First Amendment speech according to the U.S. courts, which upheld the ban of sharing DeCSS on the Internet (Electronic Frontier Foundation, 2002; Spinello, 2003). The courts determined that the code that instructed the computer to decrypt the software protection system was not considered speech under First Amendment protections (*MGM v. Grokster,* 2002) Other court cases are addressing other instances where protection systems have been circumvented under the rubrics of right of first sale and fair use. The decisions of these cases will add to the body of case law clarifying what constitutes legal access while interested parties attempt to limit censorship through limiting or prohibiting copying and sharing of materials on the Internet.

A substantial section of the DMCA addresses the implementation of treaties established by the World Intellectual Property Organization (WIPO). These treaties address the issues of copyright protection systems and copyright management information and the evaluation of the impact of copyright law and amendments on electronic commerce and technological development. The WIPO focuses on two major intellectual property fields. The first pertains to industrial property questions concerning inventions (patents), trademarks, industrial designs, and geographic indications of source. The second field concerns copyright and concomitant rights related to copyright and the collective management of copyright. The WIPO has been in existence in some form since the late 1800s. Currently, membership countries number 179, which is over 90% of the countries worldwide. One of the main purposes of WIPO is to broker international treaties that affect creators and owners while encouraging dissemination and use of IP. Recent treaties, such as the 1996 WIPO Copyright Treaty and the 2000 Patent Law Treaty, affect access to Internet materials. The WIPO Arbitration and Mediation Center maintains a list of professionals worldwide who are trained to deal with domain disputes. In 2000, the Center handled 1,850 cases (WIPO, 2002). Domain name arguments traditionally take the form of "twins," where multiple parties lay claim to the same name. Anyone with the name "McDonald" could run into problems when trying to claim a domain name where the major U.S. corporation of the same name might dispute ownership. "Parasites" are those who register a domain name that sounds very similar to a famous company name with the intent to enhance sales of their own products. "Cybersquatting" is when individuals or groups register the same or similar domain names of famous entities with the intent resell them back to the well-known people or companies (Spinello, 2003).

Software development falls under both the patent and copyright areas. Ideas cannot be patented or copyrighted, but representations of ideas can be. Patents often are classified under the "first to file" rule and ownership may not be registered with the originator of the product. As a result, who controls access to certain software code becomes controversial. Some software developers and owners believe in open source code sharing that others do not.

Universal consensus does not exist as to whether a right to ownership of information exists on the Internet in the same way it does in traditional formats. The WIPO is affiliated with the World Trade Organization (WTO) and there are those worldwide that disagree with the efforts of both organizations. The Against Intellectual Property Homepage (2002) claims that information as property and information ownership are not beneficial to society as a whole. Whether copyright is even ethical is addressed by Warwick (2001), who argues that traditional U.S. copyright laws have valued the benefit of created works to society as a right over the value and rights of creators. The Electronic Frontier Foundation (EFF, 2002) Web site frequently addresses the issue of IP and information access and content control.

Deep-linking is the practice of bypassing a Web site's homepage and linking to a lower level page within the site (*Deep Linking,* 2002). Legal cases in Germany, Denmark, and the United States illustrate how this issue can be construed as censorship when governments become involved in determining copyright infringements. In the case in Germany, copyright law awards exclusive dissemination and reproduction rights to the maker of an online database. A German publishing group, Verlasgruppe Holtzbrinck, sued NewClub.de that produces a search engine for news stories by aggregating headlines, claiming that the search results and linkages directly to articles violated and abridged their property rights. According to the publishing group, their news headlines constitute a database and thus these links violate their right to control dissemination and reproduction. The German courts are working to clarify "temporary copying acts" and allowing search engines to link to subordinate Web pages. A similar case in Denmark also involved search engines that link to newspaper articles directly, bypassing the main homepages of the news providers. The Danish Newspaper Publishers Association claims that bypassing their homepage diminishes revenues to advertisers when readers are diverted to lower level pages. In an earlier case in the United States, Tickets.com won the right to link to a subordinate page on Ticketmaster's site without linking to the main homepage. The court's decision compared the unique Internet linking system more to a library card catalog than a violation of copyright law. As long as the search engine results make it clear who is responsible for posting the Web content, no copying violation has occurred. Other legal cases will further clarify the relationship between search engines, advertising, linking, and copying both nationally and internationally.

National and international legal cases, in addition to national and international discourse, aided greatly by Internet sharing, will further the clarification of competing rights of information access between users and creators. Whether the prohibition of access to Internet information meets the level of censorship claims abridging fair use or the right of first sale, or whether intellectual property rights being upheld will continue to be debated in courts of law and on the Internet itself remains to be seen.

CENSORSHIP, TECHNOLOGY, AND INTERNET PRIVACY

Internet privacy and security issues are peripheral censorship issues but ought to be included in discussions on rights of information access and content control. Conflict arises between advocates of information being made available in its freest forms versus claims of right to privacy or security concerns. In the United States the Constitution does not designate a particular right to privacy, but over time case law and statutes have carved out more and more privacy rights pertaining to personal information of individuals. The Fourth Amendment protects the rights of individuals to be secure in their homes, papers, and effects against unreasonable search and seizure without probable cause issued by a court of law. Computer records would presumable fall under the category of "papers and effects" as being constitutionally protected. However, with the passage of the Patriot Act in 2001, what merits governmental right of access to personal information has broadened greatly, giving sweeping powers to international intelligence agencies in addition to domestic agencies. Expanded government surveillance with a very low level of evidence to substantiate probable cause has been established. Software, called "spyware" has been developed for the sole purpose of surveillance and logging Internet activities and allows not only the government but other individuals to monitor computer use. Internet service providers may voluntarily furnish user information, not including content information, without a court order or subpoena. The Patriot Act also allows the government to investigate computer trespassing without requiring a court order. Such expanded government surveillance of computer use might result in a chilling effect on Internet users who might self-censor what information they seek and which Web sites they view.

In a case that combines several aspects of Internet censorship, the Nuremberg Files case involves Internet content, privacy, hate speech, and access. The Web sites are the work of Neal Horsely, who maintains the sites (Nuremberg Files, Christian Gallery, 2002) through Pathway Communications. The sites are provided in order to collect information on those associated with abortion clinics for future court actions, according to the Web sites. Past sites had posted "wanted posters" containing personal information including names, home addresses, home phone numbers, and photographs mainly of physicians who performed abortions and clinic staff and owners. As persons on this list were killed or wounded, an X would be drawn through their photo. Also a line would be drawn through the deceased person's name. Those wounded were noted in gray type. In 1999 a Planned Parenthood organization along with an abortion clinic and four doctors filed suit in Oregon under the government's Freedom of Access to Clinics Entrances Act, which prohibits threats of violence to deny access to clinics. An Oregon federal jury found for the plaintiffs and awarded $109 million dollars in damages to Planned Parenthood, the clinic, and the doctors (*Planned Parenthood v. American Coalition of Life Activists,* 2002). A three-judge appellate court overturned the ruling in 2001, stating that the Web site speech was protected by the First Amendment. The court further ruled that even though the speech could be considered highly offensive, that the language could not be construed as representing a true threat as the site was conceptualized to provide public advocacy and was not seen as a call to action. In 2002 the Ninth U.S. Circuit Court of Appeals overruled the previous court decision, determining that "violence is not a protected value" (*American Coalition of Life Activists v. Planned Parenthood,* 2002). Court justices in the majority stated that the site demonstrated intent to inflict bodily harm existed and that there was a pattern of violence in the way the content was portrayed. The court did not uphold the monetary award of $109 million and the lower court is currently reconsidering that issue. The case is expected to be appealed to the U.S. Supreme Court. This case includes the issue of content by including wanted posters of abortion clinic affiliates on the site. Privacy rights are called into question by including personal information on where people live and their phone numbers and includes information about family members and friends. Access issues are also part of the case as the site content provider, Neal Horsely, has had difficulties in finding Internet service providers who will allow the site to be offered.

Again, questions over who owns information and whether this information can be shared easily over the Internet are called in to question where personal information is concerned. Traditionally, medical information is owned by the doctor or hospital providing service to an individual who had no control over the transfer of this information. Recently however, patients have been accorded more rights over what information health care providers may share without their permission. Individuals do not own their addresses or phone numbers so the U.S. Post Office and phone companies can sell this information to others without the permission of individuals. Banks and credit report companies often share personal financial information of individuals. Personal consumer habits and behaviors are tracked and sold without the knowledge of the individual. The Internet facilitates sharing information, which is of value to others and much of this sharing is beyond the control of the individual. In some cases, companies offer individuals to "opt in" or "opt out" of information sharing. Identity fraud is on the increase with so much personal information available via Internet sources. The government is examining many of these areas, especially in the medical and financial sectors, to balance competing rights of ownership versus access. Whether the attempt to limit the sharing of personal information is censorship is still under discussion in many arenas on many levels. Courts will need to balance information ownership issues against the rights of owners to share or sell information against the security needs of individuals. For a more in-depth discussion of privacy and security as censorship concerns, see Spinello (2003) on regulating Internet privacy.

CONCLUSION

The Internet is a complex connection of numerous computer networks originally developed to facilitate the ease of sharing information. As the number of users has grown exponentially worldwide, censorship issues have also grown almost at the same exponential rates. The amount of information available has also grown at a tremendous pace. Technology, mainly in the form of software, continues to proliferate rapidly in an intricate point–counterpoint trend where access to information is expanded, followed by software that then limits access. Who should have access to what material is the critical question being asked over and over again in evolving iterations. Many Internet users view the denial of access to information as censorship. In the United States, censorship questions have been traditionally answered legally in courts of law and based on constitutionally protected First Amendment rights of free speech. Internationally, rights of free expression vary widely by geographic area and cultural norms. Global jurisdictional questions are still being debated and will probably continue to be discussed for a long time to come, which is not inherently a negative pursuit. A long-term debate concerning whose laws take precedence or whether new international laws need to be promulgated are needed. We hope the discussions will include a wide range of stakeholders that should result in better legal and ethical decisions regarding Internet access and content control.

GLOSSARY

Censor Legally, a person who acts in an official capacity to preview or inspect works including books and other publications, films, etc., in order to purge inappropriate material; commonly, anyone attempting to prohibit access to information, books, films, etc.

Censorship Legally, the abridgement of the freedom of speech or freedom of the press by government action; commonly, the prohibition of access to information, books, films, etc.

Child Internet Protection Act (CIPA) Successor to both COPA and CDA and signed into law in 2000; restricted the disbursement of e-rate (see e-rate) subsidies to public schools and public libraries to only those agencies that have installed computer filtering software; overturned at circuit court level in 2002 mainly due to concerns over faulty technology only in terms of public libraries. Schools are still required to implement the law as written. Currently under appeal in the U.S. Supreme Court.

Child Pornography Materials in which minors under the age of 18 are depicted participating in sexual acts. Not protected under the First Amendment in the United States as such materials are considered obscene even if they do not technically meet the legally established obscenity standards.

Child Pornography Protection Act (CPPA) Signed into law in 1996; expanded child pornography statutes to include graphic depictions of computer-simulated images of minors engaged in sex acts or the use of adults who appear to be underage. In 2002 the Supreme Court overturned the law for being overbroad, as it would, in part, prohibit legal portrayals of teenage sexual behavior in art and literature. Considers that the creation of an image that does not involve actual children is not child abuse.

Children's Online Protection Act (COPA) Successor to Communications Decency Act and signed into law in 1998; aimed at commercial Web vendors and required some form of proof of adult status and identification be collected prior to disseminating pornography. The law was challenged on the basis that it impairs access by making adults provide identification to access legally protected speech. In 2002 the Supreme Court reviewed the case and subsequently remanded it back to the lower court to review whether the act is constitutional or overbroad under the community standards principle.

Communications Decency Act (CDA) Signed into law in 1996 in an attempt to regulate the distribution of pornography on the Internet with special concern regarding access by children. Ruled unconstitutional by the U.S. Supreme Court for being vague and overbroad in scope concerning indecent speech and for potentially limiting Internet content for adults to only that which is appropriate for minors.

Copyright A right granted that allows the creator of a work, the work-for-hire of a creator, or one who has purchased the right from a creator to control the dissemination of the work.

Cryptology The practice of encryption and decryption or decoding (see "encryption").

Deep-linking The practice of bypassing a Web site's homepage to connect to a subordinate page.

Digital Millennium Copyright Act (DMCA) Passed by the U.S. Congress in 1998; addresses issues of copyright privileges and rights in the digital arena. Key areas focus on prohibiting the development of technological tools to unlock digital protections placed on electronic materials to deter copying and sharing.

Encryption The practice of coding information in such a way that renders others unable to decrypt without the use of a secret key or password.

E-rate Telecommunications discounts offered by the U.S. Government to public schools and public libraries.

Fair use Establishes provisions for copyright law under which materials may be copied and distributed for educational purposes. In the digital arena the issue of "fair use" is being reconsidered.

First sale Establishes provisions under copyright law where the consumer is granted the right to subsequently loan, sell, or give that copy to someone else. Also under review in terms of the digital arena.

Inappropriate materials Books, movies, images, etc. that are deemed not proper or suitable. Most often used to distinguish "age" appropriate materials for children.

Indecent materials Books, films, images, etc. that offend public norms and mores.

Intellectual property Property rights granted to creators or rights holders of copyright, patents, trademarks, trade secrets, and fair competition.

Obscene A legal designation of materials that fall outside the bounds of community standards of decency. Usually such materials appeal to lewdness and lasciviousness.

Open source code Software where the programming code is published openly so that a wide range of people can access the code and add to it; opposite of where code is considered proprietary and thus considered a trade secret.

Patent A federally granted right to control the dissemination of something that someone has invented, improved, or produced.

Patriot Act Passed by Congress in 2001 and affords the U.S. government extended powers in gathering information regarding possible terrorist threats by lessening individual civil liberties concerning privacy and search and seizure.

Pornography Materials depicting sexual or erotic acts or behaviors with the sole intent to arouse or excite in a sexual manner.

Trademark A distinctive label, either a motto, name, symbol, etc., that an organization uses to identify or advertise its products. The Patent and Trademark Office in the United States is in charge of issuing patents and trademarks in addition to overseeing and monitoring conflict.

CROSS REFERENCES

See *Copyright Law; Cyberlaw: The Major Areas, Development, and Provisions; International Cyberlaw; Legal, Social and Ethical Issues; Patent Law; Privacy Law; Trademark Law.*

REFERENCES

Against Intellectual Property Homepage (2002). Retrieved May 8, 2002, from http://www.andrew.cmu.edu/~ctb/anarchy/intelprop.html

American Library Association (ALA) v. U.S. Civil Action, 201 F. Supp. 2d 401 (2002).

Ashcroft v. American Civil Liberties Union, 00-1293 217 F.3d 162 (2002).

Ashcroft v. Free Speech Coalition, 122 S. Ct. 1389, 152 L. Ed. 2d 403 (2002).

Child Pornography Prevention Act, PL 104-208, Title I, §121 (1996).

Children's Internet Protection Act, PL 106-554 (Part of the Consolidated Appropriations Act of 2001).

Children's Online Protection Act, PL 105-277 (1998).

Communications Decency Act, PL 104-104, Title V, Subtitle A (1996).

Deep Linking (2002). Retrieved September 7, 2002, from http://www.ala.org/alaorg/oif/deeplinking.html [Links to articles and commentary about deep linking selected by the American Library Association.]

Digital Millennium Copyright Act, PL 105-304 (1998).

Electronic Frontier Foundation (2002, January 10). Norway indicts teen who published code liberating DVDs. Retrieved September 7, 2002, from http://www.eff.org/IP/Video/DeCSS_prosecutions/Johansen_DeCSS_case/20020110_eff_pr.html

Ginsberg v. New York, 390 U.S. 629, 88 S. Ct. 1274, 20 L. Ed. 2d 195 (1968).

Hazelwood School District v. Kuhlmeier, 484 U.S. 260, 108. S. Ct. 562, 98 L. Ed. 2d 592 (1988).

MGM v. Grokster, CV 01-08541 SVW (2002) (Again, this is the only citation I could find, don't know what format, info is needed)

Miller v. California, 413 U.S. 15, 93 S. Ct. 2607, 37 L. Ed. 2d 419 (1973).

New York v. Ferber, 458 U.S. 747, 102 S. Ct. 348, 73 L. Ed. 2d 1113 (1982).

Nuremberg Files (n.d.). Retrieved September 7, 2002, from http://www.christiangallery.com/atrocity/

Patriot Act, PL 107-56 (2001).

Planned Parenthood of the Columbia/Willimette, Inc. v.

American Coalition of Life Activists, No. 99-35320, No. 99-35325, No. 99-35327, No. 99-35331, No. 99-35333, No. 99-35405, U. S. Ct. of Appeals (9th Cir. 2002).

Reno v. American Civil Liberties Union, 521 U.S. 844, 117 S. Ct. 2329, 138 L. Ed. 2d 874 (1997).

Spinello, R. A. (2003). *Cyberethics: Morality and law in cyberspace* (2nd ed.). Boston: Jones and Bartlett.

Stratton Oakmont, Inc. v. Prodigy Services Co., 23 Media L. Rep. (BNA) (N.Y. Sup. Ct., 1995); motion for renewal denied, 24 Media L. Rep. (BNA) (1995).

UEJF & LICRA v. Yahoo!, Tribunal de Grande Instance de Paris, Ordonnance de RÈfÈrÈ (22 May 2000)

United States v. Thomas, 74 F. 3d 701 (1996).

Warwick, S. (2001). Is copyright ethical? In R. A. Spinello (Ed.), *Readings in cyberethics* (pp. 263–279). Boston: Jones and Bartlett.

World Intellectual Property Organization (WIPO) (2002). Retrieved May 7, 2002, from http://www.wipo.org

Yahoo! Inc., et al. v. La Ligue Contre Le, et al., No. 01–17424, U.S. Ct. of Appeals (9th Cir, San Francisco 2002).

Internet Etiquette (Netiquette)

Joseph M. Kayany, *Western Michigan University*

INTRODUCTION

Netiquette (net + etiquette) can be defined as the informal guidelines developed by the users of the Internet for acceptable online behavior. These are developed over time in a variety of virtual environments and Internet applications. This chapter reviews the generally accepted standards of appropriate behavior for popular applications such as e-mail, newsgroups, chat, FTP, and the World Wide Web.

Defining Netiquette

Netiquette is a blend of common sense, common courtesy, and dictates of the computer technology and culture established by Internet users (Miller, 2001). The norms of behavior are adaptations of real-life etiquette to the unique features of the online technology and environment. Although many of these guidelines have evolved into formal rules that system administrators and owners of proprietary systems enforce, netiquette, in the broader arena of the Internet, is still based on an informal honor system. Etiquette comes from the French word for "ticket" (Shea, 1994), suggesting that compliance with socially defined behavioral norms is the ticket for entry to a network or society. Violators may be ignored and ostracized by the members of the group. Moreover, it is a dynamic social practice that is being continuously interpreted and redefined within the cultural milieu of a collectivity. These norms are rarely universal but specific to each social group or culture, both real and virtual.

Significance of Netiquette

Rules and norms are part of the unique identity of any social collective. Respecting group norms demonstrates fundamental respect to the group itself (when in Rome, do as the Romans do) and is essential to building a community. The Internet as a social collective is, however, very different from other groups. First, it is not grounded on affinities of culture, race, or physical proximity. Secondly, it is growing faster than any other community in history. Traditions and cultural norms are useful in dealing with such rapid pace of change.

As a communication medium, the Internet possesses certain unique features. On the one hand, it is unable, as of now, to convey the full range of nonverbal communication cues, and on the other, it gives users greater control over both the temporal and spatial coordinates of communication. That is, it sustains both synchronous and asynchronous interaction with individuals both far and near. On the one hand, its users can disguise their identities, but on the other, the messages exchanged on the Internet are guaranteed to be neither private nor secure. These unique features do influence the process and outcome of communication that occur through the Internet. The conventions that develop over time around a communication technology are often the by-products of the collective experience of those who have explored its strengths and constraints in diverse communication contexts. Therefore, knowledge of these conventions and practices is essential to anyone who proposes to use these technologies effectively.

Origins of Netiquette

Although early computer networks were primarily developed to enable researchers and academics to exchange research data and documents, those with access to the network often used it for social interaction, as a convenient, informal, and quick substitute to mailed letters or phone tag. In addition to one-on-one communication, store-and-forward applications such as electronic bulletin boards and Usenet newsgroups that allowed the simulation of a group discussion also became popular. Network users could access virtual locations where messages classified by topic were stored in chronological order. Geographically dispersed users were drawn to these shared spaces because of shared interests. Their posts and responses were experienced as group communication activities that led to group bonding, affinity, and identity. Such spatially dispersed collectivities were termed "virtual communities" in opposition to geographically bound real-life communities.

Unregulated Environment in Virtual Communities

Unfortunately, life in these virtual communities was often not as tranquil and congenial as expected. Rebukes, insults, name-calling, and even profanity became frequent on the Net, a phenomenon that came to be known as "flaming." The uncharacteristically unruly behavior on the Net had several explanations. Some argued that the type of people who inhabited the Net in the early days of computer-mediated communication were responsible (Lea, O'Shea, Fung, & Spears, 1992). Early Net users were almost exclusively young computer professionals and hackers, who not only considered themselves at the cutting edge of technology but also of social transformation. The Internet allowed them to break free of the traditional spatial boundaries of communication channels and enabled them to interact outside their cultural boundaries. Flaming was seen as a by-product of this new adolescent hacker culture that defied traditional definitions of appropriate behavior. Flaming was also thought of as a consequence of the nature of the medium itself. The Internet is a text-based communication channel that is low in "social presence"—a measure of the extent to which a person experiences the other person's presence during the communication process. Since the chance of a face-to-face encounter between users is limited, it's easy to feel insulated and anonymous. People are more willing to engage in socially inappropriate behavior when they perceive the other person to be distant—often on another part of the globe (Kiesler, Siegel, & McGuire, 1984).

It was not rare to find such virtual communities becoming war zones where civility and community building gave way to acrimony and hostility. In this context it became clear to many Internet users that civility and mutual respect were vital to maintain the Internet as an acceptable form of social communication. Groups agreed on standard practice and communicated it in publicly accessible files called FAQs (Frequently Asked Questions). Newcomers were expected to read these and comply with the suggested standards. Ironically, many flames were initiated by "old timers" trying to chastise "newcomers" for violating the published norms of online behavior.

Flaming

Sending an impolite and insensitive message that is perceived as an insult or violation of group norms is known as "flaming." A flame can initiate a series of flames or a flame war. Flaming is often an emotional reaction to a post deemed inappropriate by a member of the group. It could be because a new member is perceived to have violated the accepted norms of a group, or because a member is deemed to have wasted everyone's bandwidth by long posts, repetitive questions, indiscriminate cross-posting, or commercial spamming. Sometimes, people respond in anger to any violation of another's privacy when someone divulges personal information or when someone violates standard practices of quoting correct sources (McLaughlin, Osborne, & Smith, 1995).

Flaming is very often sparked by a misunderstanding of the tone of the message. A joke, which would be recognized as such in a face-to-face situation, could be interpreted as offensive by any one of the hundreds of users and it could lead to a flame war. The inability to convey emotions and mood has been a major drawback of the medium.

Emoticons or "Smilies"

One of the ways the online community addressed the lack of nonverbal cues was by developing keyboard-generated symbols that represent various emotions. These nonverbal surrogates are popularly known as "smilies" or "emoticons." For instance, a smiling face created with a colon, hyphen, and right parenthesis on the keyboard :-) looks like a smiling face when you bend your head to the side. A winking smile can be created with a semicolon, hyphen, and right parenthesis ;-) that says "I am only joking" or "I am being ironic." When this combination of keystrokes is added after a text, it tells the reader that the text that precedes it is being written in a jovial mood and ought to be taken in that spirit. Table 1 shows some of the most popular smilies.

Netiquette and the Real World

The early expectations of online media were based on a unique cosmopolitan cyber-culture governed by the value system of a sophisticated educated minority who wished to congregate, communicate, and relate to others like themselves. However, in the 1990s, there was a radical shift in the direction of the Internet's growth. The introduction of the World Wide Web's user-friendly interface was one of the primary reasons for its enormous popularity among the public. The original metaphor of

Table 1 Common Smilies or Emoticons

:-) or :)	Happy; jovial
;-) or ;)	Wink, light sarcasm: "I am only joking," "I am being ironic"
:-(	Sad, disappointed
:-|	Indifference
:-D or :-o	Surprise
:-@	Scream
:-O	Yell
:-*	Drunk
;-}	Leer
:-||	Angry

virtual community gave way to a new way of thinking about the Internet—as an information superhighway and a vehicle for e-commerce. As more people flooded to the Internet for its "informational and commercial" potential, the public's attention was diverted from its potential for companionship and community. The Internet increasingly became another source of information like the radio, TV, newspaper, or the library and another channel of communication like the telephone or conventional mail (often referred to as "snail-mail" because of its physical limitations). Online communication became routine at the workplace, schools, and home. People used e-mail to interact not only with strangers around the globe, but more frequently to keep in touch with their colleagues, neighbors, and friends. As a result, the perception of online media as the lifeline to a unique cyberspace existence and of netiquette as the universal norms of behavior in the virtual world gave way to real-life norms of behavior adapted to the multitude of social contexts. As a result, it is increasingly difficult to talk about a unique set of netiquette guidelines that apply to a homogenous online environment. Instead, netiquette is increasingly a reflection of the physical world, with highly contextualized norms of behavior specific to a group, corporation, society, or community. These norms are extensions of the existing culture applied to the practice of online communication.

Despite the trend toward context-specific norms, an understanding of the traditional practices is useful because they are based on a clear grasp of the strengths and weaknesses of the technology and founded on the core principle of all etiquette—respect for the other. The remaining sections of this chapter are devoted to a discussion of the norms that have found broad acceptance among the users of Internet applications such as e-mail, bulletin boards, chat rooms, FTP, and the World Wide Web. Perhaps the best use of these guidelines would be as benchmarks while individuals and groups discern appropriate behavior in particular communication contexts.

E-MAIL NETIQUETTE

E-mail netiquette guidelines revolve around the central theme—respect. While trying to effectively present the message and ourselves, we should take care to respect the receiver, his or her convenience, bandwidth, and privacy as well as the e-mail conventions and symbolisms developed over time. It is worth reiterating, however, that these norms are neither universally applicable nor appropriate to all communication situations. Interaction history is a key factor that governs decisions regarding appropriate behavior in person-to-person e-mail, whereas in e-mail to groups, factors such as the purpose, size, and geographical dispersion may contribute to the dynamic of discerning what is appropriate in a specific context.

Respect the Other

Salutations

It is safe to use the same salutation that you would use in other forms of communication such as a written letter, telephone, or face to face. For instance, if the correspondent signs his/her first name in the e-mail to you, you can use the first name salutation while replying, although conventions regarding the use of first name ought to be carefully examined while e-mailing to a woman from a different cultural background.

Aim at Consistency

You may check and respond to your e-mail once a day or once a week, but what is important is that you be consistent. People develop expectations on how soon they can expect a reply from you. Inform others if you not available to maintain the routine. Use the auto-responder function of e-mail programs to inform those expecting a reply from you that you are unavailable. However, avoid using the auto-responder option if you are subscribed to discussion groups and listservs.

Be Kind to New Users

Ignore the netiquette violations of others, especially of newcomers. If you must point out errors, do it with compassion, always through a message sent directly to the e-mail address of the offender and not to the entire group.

Redirect Missed Deliveries

If you receive in error a message intended for another, forward it to the appropriate person with a note. If you do not know the e-mail address of the intended recipient, notify the sender.

Signature File

Make sure that pertinent sender information is appended to the end of each message. It is common practice is to use a "signature file" that gets automatically appended to each message. It contains all the necessary information about the sender—name, title, organizational affiliation, mailing address, e-mail address, and phone and fax numbers. Limit the size of signature files to four lines. Avoid long quotes or images built out of keyboard characters that will often appear as gibberish if the default font setting of the receiver is different from the sender's.

Resist the Urge to Respond to Flame

If you are the victim of a flame, do not respond while you are still emotional. It is safe to leave the reply in your outbox for a day before sending it. If a flame was sent to a list of people and you feel compelled to respond to the flame, do not send the reply to the whole group; instead, e-mail your response to the person directly.

Protect Privacy of E-mail Messages

Any personal e-mail you receive is intended for "your eyes only." It is not appropriate to forward such e-mail without the sender's permission.

Protect Privacy of E-mail Addresses

While e-mailing a list of recipients, if you put the e-mail addresses in the "To:" or "CC:" (carbon copy) fields, the e-mail addresses of all the recipients will be displayed to each recipient at the top of the message. When you are sending a common e-mail to a group of people who may not be known to each other, you risk disclosing the e-mail addresses of those who may not want it to be published. This can compromise the privacy of the recipients. No one has the right to give out another person's e-mail address. You can address this problem by putting the e-mail addresses of all the recipients in the "BCC:" (blind carbon copy) field. The list of e-mail addresses will then not be displayed to all recipients.

When members of the group are known to each other, displaying the e-mail addresses of all recipients may have value. For example, in arranging a meeting of a small working group, it may help members to know who has been invited or copied. Nevertheless, it is annoying when one must scroll through a long list of addresses to get to the message.

Respect Copyright

Do not forward anything, articles, graphics, music, or multimedia files that are copyrighted. Remember that personal e-mail is also copyrighted. If you are not sure about the copyright, verify. If the author grants permission to distribute the materials, include the copyright notification.

Respect Another's Convenience

Be Brief

A general netiquette principle is that one's communications should not waste other people's time. There are people who receive hundreds of e-mails daily. Many people pay for Internet connection by the hour. Hence, keep the messages brief and to the point.

Include Meaningful Subject Headings

Use detailed subject headings. Instead of the generic "Greetings" or "Hello" or "Question for you," the subject line should contain a clear summary of your message that allows the receiver to know what the message is about without having to open it. It allows the receiver to prioritize the hundreds of messages he/she receives each day. Many people look for clear subject lines to evaluate the legitimacy of the message because there have been many instances of e-mails with viruses in attachments that arrive in people's mailboxes with innocuous-looking generic subject lines.

Use One Topic per E-mail

Make sure that each e-mail addresses only one topic. Multiple subjects in the same e-mail are harder for the receiver to file and retrieve. Further, if an e-mail with multiple topics must be forwarded to different people for appropriate action, the receiver will have to divide it into separate topics.

Increase Scannability

Increase the ability of recipients to scan the message by writing short paragraphs and by including subheadings if the message is more than a screen-full long.

Use Plain Text

It is tempting to use the formatting capabilities of the new e-mail programs and create a message in HTML with fancy fonts, colors, and images. In addition to clogging the network and increasing the download time, many e-mail programs are unable to display formatting. Often such HTML formatting appears as gibberish making the message unreadable. However, if all the recipients are capable of receiving HTML formatted messages, you may use such formatting. You can avoid the hassle of having to keep track of each person's computer capabilities by sending messages as plain text only.

Table 2 Common Acronyms on the Net

BTW	By the way
IMHO	In my humble opinion
FYI	For your information
LOL	Laughing out loud
ROTFL	Rolling on the floor, laughing
TTYL	Talk to you later
TTFN	Ta Ta for now
Gr8	Great
Ty or 10q	Thank you
OMG	Oh my God
WB	Welcome back
BBL	Be back later
HB	Hurry back
NP	No problem
kotc	Kiss on the cheek

Use Acronyms and Slang with Care

Acronyms and slang should be used only when you are sure that the other person knows the meaning. It is easy to assume that the receiver is familiar with them. Acronyms save keystrokes for the sender but they might make comprehension of the message more difficult for the receiver. Slang may hold different meanings and lead to misunderstandings especially in international settings. Table 2 shows some of the popular acronyms on the Internet.

Refer to the Message in the Reply

When replying to a message, include portions of the original message so that the receiver will immediately know the reference of your response. However, include only those portions of the message relevant to the response.

Respecting Another's Bandwidth

Think of Those with Slow Connections

A majority of Internet users rely on slow modem connections to access the Net. Many people pay for the Internet connection by the hour. Hence, it is important that your e-mails do not result in loss of time or money for the recipients.

Avoid Empty Shells

Do not include the whole text of a long message if you have nothing of substance to say than "OK," "yes," or "me too." In this respect, it is important to respect the customs of the group. If the group values such brief notes of support, encouragement, and congratulations as a way to build group cohesion, they may be perfectly appropriate.

Ask Permission for Attachments

When sending attachments, ask the recipients if they are interested in receiving the file. Always send the file in a format compatible with their system and software.

Verify That Attachments Are Virus-Free

There are now viruses that propagate via people's address books so you may appear to be receiving an attachment from a friend when it is actually a self-generated automated virus. Thus, when sending an attachment to someone it is important to explain what the attachment is and that you certify it is virus-free.

Be Stingy with Copies

Limit distribution of copies of e-mails based on the need to know. Send copies only to those who really need to get a copy.

Be Judicious with the Forward Option

Many a netiquette violation occurs because of the ease of forwarding e-mails. It is so effortless to forward a joke, an inspirational slide show, chain letters, inspirational stories, or virus warnings. Refrain from forwarding anything that you would not spend first class postage if you were to mail it out. As for virus warnings do not forward them unless you have verified them.

Respect E-mail Conventions and Symbolisms

Avoid All-Caps

Use mixed case to type the message. Don't type the messages in all caps. Not only is all-caps text harder to read, but it is also considered the cyberspace equivalent of shouting. This is one of the conventions accepted from the early years of Internet. However, caps may be effectively used to emphasize a word, or at most a sentence.

Avoid All Lowercase

Avoid all lowercase as well because the receiver may interpret it as a sign of sloppiness. All lowercase may sometimes be deliberately chosen to convey that the message should not be construed as formal. Acronyms, informal grammar, and all lowercase are the online equivalent of a scribbled note as opposed to a carefully crafted e-mail that conveys the formality of a typed letter.

Incorporate Emotions

Always be aware that e-mail is a text medium. In the absence of facial expression or voice, the tone of the message can easily be misinterpreted. Smilies may help convey some of the emotions but they should be used sparingly. Use only the most common smilies; it is possible that some of your recipients may not understand the meaning of a smilie (Table 1). Therefore, a verbal indication of the emotion being expressed in brackets can be used instead of smilies to make sure that the receivers are clear about the tone of your message. (Example: {smile}. Some people have begun using pseudo-HTML tags around text to indicate these emotions. Example: <smile> This text is silly </smile.) Do not, however, use smilies in communication environments where formal communication is expected.

Use Humor with Extreme Caution

Body language cues are essential to the interpretation of humor and sarcasm. In the absence of such cues, exercise *extreme* caution while using humor in Internet communication.

Pay Attention to the Language

It is rude to make mistakes in grammar, spelling, sentence structure, and punctuation. Re-read the messages before clicking the send button. However, it is also considered rude to criticize spelling mistakes, typos, and mistakes that suggest the sender is not a native speaker of English.

Respect the Network

Do Not Forward Executable Files

Never forward any executable files. Many viruses are transmitted as attachments whose names end with the suffix ".exe."

Keep Your System Virus Free

One of the dangers of computers being connected to each other is the possibility of a virus-infected computer infecting others. Hence, one of the important netiquette duties of every computer user is to keep his/her system healthy with regular virus checks. However, if you notice that your system has been infected, send a warning to all your e-mail contacts about a possible virus infection that may have reached them through you.

Break the Circle of Hoaxes, Urban Legends, and Chain Letters

These are messages that originate from some evil genius and get recirculated by the newbies and the naïve. Old hoaxes never die; they just get a new life cycle. Some hoaxes have been reborn so many times that they are often considered "urban legends." For instance there is a proprietary cookie recipe that the sender is apparently circulating to get even with the owners of the recipe. Then there is a story of a kid who wanted to get a postcard from as many people as possible before he died. There are plenty of phony messages and pranks on the Net especially as we get closer to April 1 each year. The best thing to do is never forward them or respond to them.

Check the Validity of Virus Hoaxes

The most popular of such practical jokes is a virus hoax that asks the recipients to forward it to as many people as they know. Don't forward it before you check with virus system administrators to ascertain the veracity of a virus warning. This will help you protect your own system as well as those of the others.

Reject Chain Letters

A typical chain letter asks recipients to forward it in exchange for some spiritual or material reward, often with a threat of harm if you don't forward it. Don't forward such messages to anyone unless you know the message is true, or you can authenticate the message and the identity of the sender. In all likelihood, it will be impossible to reach the originator because most pranks and hoaxes have forged headers and signatures.

Respect and Commercial E-mail

Early e-mail users zealously protected the medium as a noncommercial forum so that any attempt to use e-mail for advertising or sale of products and services was considered the most serious violation of netiquette. However, the success of direct mail as a very effective advertising method has encouraged people to blatantly disregard this principle. Advertisers have taken on e-mail because of the ease of distribution and low cost. Despite all protests by those who resent such commercial use of e-mail (known as spamming) as a violation of their privacy and an encroachment on their bandwidth, spamming is not going to go away because it does hit the few interested people. However, every sales person recognizes that respecting the customers and dealing with them honestly generates customer good will. It is the sneaky spammer that people resent the most.

Use Honest Subject Lines

State clearly the commercial purpose of the e-mail. Some advertisers prefix the subject line with "Adv:," which lets the receiver know that the message is an advertisement.

Use One-Line Solicitation

Instead of mailing out a detailed description of the product or service, mail a one-line solicitation asking those interested to e-mail for details. Even those not interested in what you are selling would appreciate your respect for their bandwidth.

Suppress E-mail Addresses

Always use e-mail software that suppresses the e-mail addresses of the other recipients.

Customize Lists

Use customized databases of recipients. Direct your e-mails to those most likely to be interested in your product or idea. It is also appropriate to conduct initial research on how commercial messages are likely to be received and to spare those forums that have clearly stated that commercial mail is unwelcome.

Provide a Remove Option That Works

Give a legitimate option for people to remove their name from your mailing lists. Often when people use the "remove" button to request removal of their e-mail address, they start getting more spam because the remove button was a trap. By requesting removal they have verified that theirs is indeed a "live" e-mail address. This is unethical.

Know the Law

There may be regulations prohibiting commercial e-mail in some countries, in organizations, or in electronic forums. Be aware of these laws before spamming. Even when no laws exist prohibiting such activity, remember that excesses by spammers might result in eventual legislation that will make it difficult to use this technology as a cost-effective marketing tool.

Safeguard Your Own Interest and Image

E-mail Is Not Ephemeral

Do not e-mail anything that you would not say face to face to a person. E-mails can be saved and circulated. Hence do not e-mail anything that can cause you embarrassment if your boss, spouse, friends, or children read it, now or years later. Don't e-mail anything that you don't want to see in print, posted on a bulletin board, in a newsletter, or in front of a jury. Always be aware of the privacy policies of your Internet service provider (ISP) or the corporation that provides you Internet access.

Seek Clarity

Very often e-mail communication occurs with people we do not know in real life. Our impressions of each other are based exclusively on the way text is presented in e-mail. Hence composition is an important factor in the appropriate presentation of the message and of us.

GROUP NETIQUETTE

In this application of computer networks, simulation of group interaction is achieved through store-and-forward computer servers that accept messages from participants. There are three major types of group communication platforms (Table 3), which differ in terms of how they regulate access and deliver messages. Listservs and newsgroups store the messages in the host computer. Newsgroups have an open access policy that allows anyone to access the messages. Listservs restrict access to subscribers only. Access to messages on mail servers is also limited to subscribers but they differ from listservs in how they deliver messages. Listservs store the messages in the host computer that subscribers can access, whereas mail servers forward copies of the messages to the e-mail inboxes of subscribers.

The e-mail netiquette guidelines discussed in the previous section E-mail Etiquette apply to group discussion settings because group posts are essentially e-mails directed at a group. In groups, it is rare that we know every participant. We have no way of knowing who the participants are. Therefore we ought to pay greater attention to netiquette so that we do not cause discomfort to any one

Table 3 Types of Group Communication Platforms

	Access to group	Message delivery
Newsgroups	Public access	Stored in the host computer
Listservs	Access through subscription	Stored in the host computer
Mail servers	Access through subscription	Copies e-mailed to subscribers

of the members. We must be as cautious of our behavior as if we were among strangers in a country with a different culture and language.

Respect the Group

Read FAQs and Lurk

Do not jump right into the discussion. Read the FAQs and the group archives. Introduce yourself if new members are expected to introduce themselves. Then lurk around and listen to the conversation to get a sense of who the group members are, what interests them, and what is acceptable.

Pay Attention to Procedure

Read the instructions on the procedure for subscribing and unsubscribing. Pay special attention to the two separate e-mail addresses commonly used by mailing lists and listservs. The first is for sending "subscribe" and "unsubscribe" requests. The second is for posting messages to the group. Avoid the common mistake of e-mailing to the group a "subscribe" or "unsubscribe" command.

Avoid Empty Messages

If your post does not add value to the discussion and is not of interest to the whole group, don't send it. Avoid messages that contain "test," "hello," or "me too." You can post test messages in test groups. If you want to congratulate or thank a group member, use individual e-mail.

Freeloaders

Do your own homework and try to find the answer before posting a question in a group. Posting a request for help should not be the lazy way out of searching the Internet or going to the library or e-mailing an expert directly.

Auto-responders

Turn off the auto-responder function of the e-mail addresses used to subscribe to a mailing list. Otherwise, to every message forwarded to the e-mail address by the mail server, a receipt response will be send back to the server, which in turn will be distributed to every member, thus creating a large volume of unwanted e-mail in the group.

Cross-Posting

Generally, it is not considered appropriate to send the same post to more than one group. When a message is cross-posted to several related groups to assure that it reaches all those potentially interested in it, people subscribed to several related groups end up getting multiple copies. If you decide to cross-post, you can minimize the annoyance of those who receive multiple copies by making the message brief and providing a link to more detailed information.

Be Generous

Offer answers and help if you are in a position to do so. The usefulness of a group depends on what group members are willing to contribute. Share your expertise and give it generously.

E-mail Responses

If you post a question, don't ask for the answer to be e-mailed to you. Other people may have similar questions and might benefit from the answers posted in the group. Moreover, members might point out a wrong answer if it is posted to the group. Requests that information be e-mailed to you because you don't normally read the group are considered rude.

Summarize

If you've asked a question and received several responses, summarize them and make the summary available so that others might learn.

No Personal Messages

Do not post personal e-mails intended for an individual in the discussion group. Such messages waste the bandwidth of the rest of the group. Messages not of interest to the group should not be posted to the group.

Be Patient with Newcomers

Be tolerant of common netiquette violations of newcomers to the group. If you are compelled to send an emotional response to a member that could spark a flame war, e-mail the message directly to the member and not to the group.

Avoid Commercial Uses

Most group participants resent any effort to use the forum for commercial purposes. Hence, avoid any form of sale, advertisement, or promotion. On the other hand, you may use your expertise in an area of interest to the group and help group members. This in turn can create good will among group members who might eventually become clients or customers.

Respect Copyright and Privacy

Do Not Post Copyrighted Material without Permission

Don't post copyrighted material to the group without explicit permission to do so. Remember, personal e-mail is copyrighted. Never ever post a personal e-mail from another person to the group without permission.

Protect Privacy

Don't disclose personal information—including passwords, credit card numbers, home addresses, or phone numbers to anyone online. Do not post other people's personal information anywhere online.

Public Nature of Groups

Remember that discussion forums are public forums and that messages are archived. Hence, do not post a message that may cause you embarrassment if retrieved and published years later.

CHAT AND INSTANT MESSAGING NETIQUETTE

Chat and instant messaging are Internet applications that allow users to interact with each other in real time. With these applications, simultaneous or synchronous interaction is made possible either in a public forum or in a private space between two or more individuals. Chat and instant messaging are used primarily in an informal, entertainment-oriented social interaction setting, although instant messaging may also be used as a way to manage time and get a quick turnaround on something in a work environment.

Many of the e-mail and group netiquette guidelines discussed already apply to these synchronous communication situations. Those guidelines specific to chat context or instant messaging are those norms based on the unique features of the real-time environment.

Be Brief

Instant messaging is not meant for long messages. Some chat systems limit length of messages. If you have to post long messages, break it up into multiple messages with each ending with a "MORE. . . ."

Choose Appropriate Handles and Nicknames

Some chat programs require you to choose a "handle" or "nickname" in order for users to chat without revealing their true identity. Choose a name that reflects well on you. Do not select names for their shock value. Do not use this facility to impersonate someone else. It is also not considered appropriate to keep on changing one's handle frequently during the same session.

Lurk after Introductions

If you are new to the group, lurk for a while to familiarize yourself with the topic being discussed, the tone of conversation, and above all the culture of the group. Unlike in bulletin boards, however, one cannot lurk anonymously in a chat room. Others are aware that you have entered the room. Hence, if you plan to lurk rather than participate in discussion, it is appropriate for you to introduce yourself and state your reasons for lurking.

Greeting the Group

When you enter a room, greeting the group is appropriate. The form of greeting will depend on familiarity with the group participants. With friends, a *kotc* or a *hugs* is common. (Please see Table 2 for common acronyms.) After the greeting, it is appropriate to wait to be acknowledged by the group members before jumping into discussion.

Label Personal Messages

If a post is directed at one of the members, address them by name so that the group knows that it is meant for one of them.

Respect Private Rooms

Creating a "private" room to communicate with specific individuals in the chat room is a legitimate use of chat rooms. Hence, do not "barge" into "private" rooms where your presence is unwelcome or where you are not invited.

Don't Flood

In a chat environment, flooding is a form of disruptive behavior. Flooding is the sending of a large amount of text to a channel in a short amount of time. One can cause flooding by hitting the return key repeatedly or by inputting large images. This causes the chat room screen to "scroll" so fast that it becomes difficult for other members to read the screen.

In voice chat-rooms, any behavior that would make it difficult to participate in the chat is considered a netiquette violation. For example, leaving the microphone open and playing loud music makes chat difficult for others.

Unsolicited Advertising

Do not post or transmit any unsolicited advertising, promotional materials, or any other forms of solicitation in chat rooms, except in those areas (e.g., shopping rooms) designated for such a purpose.

Know That Chat Can Be Logged

Although a chat environment simulates a face-to-face interaction where the words spoken are ephemeral, chat room conversations and instant messages can be logged and circulated. Hence, judiciousness appropriate to e-mail is applicable to chat messages as well. Sloppy writing and spontaneous flames may come back to haunt you years later.

Say Your Goodbyes

When ready to leave the room, say you are about to leave. Wait a while before you quit so that you will have chance to respond to any messages that were already in transit when you announced your intention to leave. Your last message should clearly state that it is your last post.

Use Alerts

Most chat programs allow the user to indicate his or her status and availability. Use these alerts consistently, so that those wishing to contact you can check on your availability.

Respect Another's Convenience

Technically, instant messaging and chat enable real-time interaction but remember that it is contingent upon the availability and convenience of both parties. Your need to chat or instant message should not disrespect the other party's need or desire not to engage in it at that time.

FTP NETIQUETTE

File transfer protocol (FTP) is an application that allows Internet users to move documents, programs, graphics, etc. to and from a network server. Authorization is required to access most network computers; however, documents and programs for public distribution are saved in a section of these computers that can be accessed without username or password. The following netiquette guidelines pertain to the downloading of documents and programs from this public space through a process known as "anonymous FTP."

Sign on with E-mail Address

While logging in as "anonymous," it is customary to respond to the "password" prompt with your e-mail address, although it is not necessary to gain access to the server.

This allows the system administrators to track the level of FTP usage, if they choose to do so.

Download Large Files during Off-peak Hours

When possible limit downloads, especially large downloads (1+ MB), for off-peak hours.

Minimize Connection time

Keep FTP connection times minimal. Don't leave an FTP connection open and unattended if you're not using it. You will be tying up a line for another user trying to access the site. You should respect the time restrictions of the sites you visit, in order to enable other users who want to use the site.

Pay Copyright and Shareware Fees

Check for copyright or licensing agreements before downloading programs. If there is a registration or shareware fee associated with the program, pay the fee if you decide to keep the program after the prescribed free trial period. Sometimes copyrighted software may have been illegally uploaded into FTP archives. Hence, check the copyright. If there is any doubt regarding the copyright of a program, don't download it.

Upload with Permission

Do not upload or download registered or copyrighted software to FTP sites. Don't upload files without the permission from the system administrator of the FTP site.

Check for Viruses

Do not upload virus-infected programs to FTP sites, and conversely, be sure to virus-check all programs you download from FTP sites.

WEB COMMUNICATION NETIQUETTE

Web development has gone through various stages of evolution. In the early HTML days of the Web, when it was a text-based system designed to create an interlinked knowledge pool, content was at the heart of Web development. Later, when it evolved into a broadcast tool to disseminate information, design and presentation became central. Many of the designers took to the Web as a form of creative self-expression than as a medium to communicate information. In recent years, however, there has been greater attention on the usability of Web sites as well as the responsiveness of organizations via their Web sites. The emphasis seems to have shifted from publication of information to effective communication. When the emphasis is on communication, netiquette becomes a key factor. From the point of view of the designers, it involves respect for the users and their information needs.

Design with the User in Mind

Usability pertains to designing and maintaining a Web site such that visitors will be able to meet their information needs in an efficient and user-friendly environment. It involves minimizing download time, increasing scannability of pages, and providing valuable up-to-date information. A good site is logically organized so that visitors can get to the information in three clicks or less.

Make Each Page on the Site Freestanding

Most Web users arrive at a site via search engines. This means that they do not often enter a site through the front page or the splash page. Hence, it is important that every page on the site has the information regarding the organization affiliation of the Web site, the physical location/address of the organization, the date the page was last updated, an e-mail link to contact the webmaster or the owner, and a link to the splash page of the site.

Inform Those Listed as External Links

Many Web sites increase value to their sites by providing external links. In the early years, it was recommended that you request permission for linking be sought. Doing so may not be practical any longer because some of the popular sites may get flooded with requests for permission. However, it is a good etiquette from your perspective to inform those you have linked to your site. The owners of these sites may choose to ignore you or they may reciprocate by linking to your site.

Permission for Borrowed Content

Anything on the Web—graphics, text, audio, video, and the design—is copyrighted. Although legal action can be pursued only if the Web site is registered, it is certainly against all netiquette guidelines to steal from others' Web sites without permission.

Direct Link to Graphics on Another Site

A worse offense is to provide a direct link to a graphic on your Web site. This means that every time a user clicks to your page, the user's browser goes to the linked site to download the graphic. With every hit to your site, the other Web site's bandwidth is being used without any sort of remuneration.

Warning Users of Controversial/Age-Appropriate Content

If the Web site contains material inappropriate for young people, in addition to providing a splash page that warns visitors of the nature of the content, designers should make it easier for people to block Web sites if they consider that the content on your site is inappropriate for themselves or their children. Appropriate key words and description should be included in the metatags to facilitate such filtering.

Consider Bandwidth Limitations

Despite the hype about the broadband Internet solutions, most people access the Internet using modem connections. High-resolution graphics, audio files, and video increase download time. Ten seconds is considered the threshold of frustration. Designers should make an effort to design pages low on graphic content so that their pages will load under 10 s. Thumbnails of larger images should be used so that users can decide whether they want to open the large picture.

Design for Cross-Browser Compatibility

Colors, graphics, and text are displayed differently on different browsers and monitors. Designers should seek to achieve cross-browser compatibility by including only those features available on all browsers. Web-safe colors and generic fonts should be used as far as possible. Do not design for the latest versions of browsers, plug-ins, or monitors. If tables are used to design pages, make sure that the table width is not greater than 600 pixels. Be wary of using design features such as frames that older versions of browsers do not support. If a page requires a plug-in or program to correctly open it, provide the link to the site where the program can be downloaded.

Make the Site Accessible

When providing audio and visual content, provide alternate text content that conveys essentially the same function or purpose as auditory or visual content. Add ALT tags to every nontext element on the site to make the site accessible to the vision-impaired, those using text browsers, those using nongraphic channels such as pagers and mobile phones, or those who have turned off the graphic display capability of their browsers. When using color, make sure that text and graphics are understandable even without color. If you use tables for layout, provide a nontable version to allow technologies such as screen readers to access the content.

Privacy Issues

Web designers have developed ways of tracking the users by implanting cookies in the user's computer. Cookies can make life easier for the users by remembering previous settings, passwords, etc. However, companies often use them to gather valuable data on the Web-browsing habits of users. When done surreptitiously, it is not only a violation of netiquette but unethical as well.

Increase Responsiveness

It is common practice among Web designers to include an e-mail address to contact the owner of the Web site. In addition to the convenience associated with it, publishing the e-mail address is also symbolic—that you welcome communication from the users. As a result, users expect a reply. A timely response is good netiquette. Many organizations have not yet recognized the significance of being responsive to e-mails that originate from their Web sites. It is probably more important in creating a good impression of the company than an expensive and sophisticated Web site.

Copyright Issues for Web Users

The area where Web users violate netiquette most frequently involves copyright. Browsers have made it easy for people to download documents, graphics, audio files, and video. It is easy to assume that when there is no clear copyright statement, it is not copyrighted. You may not use graphics, text, audio, or video from other Web sites on your own site unless the owners clearly indicate that the material is not copyrighted.

INTERNATIONAL CONSIDERATIONS

Because Internet technology and technical standards were developed in North America, it is no surprise that English became its lingua franca and norms of behavior were developed in the context of the cultural and ethical standards of the West. As the Internet becomes a global medium of social communication, it is likely that netiquette guidelines will become increasingly situational and culture-specific. This will make it harder for Net users to be sure that they do not give offense. In the case of e-mail it is not so hard to be sensitive to the cultural and social context of the recipient because we know who the recipient is. However, it becomes more difficult in discussion groups and Web sites where the recipients are not known. In these contexts, consider the following steps to avoid any misunderstanding among international audiences.

Use Generic Language

Remember that the majority of people who read and understand English are not native-born English speakers. Hence, as far as possible, use simple language. Avoid slang, acronyms, and contextual innuendo. It is rare that people in one part of the world will understand local slang or be aware of an acronym in another part of the world.

Use Universal Time and Measurements

Use the date and time format that will be understood by all when e-mailing to someone in another country or to a group, which is accessed by people from different parts of the world. Use September 11 instead of 9/11, which may be confused for 11th September or 9th November. Time is best written as 2:30 p.m. GMT or 14:30 GMT. When indicating measurements, use both metric and English systems (kilometers and miles, pounds and kilograms).

Use Humor with Extreme Caution

Even the funniest stories can offend someone in some place. Hence, use humor with extreme caution when an international audience has access to the post.

Deciding on an Appropriate Level of Formality

Outside North America, people in positions of authority and those who are older expect certain deference from those younger. Hence, you must consider age and rank while deciding the tone of the e-mail. In some parts of the world, all forms of written communication are considered formal. As long as you are not sure how formal you are expected to be, start the initial e-mails in a formal tone. Eventually, you will know the extent to which you can become informal.

ENFORCING NETIQUETTE

Increasingly, netiquette norms are becoming formal rules that require compliance from users. Group moderators and ISPs have set them up as conditions for access. Courts have upheld the right of the ISPs to suspend services to those who violate these norms of behavior. The decentralized nature of the Internet prohibits centralized enforcement mechanisms or netiquette-enforcement agencies that supervise online behavior. On the other hand, informal enforcement of netiquette has been fairly successful.

However, this system of peer enforcement requires active engagement by the Internet community. Instead of leaving enforcement to a handful of netiquette vigilantes who flame, pursue, and track down violators, each Net user is encouraged to educate and enforce appropriate Net behavior.

Help Educate/Inform the Newbie

Most netiquette violations are committed by newbies who do not have the patience to read the FAQs. Politely inform them of violations through direct e-mail. Do not post a flame to the group.

Ignore the Flamer and the Evangelist

It is wiser to provide silent treatment to those flooding the Internet forums with a constant flow of propaganda. Today's e-mail and bulletin board programs allow us to filter out messages from e-mail addresses and domains with whom we do not want to have interaction. Known traditionally as the "kill file," this feature can be used to effectively ban people from your network.

Report Violators

Despite having strict rules, most ISPs do not have a mechanism to identify violators. Hence, they want members to report violators. Most ISPs have an e-mail address for reporting violations. Usually, it is abuse@hostname.domain. For instance, AOL's abuse report address is abuse@aol.com. Whenever someone consistently violates rules regarding flaming and spamming, take time to report them to the appropriate address.

CONCLUSION

Because of the asynchroneity of the medium, the Internet is perhaps the most polite and considerate form of communication. E-mail messages don't interrupt recipients during dinner (Miller, 2001). Respect for the other is ingrained in the very nature of e-mail.

Despite the speed of delivery and apparent informality, the Internet can also be a tool for deliberate communication. It helps us avoid spontaneous, thoughtless blurting out of words that we may later regret. The Internet allows us to compose a message with thought and consideration for the other person. The absence of nonverbal cues is often cited as a weakness of computer-mediated communication. The lack of social presence helps us avoid unpleasant situations. We can walk away from a confrontational situation by ignoring an insult or avoid people with whom we do not want to interact without being obvious about it.

As the Internet becomes a popular medium of communication, it will take its rightful place among the channels of communication as one of the options available to us. Thus, choosing a medium respectful of the communication situation and the people involved in it will itself become the ultimate netiquette guideline.

GLOSSARY

Bandwidth The information-carrying capacity of the network connection. The greater the bandwidth of the connection, the shorter the download time.

Blind carbon copy (BCC) A feature that hides the e-mail addresses to which the message is sent.

Bulletin board A location in the computer network where people post their messages so that others who access the site can read them.

Cross-post Posting the same message across several bulletin boards and newsgroups.

E-mail lists Grouping of e-mail addresses under one title in order to distribute the same e-mail automatically to this group of addresses.

Frequently asked questions (FAQ) A document available in most discussion groups/chat rooms that describes the purpose of the group, the rules, and procedures for membership and participation; may also be found at a Web site, providing basic information about the site, its services, products, and terms and conditions.

Flame An emotionally charged message posted in a discussion group or sent via e-mail that insults or rebukes others.

Flame war A sequence of flames.

Flame-bait A provocative post intended to upset readers in order to elicit an emotional response.

Flood Sending a large amount of text to a channel in a short amount of time; considered rude in a chat environment.

Frames A style of Web design that creates on the same browser screen multiple windows to display different pages simultaneously.

Hypertext markup language (HTML) The coding scheme, developed by Tim Berners-Lee, used to format Web pages.

Kill file The list of e-mail addresses or domain names that a person has assigned to be excluded from his/her communication network.

Listserv A small program used to automatically redistribute e-mail to names on a mailing list, commonly used to sustain a discussion group on the network.

Lurking Observation of the proceedings in discussion groups and chat rooms; that is, reading the discussion and chat threads without contributing to the conversation.

Metatags The information in the ;<HEAD> section of an HTML document that contains data about the HTML document that is not displayed by the browser. Metatag key words and description are useful to search engines to classify and describe Web sites.

Newbie Newcomer to the Internet in general or to an Internet application.

Newsgroup Name of a discussion group on Usenet.

Spamming Unsolicited distribution of the same message to many e-mail addresses to promote an idea, product, or service; also used maliciously to disrupt communication.

Splash page The opening page of a Web site that uses an animated design and serves as the title page of the site.

Store and forward Communication network systems that store the messages for a time before transmitting them to the ultimate recipients.

Usenet A world-wide system of discussion groups that predates the Internet.

Virtual communities An online collectivity of geographically dispersed users who are drawn to the group because of shared interests.

CROSS REFERENCES

See *E-mail and Instant Messaging; Internet Literacy; Internet Navigation (Basics, Services, and Portals); Online Communities.*

REFERENCES

Kiesler, S., Siegel, J., & McGuire, T. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist, 39,* 1123–1134.

Lea, M., O'Shea, T., Fung, P., & Spears, R. (1992). "Flaming" in computer-mediated communication: Observations, explanations, and implications. In M. Lea (Ed.), *Contexts of computer-mediated communication* (pp. 89–112). New York: Harvester-Wheatsheaf.

McLaughlin, M. L., Osborne, K. K., & Smith, C. B. (1995). Standards of conduct on Usenet. In S. Jones (Ed.), *Cybersociety* (pp. 90–111). Thousand Oaks, CA: Sage.

Miller, S. (2001). *E-mail etiquette: Do's, don'ts, and disaster tales from People magazine's Internet manners expert*. New York: Warner Books.

Shea, V. (1994). *Netiquette*. New York: Albion Books.

FURTHER READING

General

http://www.fau.edu/netiquette/net
http://www.dtcc.edu/cs/rfc1855.html
http://www.rand.org/publications/MR/R3283
http://songweaver.com/netiquette.html
http://www.bspage.com/1netiq/Netiq.html
http://www.templetons.com/brad/emily.html
http://www.learnthenet.com/english/html/09netiqt.htm
http://www.ukindex.co.uk/begin8.html
http://www.albion.com/netiquette

E-mail Etiquette

http://www.askmen.com/fashion/how_to/1_how_to.html
http://unquietmind.com/email.html
http://www.learnthenet.com/english/html/65mailet.htm

Hoaxes

http://www.urbanlegends.com/
http://www.nonprofit.net/hoax/
http://www.vmyths.com/

List-servs, Bulletin Boards, Newsgroups

http://list-etiquette.com
http://www.cs.tut.fi/~jkorpela/usenet/dont.html
http://www.nmt.edu/tcc/help/news/idiot.html
http://advisor.uchicago.edu/docs/general/g_news-eti.html
http://www.faqs.org/faqs/usenet/posting-rules/part1
http://www.cs.indiana.edu/docproject/zen/zen-1.0_toc.html

Chat and Instant Messaging

http://www.cyfernet.org/cybercamp/chatrule.html
http://www.aol.co.uk/communities/etiquette.html

FTP

http://www.plu.edu/~libr/workshops/ftp/netiquette.html
http://www.columbia.edu/acis/fifteen/basics/netiqete/sld01.html

Web Netiquette

http://builder.cnet.com/webbuilding/pages/Business/Rules/
http://www.dreamink.com/design2.shtml
http://bobby.watchfire.com/bobby/html/en/index.jsp
http://www.w3.org/TR/WCAG10/
http://www.useit.com
http://www.cwrl.utexas.edu/currents/spring02/slatin.html

Internet Literacy

Hossein Bidgoli, *California State University, Bakersfield*

INTRODUCTION

This chapter provides a basic literacy of the Internet and Web technologies. It provides a brief history of the Internet and then explains domain name systems, navigational tools, and search engines. The chapter defines intranets and extranets and compares and contrasts them with the Internet. The chapter concludes with a brief survey of popular applications of the Internet including tourism and travel, publishing, higher education, real estate, employment, banking and brokerage firms, software distribution, healthcare, and politics. Other chapters throughout the *Encyclopedia* will further explain most of the topics presented here.

THE INFORMATION SUPERHIGHWAY AND THE WORLD WIDE WEB

The backbone of the information superhighway and electronic commerce (e-commerce) is the Internet. The Internet is a collection of millions of computers and network systems of all sizes. Simply put, the Internet is the "network of networks." The information superhighway is also known as the Internet. No one actually owns or runs the Internet. Each network is locally administered and funded, in some cases by volunteers. It is estimated that, in 2003, more than 200 countries are directly or indirectly connected to the Internet. This number is increasing on a daily basis and makes global e-commerce a reality.

The Internet started in 1969 as a Defense Department Advanced Research Projects Agency project called ARPANET. It served from 1969 through 1990 as the basis for early networking research, and as a central backbone network during development of the Internet. Since the Internet began in 1987, it has grown rapidly in size.

ARPANET evolved into the NSFNET (National Science Foundation Network) in 1987. NSFNET is considered the initial Internet backbone. The term Internet was derived from the term "internetworking," which signified the networking of networks. NSFNET initially connected four supercomputers located at San Diego, Cornell, Pittsburgh, and Illinois to form the backbone. Other universities and government labs were subsequently added to the network. These backbones linked all existing networks in a three-level structure:

Backbones;
Regional networks; and
Local area networks (LANs).

Backbones provide connectivity to other international backbones. The NAPs (network access points) are a key component of the Internet backbones. A NAP is a public network exchange facility where Internet service providers (ISPs) can connect with one another. The connection within NAPs determines how traffic is routed over the Internet and they are also the focus of Internet congestion. Local area networks (LANs) provide the standard user interface for computers to access the Internet. Phone lines (twisted pair), coaxial cables; microwaves, satellites, and other communications media are used to connect local area networks to the regional networks. TCP/IP (transmission control protocol/Internet protocol) is the common language of the Internet that allows the network systems to understand each other. TCP/IP divides network traffic into individually addressed packets that are routed over different paths. Protocols are conventions and rules that govern a data communications system. They cover error detection, message length, and speed of transmission. Protocols provide compatibility among different manufacturers' devices.

The NSF and state governments have subsidized regional networks. NSFNET's acceptable use policy initially restricted the Internet to research and educational institutions; commercial use was not allowed. Due to increasing demand, additional backbones were allowed to connect to NSFNET and commercial applications began.

The World Wide Web (WWW or the Web) changed the Internet by introducing a true graphical environment. It has been around since 1989, proposed by Tim Berners-Lee at CERN. WWW is an Internet service that organizes information using hypermedia. Each document can include embedded reference to audio, images, full motion video, or other documents. The WWW consists of a large portion of the Internet that contains hypermedia documents. Hypermedia is an extension of hypertext. Hypertext allows a user to follow a desired path by clicking on highlighted text to follow a particular "thread" or topic. This involves accessing files, applications, and computers in a nonsequential fashion. It allows for combinations of text, images, sounds, and full-motion video in the same document. It allows information retrieval with the click of a button. Hypertext is an approach to data management in which data are stored in a network of nodes connected by links. The nodes are designed to be accessed through an interactive browsing system. A hypertext document includes document links and supporting indexes for a particular topic. A hypertext document may include data, audio, and images. This type of document is called hypermedia. In hypertext documents the physical and logical layouts are usually different. This is not the case in a paper document. In a paper document the author of the paper establishes the order and readers are instructed to follow the predetermined path.

A hypertext system provides users with nonsequential paths to access information. This means that information does not have to be accessed sequentially as in a book. A hypertext system allows the user to make any request that the author or designer of the hypertext provides through links. These links choices are similar to lists of indexes, and can lead the reader on a "custom path."

Any computer that stores hypermedia documents and makes them available to other computers on the Internet is called a server or a Web server. The computers that request these documents are called clients. A client can be a personal computer at home or a node in a local area network at a University or at an organization. The most exciting feature about the Internet and the WWW is that these hypermedia documents can be stored anywhere in the world. A user can as easily jump from a site in the United States as to a site in Paris, France, all in a few milliseconds.

DOMAIN NAME SYSTEMS

Before a user can begin to navigate the Internet and use it for personal use or e-commerce applications, an understanding of domain name systems (DNS) (also called domain name servers) is essential. Domain names are unique identifiers of computer or network addresses on the Internet. The following are examples of domain names:

Netscape.com

Microsoft.com

UN.org

Whitehouse.gov

They come in two forms: English-like names and numeric or IP (Internet protocol) addresses.

Table 1 Organizational Domains (Generic Top-Level Domains) (gTLD)

.com	Commercial organizations (e.g., Microsoft)
.edu	Education and academic organizations (e.g., California State University)
.int	International organizations (e.g., United Nations)
.mil	U.S. military organizations (e.g., U.S. Army)
.gov	U.S. government organizations (e.g., Internal Revenue Service)
.net	Backbone, regional, and commercial networks (e.g., The National Science Foundation's Internet Network Information Center)
.org	Other organizations such as research and nonprofit (e.g., The Internet Town Hall)

The Internet Corporation for Assigned Names and Numbers (ICANN) is the nonprofit corporation that assigns and keeps track of these addresses. This was previously performed under U.S. Government contract by IANA (the Internet Assigned Numbers Authority) and other entities.

IP addresses are less convenient because numbers are more difficult to remember. The English-like names are electronically converted to IP addresses for routing (transferring information from one network to another network). Domain names are used in URLs (uniform resource locator or universal resource locator) to identify a particular Web page. A URL is basically the address of a file or a site on the Internet. For example, in the URL http://www.csub.edu/~hbidgoli, the domain name is csub.edu. Every domain name has a suffix that indicates which top-level domain (TLD) it belongs to. In the above example the suffix is edu, for educational institutions. Combinations of the letters of the alphabet as well as the numbers 0 through 9 can be used in domain names. The hyphen is the only other character utilized; spaces are not allowed.

The top-level domain is the field on the far most right. It denotes the type of organization or country the address specifies. TLDs are divided into organizational (generic) and geographic (country code) domains. (See Tables 1 through 3.)

This system makes it easy to identify the type or location of the organization by looking at the last section of the domain name. Organization, which is the second field from the right, refers to the name of the organization. A name for a small company is as easy as a company name. The two left-most fields of the domain name refer to the computer. This is relevant for large organizations with several levels of subdomains. An example of a relatively complete Internet address is the address of a document in the Virtual Tourist Web site: http://www.vtourist.com/vt/usa.htm. A brief explanation from left to right follows:

Table 2 Some of the Proposed New Domain Names

.aero	For aviation industry
.arts	For entities emphasizing cultural and entertainment activities
.biz	For businesses
.coop	For cooperative or cooperative service organization
.firm	For businesses or firms
.inc	Corporations
.info	For entities providing information services
.law	For those in the legal profession
.museum	For a museum or professionally affiliated personnel
.name	For a noncommercial site associated with a private individual
.nom	For individuals or family names
.news	For news-related sites
.pro	For a site associated with a certified professional or professional organization
.rec	For entities emphasizing recreation and entertainment activities
.shop	For businesses offering goods and commodities
.store	For electronic storefronts
.web	For entities emphasizing activities related to WWW (World Wide Web)
.xxx	For adult content

http—Means of access, hypertext transfer protocol. This is how the majority of Web documents are transferred.

www.vtourist.com—This is the address of a Web site. It is uniquely defined and differentiated from any other Web sites. WWW is an Internet service that organizes information using hypermedia.

vt—This is a path or directory. A server may be divided into a series of directories for a better organization.

usa.htm—This is the document itself. The htm extension indicates that this is an html (hypertext markup language) document. It is the authoring language used to create documents on the Web. HTML defines the structure and layout of a Web document by using a variety of tags and attributes. A separate chapter in the *Encyclopedia* discusses HTML. Most hypermedia documents are written in HTML format. Servers that do not support long extensions display htm, while other servers display html.

Table 3 Sample Geographic Domains (Country Code Top-Level Domains) (ccTLD)

.au	Australia
.br	Brazil
.ca	Canada
.fr	France
.de	Germany
.hk	Hong Kong
.il	Israel
.jp	Japan
.kr	Korea (Republic)
.ru	Russia
.es	Spain
.uk	United Kingdom
.us	United States
.va	Vatican City State
.zw	Zimbabwe

NAVIGATIONAL TOOLS, SEARCH ENGINES, AND DIRECTORIES

Navigational tools allow the user to surf the Internet and search engines provide access to various resources available on the Internet such as library searches for writing a term paper or making a reservation for an airline ticket. Directories use indexes of information based on key words in the document. As will be discussed later in the chapter, Yahoo! is the most popular directory on the Internet.

The original command language of the Internet was based on computer commands and was difficult to learn by most users. Character-based languages were used for tasks such as downloading files or sending e-mails. These languages are UNIX-based, meaning the user was required to know the specific syntax of many commands. Everything was communicated in plain text, and graphics, sound, and animation data were not available. The introduction of graphical browsers such as *Netscape Navigator* changed all of this. *Microsoft Internet Explorer* and Netscape Navigator are the best-known graphical browsers available for navigating the Internet. Each of these browsers combine powerful graphics and audio and visual capabilities. Each Web server has a "homepage" or a "Web site" that publishes information about the location. Using character-based browsers such as Lynx, a user will find this information in text form, while graphical browsers such as Microsoft Internet Explorer support images and sound clips as well.

Navigational Tools

Microsoft Internet Explorer and Netscape Navigator are among the most popular navigational tools. Microsoft Internet Explorer is the most popular graphical browser in the Internet world. With strong marketing support from Microsoft and improvement in its features, Internet Explorer has gained the leadership in the browser market. Netscape Navigator is another graphical browser available for all major platforms. Netscape, similar to Internet Explorer, provides a true graphical environment that allows the user to surf the Internet using a mouse and the point-and-click technique. Similar to other Windows applications, Netscape Navigator features a standard menu bar and toolbar buttons for frequently used commands.

Directories and Search Engines

There are several search engines and directories in use. Yahoo! is the most popular directory and Google, Excite, and Infoseek are three of the popular search engines. These programs allow a user to scan the Internet and find information. A search could be research for a term paper or finding an exotic antique for a personal collection or anything in between. The following paragraphs briefly describe Yahoo!, Google, and Excite (Bidgoli, 2002).

Jerry Yang and Dave Filo founded Yahoo! in April 1994. Yahoo! is one of the best-known directories on the Internet. A directory is a search service that classifies Web sites into a hierarchical subject-based structure. For example, Yahoo! includes categories such as art, business, and entertainment. These categories are organized by topic. The user can go to a category and then navigate for specific information. Yahoo! also includes an internal search engine that can expedite the search process. Yahoo! soon expanded to offer other services and became a portal on the Internet. A portal or gateway for the WWW is a new application that serves as an information search organizer. Portals provide a single-point integration and navigation through the system. Portals create an information community that can be customized for an individual or a corporation. Portals serve as a major starting site for individuals who are connecting to the Internet. Some of the services offered by Yahoo! include Yahoo! Travel, Yahoo! Classifieds, Yahoo! Pager, and Yahoo! Autos.

Larry Page and Sergey Brin, two Stanford Ph.D. candidates, founded Google in 1998. Google helps its users find the information they are looking for with high levels of ease, accuracy, and relevancy. The company delivers its services to individuals and corporations through its own public site, http://www.google.com, and through cobranding its Web search services. To reach the Google Web site, user simply types "www.google.com" (its URL) into the location box of his/her Web browser and presses the Enter key. At the initial Google screen the user enters the desired search item(s), for example "e-commerce," and then again presses the Enter key or clicks on the "Google Search" button. In a few seconds the items that closely match the search items will be displayed. As the Web site's default language is English, the user can choose a different language by clicking the down arrow to the right.

Excite, Inc. was founded in June 1994. Its basic mission is to provide a gateway to the Internet and organize, aggregate, and deliver information to meet users' needs. The Excite Network, including the Excite and WebCrawler brands, contain a suite of specialized information services that combine proprietary search technology, editorial Web reviews, aggregated content from third parties, and bulletin boards. The Excite Network serves as a central place for consumers to gather and interact during each Web experience. Excite PAL is an instant paging service. By entering the names and e-mail addresses of friends, family, and colleagues into Excite PAL, a user can find them online.

As of March 2002, there were an estimated 114 million Internet users online in the United States at work or at home, 80% of whom are estimated to have made some type of search request during that month.

INTERNET SERVICES THAT SUPPORT ELECTRONIC COMMERCE

Electronic mail (e-mail), news and discussion groups, Internet relay chat (IRC), instant messenger, and the Internet phone are among the services offered by the Internet that could enhance a successful e-commerce program. Other chapters in the *Encyclopedia* will provide a more in-depth discussion of these services. In this chapter a brief overview of these services is presented (Bidgoli, 2002).

Electronic mail or **e-mail** is one of the most popular services available on the Internet. Using e-mail a user can create a letter electronically and send it over the communications media. New products and services can be announced to the customers using e-mail. Confirmations can be sent using e-mail and also many business communications can be effectively performed using e-mail. When a user sends an e-mail, the message usually stays in the recipient's computer until he/she reads it. In many e-mail systems, the receiver is able to store the e-mail message in an electronic folder for future reference. E-mail is fast and will get to the recipient's computer in a matter of seconds or minutes. All is needed in order to send an e-mail message is the e-mail address of a recipient on the Internet. A user can also send a single e-mail message to a group of people at the same time. A user can apply all the word processing tasks such as spell checking and grammar correction before sending the e-mail message. Document files and/or multimedia files can be attached to an e-mail message and a user could ask for delivery notification. With e-mail a user can usually establish various folders with different contents and send a particular e-mail to a specific group. Using e-mail enables a user to establish an effective message distribution system for advertising products and services.

The Internet brings together people with diverse backgrounds and interests. **Discussion groups** that share opinions and ideas facilitate this. Each person in a discussion group can post messages or articles that can be accessed and read by others in the group. **Newsgroups** can be established for any topic or hobby and allow people to get together for fun and entertainment or for business purposes. For example, a user may join a newsgroup interested in ancient civilization, or a user may join a newsgroup that can help in writing and debugging a computer program in a specific programming language. Newsgroups can serve as an effective advertising medium in e-commerce environment.

Internet relay chat enables a user to interactively communicate in a written form with other users from all over the world. It is similar to a coffee shop where people sit around a table and start chatting. The two major differences between this electronic coffee shop and a real coffee shop are that there is no coffee and the user does not see the people that he/she is chatting with. However, a user is able to participate in many different discussions with people anywhere in the world who have the same interest.

Instant messenger is a communication service that enables a user to create a private chat room with another user. Different instant messengers offer different capabilities. They typically alert a user whenever somebody on his/her private list is online, then a user may initiate a chat session with that particular individual.

Internet telephony is the use of the Internet rather than the traditional telephone company infrastructure and rate structure to exchange spoken or other telephone information. Because access to the Internet is available at local phone connection rates, an international or other long-distance call will be much less expensive than through the traditional calling arrangement. This could be a major cost saving for an e-commerce site offering hotline services, help desk, and so forth.

Three new services are now or will soon be available on the Internet:

1. The ability to make a normal voice phone call (despite whether the person called is immediately available; that is, the phone will ring at the location of the person called). In most of the technologies currently available, a "phone-meeting" must be arranged in advance, and then both parties log onto the Internet at the same time to conduct the conversation.
2. The ability to send fax transmissions at very low cost (at local call prices) through a gateway point on the Internet in major cities.
3. The ability to leave voice mail at a called number.

WHAT IS AN INTRANET?

The excitement created by the Internet has been transferred to another growing application called intranets. In simple terms, whatever that a user can do with the Internet he/she should be able to do with an organization's private network, or an intranet.

An intranet provides users with easy-to-use access that can operate on any computer regardless of the operating systems in use. Intranet technology helps companies disseminate information faster and more easily to both vendors and customers and can be of benefit to the internal operations of the organization. Although intranets are fairly new, they have attracted a lot of attention in a very short time (Bidgoli, 1999, 2002).

The intranet uses Internet and Web technologies to solve organizational problems traditionally solved by proprietary databases, Groupware, scheduling, and workflow applications. One should understand that an intranet is different from a local area network (LAN) or wide area network (WAN), although it uses the same physical connections. An intranet is an application or service (or set of applications or services) using the computer networks (the LANs and WANs) of an organization and that is why it is different from LANs and WANs. The intranet is only logically internal to the organization. Intranets can physically span the globe, as long as access is specifically defined and limited to the specific organization's community of users behind a firewall or a series of firewalls.

In a typical intranet configuration, all users in the organization could access all the Web servers. The system administrator must define the degree of access for each user. They can constantly communicate with one another and post information on their departmental Web servers. However, usually a firewall (or several firewalls) separates these internal networks from the Internet (the worldwide network).

Within these departmental Web servers, individual employees can have their own Web pages broken down by department and a series of Web pages. For example the following departments each may include several Web pages as parts of the organization's intranet program:

Finance;
Human resources;
Information services;
Manufacturing;
Marketing; and
Sales.

So what is an intranet? In simple terms, an intranet is a network within the organization that uses Web technologies (TCP/IP, HTTP, FTP, SMPT, HTML, and XML) for collecting, storing, and disseminating useful information throughout the organization. This information supports e-commerce activities such as sales, customer service, and marketing.

Employees can find internal information and they can bookmark important sites within the intranet. Furthermore, individual departments can create their own Web sites to educate or inform other employees about their departments by implementing intranet technology. For example, marketing can present the latest product information, while manufacturing can post shipping schedules and new product designs. The human resources department can post new jobs, benefit information, new promotions, and 401K plan information. The finance and accounting departments can post cost information and other financial reports on their sites. The president's office might post the next company picnic on its site. This information collectively supports a successful e-commerce program.

THE INTERNET VERSUS INTRANETS

The Internet is a public network. Any user can access the Internet assuming the user has an account with an ISP. The Internet is a worldwide network, whereas intranets are private and are not necessarily connected to the Web. Intranets are connected to a specific company's network and usually the users are the company's employees. An intranet is separated from the Internet through a firewall (or several firewalls). Intranets usually have higher throughput and performance than the Internet and are usually more secure than the Internet.

Apart from the above-mentioned dissimilarities, the two have a lot in common. They both use the same network technology, TCP/IP, and they both use browsers for accessing information. They both use documents in HTML and XML formats and both are capable of carrying documents with multimedia formats. Also, they both use the Java programming (or its derivatives) language for developing applications.

Intranets may or may not use any of the technologies beyond HTML, i.e., Java programming, JavaScript or VBScript, Active X, Dynamic HTML, or XML. One of the

Table 4 The Internet versus Intranet

Key Feature	Internet	Intranet
User	Anybody	Employees only
Geographical scope	Unlimited	Limited to unlimited
Speed	Lower than that of an intranet	Higher than that of the Internet
Security	Lower than that of an intranet	Higher than that of the Internet
Technology used	TCP/IP	TCP/IP
Document format	HTML	HTML
Multimedia capability	Could be lower than that of an intranet	Could be higher than that of the Internet

advantages of an intranet is that since the organization can control the browser used, it can specify a browser that will support the technologies in use. Beyond Web documents, the organization can also specify the use of the Internet phone, e-mail, video conferencing, and other Web technologies supported by the chosen browser. Table 4 summarizes the similarities and dissimilarities of these two technologies.

SELECTED APPLICATIONS OF AN INTRANET

A properly designed intranet can make the type of information listed in Table 5 available to the entire organization in a timely manner. This information directly or indirectly can improve the efficiency and effectiveness of an organization (Bidgoli, 1999, 2002).

Many internal applications in use today can be easily converted to an intranet or can be supported using an intranet. Human resources applications, such as job information, name and phone number lists, and medical benefits, can be displayed on a human resources Web site. The finance Web site might present information on time cards, expense reports, or credit authorization. Employees can easily access the latest information on a server. With e-mail, e-mail distribution lists, and chat lines, employees can retrieve meeting minutes.

The intranet also allows organizations to evolve from a "calendar" or "schedule"-based publishing strategy, to an "event-driven" or "needs-based" publishing strategy. In the past, companies published an employee handbook once a year. Traditionally, the handbooks would not be updated until the following year even though they may have been outdated as soon as they arrived on the users' desks. Some of these organizations sent a few loose pages as an update every so often. The employee is supposed to add these additional pages to the binder. After a while these materials become difficult to go through to retrieve specific information.

With an intranet publishing strategy, information can be updated instantly. If the organization adds a new

Table 5 Possible Information Provided by an Intranet

Human Resources Management
401K plans
Calendar events
Company mission statement and policies
Contest results
Department information
Employee classified
Employee stock options
Job postings
Job descriptions
Leave of absence and sabbatical news
Maps
Medical benefits
New hire orientation materials
Online training
Telephone listings
Time cards
Training manuals
Training schedules
Travel authorization
Organizational charts
Meeting minutes
Personnel policy
Press releases
Salary ranges
Software program tutorials
Suggestion box
Upcoming functions
Employment applications
Security policies and procedures
Web usage and e-mail policies

Sales and Marketing
Call tracking
Data regarding the latest actions taken by the competitors
Customers' information
Order tracking and placement
Newscast on demand to desktop, custom filtered to client profile
Sales tips
Product information

Production and Operations
Equipment inventory
Facilities management
Industry news
New product offerings
Product catalog
Project information
Distribution of technical drawings

Accounting and Finance
Budget planning
Credit authorization
Expense report

mutual fund to the 401K programs, content on the benefits page can be updated immediately to reflect that change, and the company internal homepage can include a brief announcement about the change. Then the employees have the new information at their desktops as soon as they look up the 401K programs.

Intranets dramatically reduce the costs and time of content development, duplication, distribution, and utilization. The traditional publication model includes a multistep process including

Creation of content;
Production of the draft;
Revision of the draft;
Final draft preparation;
Migration of the content to the desktop publishing environment;
Duplication; and
Distribution.

However, intranet technology reduces the number of steps to only two (it eliminates the duplication and distribution steps):

Creation of content; and
Migration of content to the intranet environment.

However, content still needs review and approval regardless of the medium used for delivery.

WHAT IS AN EXTRANET?

Interorganizational systems (IOSs) facilitate information exchange among business partners. Some of these systems such as electronic funds transfer (EFT) and e-mail have been used in traditional businesses as well as in the e-commerce environment. Among the most popular IOSs are electronic data interchange (EDI). Both EDI and extranets provide a secure connection among business partners. Their roles in a business-to-business e-commerce are on the rise. These systems create a seamless environment that expedites the transfer of information in a timely matter.

Some organizations allow customers and business partners to access their intranets for specific business purposes. For example, a supplier may want to check the inventory status or a customer may want to check his/her account balances. These networks are referred to as extranets. It should be noted that an organization usually makes only a portion of its intranet accessible to these external parties. Also, comprehensive security measures must ensure that access is given only to authorized users and trusted business partners.

An extranet is defined as a secure network that uses Internet and Web technology to connect two or more intranets of business partners, enabling business-to-business, business-to-consumer, and consumer-to-business communications. Extranets are a network service that allows trusted business partners to secure access to useful information on another organization's intranet. Table 6 provides a comparison of the Internet, intranet, and extranet (Bidgoli, 2002; Fletcher, 1997).

There are numerous applications of extranets in the e-commerce world. One such example is Toshiba America Inc. Toshiba has designed an extranet for timely order entry processing. Using this extranet more than 300 dealers can place orders for parts until 5 p.m. for next day delivery. Dealers can also check account receivable balances and pricing arrangements, read press releases, and much more. This secure system has resulted in significant cost saving and has improved customer service (Jones, 1998).

Another example of an extranet is the Federal Express Tracking System (http://www.fedex.com). Federal Express uses its extranet to collect information and make it available to its customers over the Internet. The FedEx Web site is one of the earliest and best-known examples of an extranet—an intranet that is opened to external users. The customer can access FedEx's public site, and enter his/her tracking number and locate any package still in the system. Using this system a customer can enter all the information needed to prepare a shipping form, obtain a tracking number, print the form, and schedule a pick-up.

Extranets provide highly secure, temporary connections over public and private networks between an organization and a diverse group of business partners outside of the organization. These groups may include

Customers;
Vendors;
Suppliers;
Consultants;
Distributors;
Resellers; and
Outsourcers, such as claim processors, or those with whom the company is doing collaborative R&D or other collaborative work, such as product design.

Extranets not only allow companies to reduce internetworking costs; they also provide companies with

Table 6 Comparison of the Internet, Intranet, and Extranet

	The Internet	Intranet	Extranet
Access	Public	Private	Private
Information	Fragmented	Proprietary	Proprietary
Users	Everybody	Members of an organization	Groups of closely related companies

a competitive advantage, which leads to increased profit. A successful extranet program requires a comprehensive security system and management control. The security system should provide comprehensive access control; user-based authentication, encryption capability, and comprehensive auditing and reporting capabilities.

An extranet offers an organization the same benefits that an intranet offers while also delivering the benefits from being linked to the outside world. Some of the specific advantages of an extranet include (Bidgoli, 2002) the following:

Coordination—An extranet allows for improved coordination among participating partners. This usually includes suppliers, distributors, and customers. Critical information from one partner can be made available so that another partner can make a decision without delay. For example, it is possible for a manufacturer to coordinate its production by checking the inventory status of a customer.

Feedback—An extranet enables an organization to receive instant feedback from its customers and other business partners. It gives the consumers an opportunity to express their views about products or services before those products or services are even introduced to the market.

Customer satisfaction—An extranet links the customer to an organization. This provides the customer with more information about products and services and the organization in general. This also makes ordering products or services as easy as a click of the mouse. Expediting business-to-business e-commerce is definitely one of the greatest benefits of an extranet.

Cost reduction—An extranet can reduce the inventory costs by providing timely information to the participants of a supply network program. Mobil Corporation, based in Fairfax, VA, designed an extranet application that allows distributors throughout the world to submit purchase orders. By doing this, the company significantly increases the efficiency of the operation. It also expedites the delivery of goods and services (Maloff, 1997).

Expedite communication—Extranets increase the efficiency and effectiveness of communication among business partners by linking intranets for immediate access to critical information. A traveling salesperson can receive the latest product information from his/her hotel room before going to a sales meeting. A car dealer can provide the latest information to a customer on a new model without making several phone calls and going through different brochures and sales manuals.

SELECTED INTERNET APPLICATIONS

Several segments of service industries have significantly benefited from the Internet and its supporting technologies. The Internet has enabled these businesses to offer their services and products to a broad range of customers with more competitive prices and convenience. The Internet offers numerous tools and advantages to these businesses to sell their products and services all over the world. Table 7 lists popular Internet applications.

Table 7 Popular Internet Applications

Marketing and advertising
Distance learning
Electronic conferencing
Electronic mail (e-mail)
Electronic posting
Healthcare management
Home shopping
Interactive games
Inventory management
News groups and discussions
News on demand
Online banking
Online employment
Online software distribution
Online training
Online politics (voting, participating in political forums, chat groups, and using the Internet for political fund raising)
Remote log-in
Sale of products and services
Telecommuting
Transferring files with file transfer protocol (FTP)
Video on demand
Videophones
Online demo of products and services throughout the world
Virtual reality games
Online request for proposal (RFP), request for quotes (RFQ), and request for information (RFI)

In the following pages some of the major beneficiaries of the Internet and e-commerce will be reviewed.

Tourism and Travel

Tourism and travel industries have significantly benefited from various Internet and e-commerce applications. As an example, the Tropical Island Vacation (http://www.tropicalislandvacation.com) homepage directs prospective vacationers to an appropriate online brochure after responding to a few brief questions about the type of vacation they would like to take. Customers simply must point and click on appealing photographs or phrases to explore further. Another example is Zeus Tours (http://zeustours.com), which has been very effective at offering unique and exciting tours, cruises, and other travel packages online. Many Web sites allow customers to reserve tickets for planes, trains, buses, cruises, hotels, and resorts. Sites such as biztravel.com (http://biztravel.com) allow its business customers to plan a trip, book a vacation, gather information on many cities, gather weather information, and much more. Expedia.com, Travel.com, Travelocity.com, Priceline.com, hotels.com and Yahoo! Travel are other examples of sites that offer all types of travel and tourism services.

Publishing

Many major textbook publishers in the United States and Europe have homepages. An interested individual can read the major features of forthcoming books or books in print before ordering them. The Web sites of some publishers include a sample chapter from specific books, or entire books that can be read online for free for 90 days, while others allow online customers to purchase portions of rather than the entire book. The Web site of John Wiley & Sons (http://wiley.com), publisher of *The Internet Encyclopedia*, allows a prospective buyer to search the online catalog based on the author's name, the title of the book, and so forth. When the desired book is found, it can be ordered online.

Higher Education

Major universities also have homepages. An interested individual can go on a tour of the university and read about different departments and programs, faculty, and academic resources. Many universities throughout the world are creating virtual divisions that offer entire degree programs on the Internet. Many professional certificate programs are also offered through the Internet. These programs and courses provide a real opportunity and convenience for individuals in remote areas and individuals who cannot attend regular classes to enroll in these courses. They also provide a source of revenue for many colleges and universities that are facing enrollment decline in their service areas. They also allow renowned experts to teach a course to a broad geographic audience.

Real Estate

Numerous real estate Web sites provide millions of up-to-date listings of existing and new homes for sale throughout the world. These sites are devoted entirely to buying and selling real estate. The buyer or seller can review neighborhoods, schools, and local real estate prices. These sites allow the customer to find a realtor, find brokerage firms, and learn many home-buying tips. Some of these sites offer or will soon offer "virtual tours." These virtual tours will enable a buyer to view a prospective property from distance. This is achieved by using virtual reality technologies. Some of the services offered by a typical real estate site are listed below:

Appraisal services;
Buying;
Checking neighborhood profiles;
Checking schools profiles;
Financing;
Home improvement advice;
Obtaining credit reports;
Posting a free listing;
Renting services; and
Selling advice and much more.

Table 8 lists examples of major real estate sites.

Table 8 Examples of Online Real Estate Sites

Prudential California (http://www.prudential.com) provides wireless listing services and property data for agents.
ERA (http://www.era.com) is an Internet-based application with listing information for agents.
Century 21(http://www.century21.com) is an electronic system for tracking agent referrals worldwide.
Re/Max (http://www.remax.com) is a contact management tool for agents that interface with Palm Pilot.
Homestore.com (http://www.homestore.com) provides a listing of more than million properties throughout the United Sates.
Mortgage Expo.com (http://www.mortgageexpo.com) provides a listing of more than 800 home lenders throughout the United Sates.

Employment

Employment service providers have established a Web presence. Table 9 provides a listing of some of the popular sites to use for finding or recruiting for a job, especially if it involves information technology.

Banking and Brokerage Firms

Online banking is here. Many U.S. and Canadian banks and credit unions offer online banking services. Although online banking has not been fully accepted by the customers, many banking-related resources are being utilized. For example, many banks use e-mail to communicate with their corporate customers. E-mail is a less expensive alternative to a telephone call, especially for long distances. Financial reports for banks can be easily distributed via e-mail to mutual-fund investors or customers.

The banking industry's ultimate goal is to carry out many of their transactions through the Internet. Consumer acceptance is the major factor that has kept this business from exploding. It is generally believed that a secure nationwide electronic banking system is almost in place. Soon people will be able to use their PCs and the Internet to do all types of banking activities.

As will be discussed in one of the chapters in the *Encyclopedia*, digital signatures are a key technology for the

Table 9 Some of the Popular Sites to Use for Finding or Recruiting for a Job

http://www.careerbuilder.com
http://www.espan.com
http://www.hotjobs.com
http://www.monster.com
http://www.webhire.com
http://www.dice.com
http://www.guru.com

Table 10 Some of the Services Available via the Internet for Banking Activities

24/7 customer service by e-mail
Accessing the old transactions
Categorizing transactions and producing reports
Exporting banking data to popular money management software
Obtaining online funding for checking accounts
Obtaining online mortgage and CD applications
Obtaining written guarantee against frauds and late payments
Obtaining instant approval for personal loans
Obtaining interactive guides to aid selection of a proper banking product or service
Obtaining interactive tools for designing a savings plan, choosing a mortgage, and/or obtaining online insurance quotes all tied to applications
Obtaining online application for both checking and savings accounts
Obtaining online forms for ordering checks, and issuing a stop payment
Obtaining free checks, and free foreign ATM use
Obtaining IRA and brokerage account information
Obtaining loan status and credit card account information online
Paying bills
Paying credit card accounts
Transferring funds
Viewing digital copies of checks

banking and brokerage industry since they provide an electronic means of guaranteeing the authenticity of the sending party and assurance that encrypted documents have not been changed during transmission. The current mergers and acquisitions taking place and the frequent downsizing within the financial industry strongly support Internet banking. Table 10 lists some of the services available via the Internet for banking activities (Carter, 1999).

Many brokerage firms offer stock and other security transactions online. They provide quotations for stocks, bonds, and other securities. To encourage more customers to use these services, they offer discounts.

Software Distribution

Several major software vendors offer software on the Internet. Customers can view listings of software available, and order and designate an installation time. Microsoft and several other software companies already offer free software via the Internet. Routine downloading of the Netscape Navigator and Microsoft Internet Explorer browser applications are two good examples. Both are relatively small programs. In contrast is the Microsoft Office Suite, which would take significantly longer to download through an online application service provider. Given today's communications throughput and bandwidth limitations, program size will definitely pose a challenge to online software distribution.

A successful application in this area is the distribution of anti-virus programs over the Internet. Several of the vendors of this software application are already using the Internet to sell their software to prospective buyers. A major advantage of this method is the frequent and the automatic updates that the vendors provide for their customers. The Internet makes this process a cost-effective venture for both the vendors and the customers.

The development of online copyright protection schemes continues to be a challenging problem. If users need an encryption code to "unlock" software, backups may not be possible. However, the odds are in favor of online software distribution as it provides an inexpensive, convenient, and speedy method of purchase and implementation (Cross, 1994; Hayes, 1995).

Healthcare

Electronic patient records on the Internet could provide complete medical information and allow physicians to order lab tests, admit patients to hospitals, refer patients to other physicians or specialists, and order prescriptions. Test and consultation results would be directed automatically to electronic patient records. The advantages of this approach include the fact that all patient information would be accessible from one central location. Another positive side of this application is that it would allow easy access to critical health information. Imagine a person who is far away from home and runs into a serious health problem due to injury or other causes. Any physician in any location will be able to download the complete medical history of this patient and prescribe a suitable treatment in a short period. However, these systems may offer disadvantages such as information privacy, accuracy, and currency.

Telemedicine (http://telemedtoday.com) may provide the medical profession with the ability to conduct remote consultation, diagnosis, and conferencing. This could result in major annual savings in travel costs and overhead for medical care professionals. As part of the information superhighway, a personal health information system (PHIS) could conceivably provide interactive medical tools to the public. Public kiosks located in shopping malls would be equipped with user-friendly computer equipment for Internet access. Patients would be prompted through the diagnosis procedure by a series of questions. Premature onset of disease could be minimized with this aggressive and proactive approach (Anonymous, 1994).

Virtual medicine on the Internet may allow specialists at major hospitals to operate on patients remotely. Telepresence surgery, as this is called, would allow surgeons to operate all over the world without physically traveling anywhere. A robot would perform the surgery based on the digitized information sent by the specialist over the Internet. Robots would have stereoscopic cameras to create three-dimensional images for the surgeon's virtual reality goggles. Physicians would operate in a virtual operating room. Tactile sensors on the robot would provide position information to the surgeon so that he/she can feel what the robot feels. Already prescription drugs are sold online and there are several Web sites that offer medical services (Bazzolo, 2000).

Politics

In the United States, in recent years the Internet has become a major promotional tool for all major political contenders in races for the White House, the House of Representatives and Senate, and other races. Political candidates use the Internet to announce the platforms that they are running on, their major differences with their opponents, their leadership styles, forth-coming debates, political events, and so forth. They even use the Internet for fund-raising.

The Internet may facilitate empowering voters and revitalizing democracy. Twenty-first century citizens may vote using a computer connected to the Internet, resulting in increased participation. Part-time legislators may have remote access to Washington and they may be able to remain geographically close to their constituents. Of course, an identification system will have to be in place, which could very likely use voice identification, face scan, finger image, or some other biometric verification technology. If such a system becomes available, then security of voting application, security of voting results, and counting accuracy must be carefully analyzed. Currently, the U.S. House of Representatives is attempting to put all pending legislation online. Presidential documents can be found on the Internet. Full-text versions of speeches, proclamations, executive orders, press briefings, daily schedules, the proposed federal budget, healthcare reform documents, and the Economic Report of the President are available. There are a number of repositories of this information that can be found using search engines.

GLOSSARY

ARPANET (Advanced Research Projects Agency Network) Started in 1969 and served through 1990 as the basis for early networking research, and as a central backbone network during development of the Internet.

Directories Use an index of information based on key words in a document. Yahoo! is the most popular directory on the Internet.

Domain name systems (or domain name servers, DNS) Unique identifiers of computer or network addresses on the Internet. Whitehouse.gov or csub.edu are two examples. The first one uniquely identifies the White House and the second identifies California State University in Bakersfield.

Extranet A secure network that uses the Internet and Web technology to connect two or more intranets of trusted business partners, enabling business-to-business, business-to-consumer, and consumer-to-business communications.

Hypermedia Allows links for combinations of text, images, sounds, and full-motion video in the same document. It allows information retrieval with a click of a button.

Hypertext Provides users with nonsequential paths to access information whereby information does not have to be accessed sequentially as in a book.

Internet A collection of millions of computers and network systems of all sizes. Simply put, the Internet is the "network of networks."

Intranet A network within the organization that uses Web technologies (TCP/IP, HTTP, FTP, SMPT, HTML, XML, and its variations) for collecting, storing, and disseminating useful information throughout the organization.

Navigational tools Allow the user to surf the Internet, Microsoft Internet Explorer and Netscape Navigator being two prime examples.

Search engines Provide access to various resources available on the Internet such as library searches for writing a term paper or making a reservation for an airline ticket. Google.com is an example.

URL (uniform/universal resource locator) The address of a file or a site on the Internet used to identify a particular Web page.

CROSS REFERENCES

See *Computer Literacy; Electronic Commerce and Electronic Business; Internet Etiquette (Netiquette); Internet Navigation (Basics, Services, and Portals); Travel and Tourism.*

REFERENCES

Anonymous (1994, August). Health care on the information superhighway poses advantages and challenges. *Employee Benefit Review,* 24–29.

Bazzolo, F. (2000, May). Putting patient at the center. *Internet Health Care Magazine,* 42–51.

Bidgoli, H. (1999, Summer). An integrated model for introducing intranets. *Information Systems Management, 16*(3), 78–87.

Bidgoli, H. (2002). *Electronic commerce: Principles and practice.* San Diego: Academic Press.

Carter, M. Internet banking incentives: What are banks offering their customers? Retrieved March 28, 2003, from http://216.239.51.100/search?q=cache:EYCUWxpsn_MC:www.icbnd.com/data/newsletter/Bank0001.pdf + % 22carter + Merkle: + www . bank & hl = en & ie = UTF-8

Cross, R. (1994, October). Internet: The missing marketing medium found. *Direct Marketing,* 20–23.

Fletcher, T. (1997, September 29). Intranet pays dividends in time and efficiency for investment giant. *InfoWorld,* 84.

Hayes, M. (1995, January 2). Online shopping for software. *Information Week,* 23–24.

Jones, K. (1998, February 8). Copier strategy as yet unduplicated. *Interactive Week.* Retrieved March 28, 2003, from http://cma.zdnet.com/texis/techinfobase/techinfobase/+Nwq_qt+8vKK_/zdisplay.html

Maloff, J. (1997, August). Extranets: Stretching the Net to boost efficiency. *NetGuide,* 62.

FURTHER READING

Bayles Kalman, D. (2003a). Intranets. In H. Bidgoli (Ed.), *The Encyclopedia of Information Systems: Volume II* (pp. 683–692). San Diego: Academic Press.

Bayles Kalman, D. (2003b). Extranets. In H. Bidgoli (Ed.), *The Encyclopedia of Information Systems: Volume II* (pp. 301–312). San Diego: Academic Press.

Bidgoli, H. (2000). *Handbook of business data communications: A managerial perspective*. San Diego: Academic Press.

Bidgoli, H. (2003). Electronic commerce. In H. Bidgoli (Ed.), *The Encyclopedia of Information Systems: Volume II* (pp. 15–28). San Diego: Academic Press.

Sullivan, D. (2001, December 18). *Search engine sizes*. Retrieved March 28, 2003, from http://searchenginewatch.com/reports/ sizes.html

Sullivan, D. (2002, April 29). *Jupiter Media Metrix search engine ratings*. Retrieved March 28, 2003, from http://searchenginewatch.com/reports/mediametrix.html

Internet Navigation (Basics, Services, and Portals)

Pratap Reddy, *Raritan Valley Community College*

INTRODUCTION

The Internet is a system of computer networks designed to send and receive data. This chapter introduces the basic concepts of Internet navigation. It explains what the Internet is, the infrastructure (both hardware and software) needed to access it, and the services and facilities available. Internet navigation includes the use of communication services such as electronic mail (e-mail), listserv/mailing lists, news groups, Internet relay chat, downloading services such as file transfer protocol (FTP), and search services like the World Wide Web and the ubiquitous Web portals.

What Is the Internet?

The Internet may be defined as a web of computer networks spanning the globe to send and receive data in a seamless fashion. The Internet is a form of organized anarchy; it's in flux all the time, and it's decentralized because no company, organization, or government runs it.

The Origin of the Internet

The Internet sprouted from the Advanced Research Project Agency's ARPANET, an undertaking of the U.S. Department of Defense in 1969. It was conceived to test the idea of a communication link that could withstand a nuclear attack. An in-depth discussion about the history of the Internet may be found in a separate chapter within the encyclopedia.

How Does It Work?

The Internet is based on TCP/IP protocol and the packet-switching technology. Transmission control protocol (TCP) is responsible for breaking up messages into sequences of packets, each approximately 1500 bytes. It makes and breaks the connection between computers on the Internet, whereas the Internet protocol (IP) takes care of routing the data packets on the Internet. At the destination, the TCP software reassembles these packets into complete messages.

As shown in Figure 1, messages are sent from the sender to the router. Then they are sent to the Internet. The router is a special purpose computer that connects different computer networks on the Internet (Andrews, 2001, p. 72).

How Can Users Access the Internet?

There are three ways of connecting to the Internet (Figure 2 depicts the three types of connections):

1. Using the Internet from a school or a college or university. Most likely, the computers at educational institutions have network cards and are connected to the institution's local area network (LAN), or they use cable modems.
2. Using the point-to-point protocol (PPP) to dial up an Internet service provider (ISP) to get connected to the Internet.
3. Using a cable or digital subscriber line (DSL) or similar service that connects to the Internet.

How Do Users Select an ISP?

Selecting an ISP is not difficult. According to Ackermann and Hartman (2002), the basic criteria that one needs to take into account before selecting an ISP are access (including dial-up cost), reliability, service, speed, fees, and features (see http://ispfinder.com for additional information about ISPs).

What Is the Internet Infrastructure?

The infrastructure of the Internet is made up of technologies that enable connections through various modalities; for example regular modem, DSL, cable modem, fiber-optic cable, wireless, and satellite.

Regular Modem

A regular modem and a dial-up number connect to the Internet at speeds of 56 kilobits per second (Kbps), in practice around 49 Kbps (the telephone network limitation).

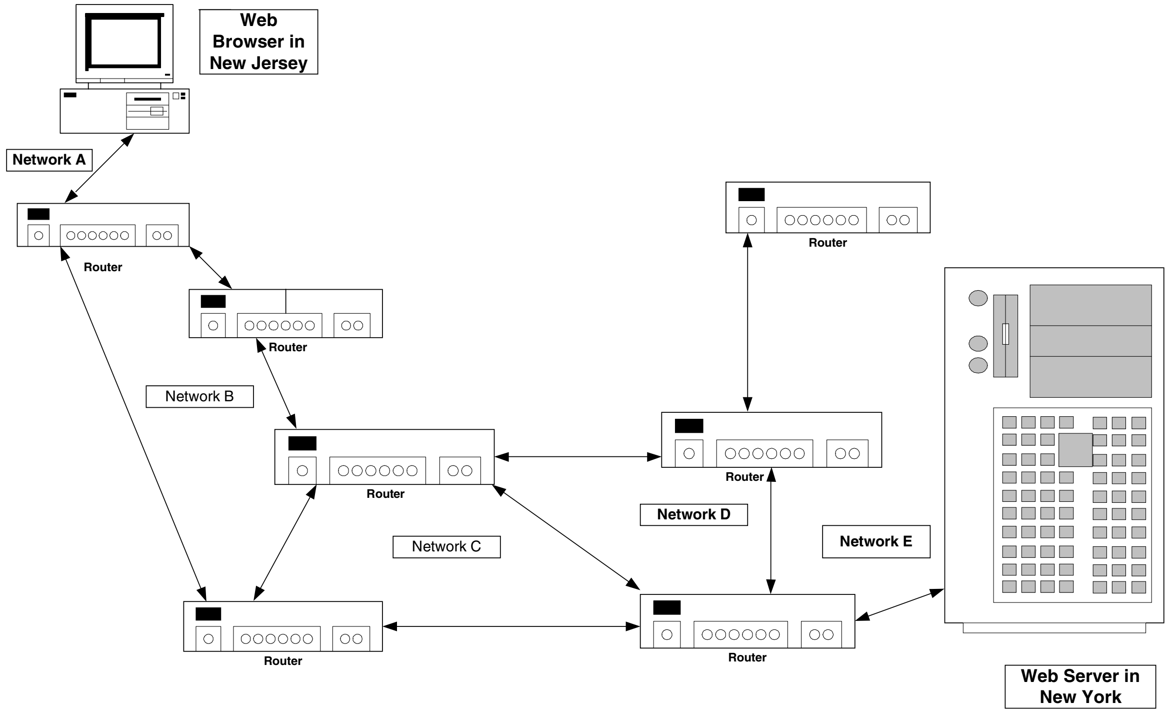


Figure 1: The Internet.

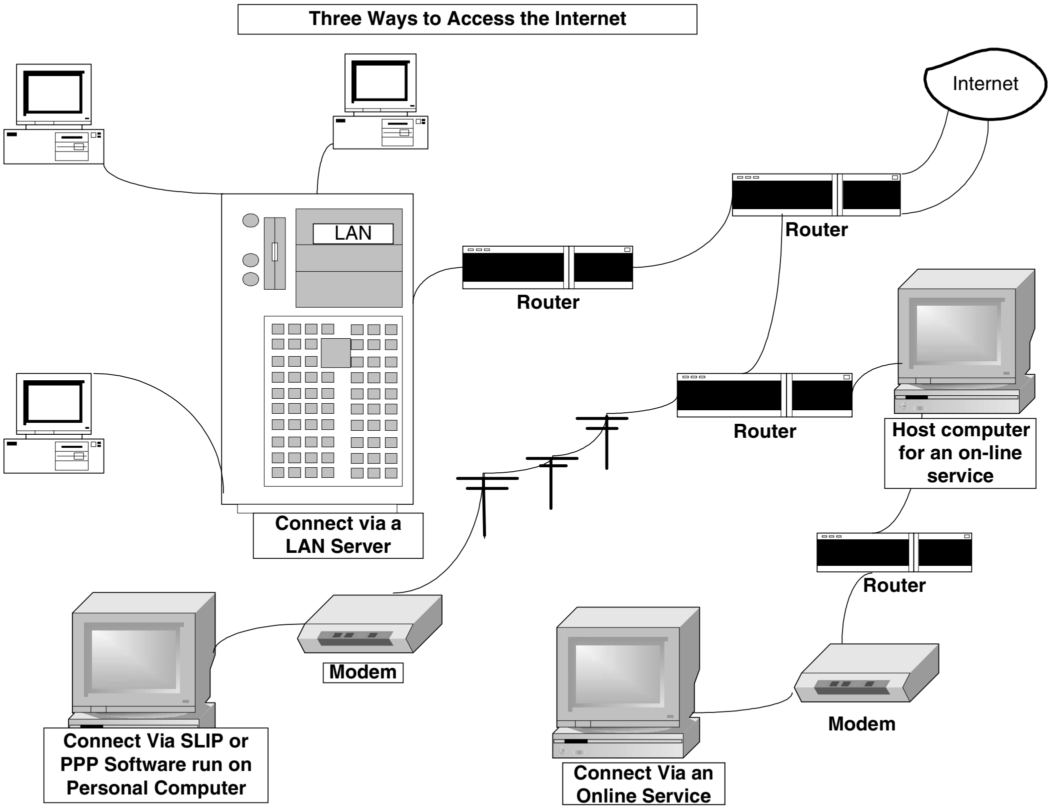

Three Ways to Access the Internet
Internet
LAN
Router
Router
Router
Host computer for an on-line service
Connect via a LAN Server
Router
Modem
Modem
Connect Via SLIP or PPP Software run on Personal Computer
Connect Via an Online Service

Figure 2: Three ways to access the Internet.

Cable Modem

The cable TV network provides speeds of 500 to 2,000 Kbps for downloading data. The upload speeds are similar to a telephone modem—48 Kbps at best. In a cable network, typically 500 to 2,000 homes are serviced by the same signal with a bandwidth of 27 to 39 mega (million) bits per second (Mbps). In the worst case scenario, if all of these users try to access the Internet at the same time, they each will get a share of the bandwidth but only 13.5 Kbps of it—worst than a dial-up modem. However, in practice, it's unlikely that all users will be on the Internet at the same time.

Types of Cable Modem Connections

One-Way. As the name implies, in this type of connection, data can only move downstream. If the user has data to be uploaded, it must go over the regular dial-up connection and thus requires dependence on the plain old telephone system (POTS).

Two-Way. Using this type of technology, the user can send and receive data. Because many customers share a cable, data collision during uploading is likely. A proper timing sequence for data transmission may prevent collisions, however. In the case of downloading, collision is not possible, because the head end or the central office (CO) can schedule downloading for each subscriber as it sees fit to avoid collision (Smith, 2002).

The major players in the cable modem business can be found at the following Web sites: Earthlink (http://www.earthlink.net), Range Broadband (http://www.the-bridge.net), RCN (http://www.rcn.com), and HSA (http://www.hsacorp.net).

Digital Subscriber Line (DSL)

DSL involves using regular telephone lines at a very high-speed connection. There are various types of DSL connections:

Very High Bit DSL (VDSL). The connection speeds for this type may approach 52 Mbps. Telecommunication companies such as QWEST offer this service in certain parts of the United States. This type of connection is effective over short distances.

Asymmetric DSL (ADSL). With this type of connection, the available frequencies of a line are divided up, based on the assumption that most Internet users either look at or download much more data than they send or upload. Connection speeds for downloads are 3 to 4 times faster than speeds for uploading. ADSL is dependent on the subscriber's distance from the provider's central office. The limit is 18,000 feet (5,460 m). Therefore, as the distance of the connection increases, the strength of the signal gets weaker, and the speed of the connection decreases. Although ADSL could theoretically provide an 8-Mbps connection speed at a distance of 6,000 feet (1,820 m), in practice the speed rarely surpasses 1.5 Mbps for downloads and 64 to 640 Kbps for uploads.

There are two subtypes of ADSL: carrier-less amplitude phase (CAP) and discrete multiple tone (DMT). ISPs increasingly use more DMTs than CAPs. Because each ISP uses a different modem, it is important to be aware of this before switching to a new service (http://www.howstuffworks.com/dsl.htm).

Symmetric DSL (SDSL): This type of connection does not allow simultaneous phone and Internet use. As the word *symmetric* implies, the upload and download speeds are the same.

Rate Adaptive DSL (RADSL): This is a variation of ADSL. In this type of connection, the modem can adjust to the speed of the connection depending on the length and quality of the line.

High-Bit-Rate DSL: This technology has long been used in T1 connections. It requires a two-line pair (four wires) as opposed to one-line pair, which is common to most DSL. In a newer type of connection, HDSL-2, a one-line pair is used but it covers shorter distances.

Integrated Services Digital Network DSL (ISDN DSL): This is ISDN-based DSL technology, which cannot be used for voice calls. It provides speeds of 144 Kbps in both directions (Smith, 2002). The following is a list of major DSL service providers in the United States (additional information about DSL service can be found at Broadband Report.com http://www.dslreports.com): AT&T (http://www.att.com), Qwest (http://www.qwest.com), Sprint (http://www.sprint.com), and Verizon (http://www.verizon.com).

The Internet Connection Through Satellites

Even today, many developing nations and some remote parts of industrialized countries lack an efficient wired infrastructure technology such as telephone, DSL, and cable to deliver high-speed Internet access. Satellite Internet is an alternative to address this issue. In such a system, the content provider beams a signal to a satellite that then bounces the signal back to Earth, where it can be received by millions of subscribers.

Although satellites could be employed for Internet communication, little research is being conducted in this area. This may be due to factors such as problems of latency and the cost. Smith (2002) observed that whereas the latency is a mere 10 to 100 ms in the case of DSL and cable modem, it is 239 ms for one-way and 478 ms for two-way satellite communication. This is because it takes 239 ms for the light to reach geosynchronous satellites at a speed of 186,000 miles per second. Two companies in particular that are doing business in satellite Internet connection are Hughes Electronics Corporation (http://www.hughes.com) and Starband Communications (http://www.starband.com/).

Internet Connection Through Wireless Technology

The evolution of wireless technology can be classified as follows: First-generation cell phones use analog technology, second-generation cell phones use digital technology (such as personal services communication [PCS]), and third-generation (3G) cell phones use the 1.9-GHz range, which is also the technology used for laptop computers. The major players in the wireless Internet

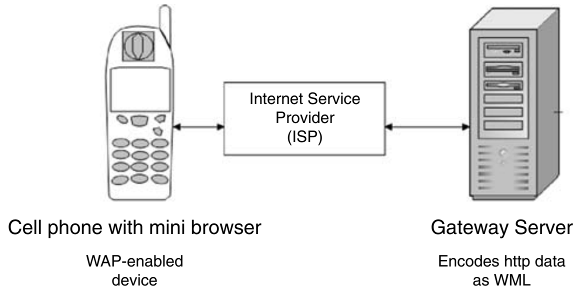


Figure 3: Wireless Internet.

field are CAI Wireless (http://www.caiwireless.com) and Winstar (http://www.winstar.com).

As depicted in Figure 3, accessing a Web site using the Internet-enabled cell phone involves the following steps: (a) The user turns on the cell phone and opens its mini browser; (b) the cell phone sends out a radio signal, searching for service; (c) a connection is made with the ISP; (d) the user selects a Web site to view; (e) a request is sent to the gateway server using wireless application protocol (WAP); (f) the gateway server retrieves the Web site information using hypertext transfer protocol (HTTP); (g) the gateway server encodes the HTTP data as wireless markup language (WML); (h) the WML-enabled data is sent to user's WAP-enabled device (cell phone or other handheld device); (i) the user sees the wireless Internet version of the Web page selected. For an in-depth discussion of wireless Internet, please see the chapter dedicated to that topic in this encyclopedia.

Internet Connection Through Fiber-Optic Cable

Consumers can get an Internet connection through fiber-optic cable via one of the following modalities:

Fiber to the Curb (FTTC)

In this method, fiber-optic cable is run from the central office of the service provider to the curbside of the consumer's office building or apartment complex. From there, an optical network unit (ONU) converts the optical signal to an electrical signal, and then the message is carried into the consumer's apartment or building.

Fiber to the Home (FTTH)

In this case, the ONU device is in the apartment complex or the building. Up to 32 subscribers can access the signal at 622 Mbps speeds for downloading and 155 Mbps for uploading.

The major player in the fiber-optic arena is Surewest Broadband (http://www.surewestbroadband.com/wf/). In a newer FTTH technology, using a dedicated fiber cable between subscriber and the central office, one could achieve 155-Mbps asymmetric speeds (Smith, 2002).

INTERNET SERVICES

There is a vast array of services, programs, and applications available on the Internet—from the ubiquitous e-mail to the arcane gopher and many in between. Some of these programs and applications used to navigate the Internet are the following:

- E-mail,
- Listservs and mailing lists,
- Usenet and news groups,
- Internet relay chat (IRC),
- Telnet,
- FTP,
- World Wide Web, and
- Portals.

This section includes a detailed discussion of Telnet, listservs and mailing lists, Usenet and discussion groups, the World Wide Web, and portals. An in-depth discussion of some of the remaining services can be found in separate chapters in this encyclopedia.

It is important to understand the domain name systems and the data types on the Internet before any discussion of e-mail. The domain name system (DNS) will take care of the translation of a domain name into an IP address or vice versa. For example, if a user were to type the URL http://www.raritanval.edu, the DNS would translate it into an equivalent IP address of 192.231.207.100. We could also type the URL http://192.231.207.100 for a quicker response, but this is not recommended because it is not easy to understand as the regular URL. The *nslookup* can be used to find an IP address of a domain name or vice versa. Users can access this from http://www.dns411.com or http://www.lasaltech.com/cgi-bin/nslookup.

How Does the DNS Work?

When a computer in the United States wants the IP address of a computer in India, it first looks at the local DNS server to find this information. If that does not work, then the U.S. server contacts the DNS server in India to get the information it needs and then provides it to the computer in United States.

Domain Names

In the URL http://www.raritanval.edu, *www* is the computer name, *raritanval* is the organization name, and the *edu* is the organization suffix or main or primary domain. A domain name could be divided further into what is called a subdomain name as follows: http://webct.raritanval.edu, where *raritanval.edu* is the domain name and *webct* is the subdomain name.

International Domain Names

Domain names outside the United States are, for example, as follows: Australia = au, India = in, Japan = jp, China = cn, United Kingdom = uk, South Africa = za. A domain for a university in Japan could be ftp.meiji.ac.jp, for example, where the domain *ac* stands for academic (college or university), analogous to *edu* in the United States, to denote colleges and universities. See the URL http://www.iana.org/cctld/cctld-whois.htm for a list of international domain names.

The following are the top-level domain names that have been in use for some time and the new ones that were approved recently:

Top-Level Domains

.com (commercial)
.edu (higher education)
.net (networking companies)
.org (nonprofit organizations)
.gov (U.S. federal government)
.mil (U.S. military)
.int (international treaties)

New Top-Level Domains

.aero (air-transport industry)
.biz (Business)
.coop (cooperatives)
.info
.museum
.name (for registration by individuals)
.pro (accountants, lawyers, physicians, and other professionals)

Communication

Electronic mail (E-mail)

Asynchronous Communication. In asynchronous communication, the sender and receiver are not necessarily online at the same time. E-mail, Usenet, and listserv are some examples of such communication. E-mail is the most widely used service on the Internet. More than a billion messages are sent on the Internet each day. A detailed discussion of e-mail can be found in a separate chapter of this encyclopedia.

Listserv

A listserv is a system that makes it possible to create, manage and control electronic "mailing lists" on a corporate network or on the Internet. Listservs have been continually improved and expanded to become the predominant mailing list system in use today (Lsoft, 2002).

A subscriber to a listserv automatically receives via e-mail all the messages that are sent to the listserv. Listservs are used on the Internet to send and receive messages on a variety of topics. Because the discussions are asynchronous (i.e., not taking place in real time), users have time to think and respond to the e-mail at their leisure.

There are two types of listservs: moderated and unmoderated. In a moderated listserv, editors screen messages for appropriateness; in an unmoderated listserv, everybody on the list receives every message posted to the list. This may result in hundreds of e-mail messages per day.

How to Send a Message to a Listserv. To subscribe to a Web Designers Listserv, send a blank e-mail message to join@alistapart.com. Do not write anything in the body; leave the CC and BCC blank.

The EDUCAUSE listserv online is a weekly e-mail notification system that summarizes news and information found on the EDUCAUSE Web site.

To subscribe, send a message to LISTSERV@ LISTSERV.EDUCAUSE.EDU. Leave the subject area blank, and in the body of the message write SUBSCRIBE EDUCAUSE-ONLINE YOUR NAME

Example: subscribe educause-online Pratap Reddy

Netiquette. When one is navigating the Internet, basic netiquette, network etiquette, is expected. It includes avoiding spamming and flaming. For additional information on netiquette, see http://www.albion.com/netiquette/index.html. Additional resources of listservs include the following:

- http://www.javascript.com/newsletters/
- http://paml.org (publicly accessible mailing Lists)
- http://www.lsoft.com/lists/listref.html (catalog of Listserv lists)

What Is Telnet?

Telnet is a program that connects a computer to a remote computer. A telnet program can be invoked from Windows 95 or higher (start–run telnet), from UNIX prompt, or from a Web browser. Telnet is used to navigate the Internet for the following purposes:

- To connect to remote servers, such as college and university computer systems, using a valid login address and password. This enables the user to access system resources that are made available to them.
- To connect to libraries and similar organizations, which allow users to login without a password.

Telnet connections do not provide for a graphical user interface (GUI); one would simply see text information. The information retrieval would be faster, however, because it does not contain any graphics. If a user simply wishes to do a library search, for example, then a telnet search would be better than a WWW search, which may involve graphics.

How to Access Telnet From a Web Browser. Type the telnet host address in the Web browser. For example, telnet://nyplgate.nypl.org. The system prompts the user to enter a login name and password. If the telnet host is a library or similar type, then one may not need a password but only a log-in name. This log-in name is usually listed on the telnet session for convenience. In the case of New York Public Library, for example, the log-in name is nypl.

After typing a log-in name and password, if required, the user will be connected to the remote host computer and is then ready to access the resources provided on the host computer.

The following telnet servers can be accessed from a user's Web browser. Log-in information could be found during the telnet session.

- telnet://hp.falcon.edu
- telnet://fedworld.gov
- telnet://freenet.victoria.bc.ca

Another resource for telnet can be found at http://www.udel.edu/interlit/chapter10.html.

USENET and News Groups

User networks also known as Usenet or news groups are asynchronous discussion forums about many topics such as computers, social issues, music, and recreation by people from all over the world. News is obtained from a NNTP (network news transfer protocol) server. For example, Raritan Valley Community College in New Jersey, gets its news from a server called *news.bellatlantic.net*. In turn, this server may get a feed from some other machine. One could subscribe to the Usenet or news groups from a Web browser by connecting to a USENET server. Discussion articles are read with the help of a newsreader. We could also use the Web browsers such as Netscape, Microsoft Explorer, or Lynx to read news. Individual contributions and articles to these news groups are known as *postings*. News groups differ from listservs and Internet relay chat. In a listserv, the messages are directly e-mailed to the subscriber's e-mail box, whereas in a news group, the messages are posted on a news server. To read them, the user has to log in to the news server. When posting messages, however, the user has the option of directly e-mailing the discussion participants. Participants of discussion groups are expected to follow *netiquette* or network etiquette.

Names for News Categories

- Alt (alternative news groups on many topics)
- Comp (computers)
- Rec (recreation)
- Soc (social)

For example, *soc.culture.indian.telugu* is a news group of Telugu speaking people of Indian origin; *alt.rec.humor* is a discussion group devoted to humor; and *comp.lang.javascript* group is for JavaScript programmers, which is under *lang* subgroup, and this subgroup is under the main group *comp* (i.e., computers).

Users can subscribe to these news groups either from an ISP, college, university, or directly through a Web browser from the URL http://groups.google.com.

To access a master list of news groups, go to http://www.magma.ca/~leisen/mlnh/mlnhtables.html.

Internet Relay Chat (IRC)

Internet relay chat is *real-time* communication. Both sender and receiver are online at the same time. An in-depth discussion about Internet relay chat can be found in a separate chapter in this encyclopedia.

Downloading: File Transfer Protocol (FTP)

The file transfer protocol is designed to upload (send) and download (receive) files. To perform FTP activities, one needs a computer or a similar device and a software client or program. Users can access anonymous FTP servers from a Web browser by simply typing ftp://sunsite.unc.edu, for example. A detailed discussion of FTP and other downloading software can be found in a separate chapter in this encyclopedia.

Searching: The World Wide Web

The World Wide Web (WWW) is a collection of hypertext and media information that is accessible through the Internet. Tim Berners-Lee, a European scientist, was instrumental in pioneering the idea of the World Wide Web. The W3C (consortium) has been formulating the guidelines for the expansion of the WWW.

To access documents from the Web server, one usually uses a Web browser, such as Netscape, Microsoft Explorer, and Lynx. The HTTP is used to make the connection between a Web client (browser) and a Web server.

Web Client Software

This is the program used to read documents on the Web. There are two types of Web browsers: graphical and nongraphical.

Nongraphical User Interface (Non-GUI) Web Browser

Lynx is a nongraphical, text-only browser that can be invoked from a UNIX server. Because it does not display audio or video, the download is faster. To learn more about Lynx refer to the following URL: http://lynx.browser.org.

Graphical User Interface (GUI) Browsers

The three most common Internet browsers are

- Netscape (version 4–6, 7),
- Internet Explorer (version 4–6), and
- Opera (version 6.03).

All of these Web browsers have features such as searching and retrieving data, performing basic e-mail functions, and facilitating synchronous (IRC) and asynchronous communication (news groups).

Web Browser Protocols

Web browsers are capable of interpreting the following Internet protocols:

- ftp://wuarchive.wustl.edu,
- http://www.raritanval.edu,
- telnet://nyplgate.nypl.org,
- gopher://gopher.ptloma.edu, and
- news:soc.culture.indian.

A Web browser displays the documents written HTML, extensible markup language (XML), and scripting languages such as JavaScript.

HTML Versus XML

The primary purpose of the HTML is markup and linking of text documents. It tells the browser how to format or style the content. In other words, it says, "make this bold, make that italic."

Unlike HTML, XML focuses on communicating content. It draws relationships between content and leaves the formatting to style sheets such as cascading style sheets (CSS) and extensible style sheet language (XSL).

Web Browser Icons

Back and Forward: If the user clicks on these icons, the previous and following document respectively will be displayed in the Web browser.

Reload or Refresh: These icons will reload or refresh a Web page from the server.

Favorites or Bookmarks: By highlighting the URL in the address or location area of the browser, the user can save the URL as "Favorites" or "Bookmarks" for future use. In Netscape, the user can edit these URLs and save them in the Personal Bar and click on them to visit them. It's a good idea to store the most frequently visited URLs on the Personal Bar.

File: This icon is used to open, save, print, and send via e-mail a document.

Edit: In Internet Explorer, this icon is used to copy, cut, paste, select all, and find something on the page. In Netscape, in addition to these functions, using the preferences option the user can configure the browser to send and receive e-mail, read and post news, and store the home page address.

View: This icon is used to increase or decrease the font size and view the Web page's HTML source code.

Mail: This icon is used to read mail and news. The user needs to set up the Web browser so it knows where to get the mail.

Home: To load a home page, click on home. The user can configure the homepage URL either from tools or Internet options of the Explorer browser or the edit and preferences option of Netscape. The next time the browser loads, it will start with this page, the home page.

History: This icon is used to display navigation activities in the recent past on the Web browser.

Enhancing Web Browser Functionality

Web browsers, particularly, the GUI types, use additional software known as plug-ins, and helper applications to extend their functionality. A plug-in is an application that extends the browser functionality. It runs within the browser. Examples are programs such as Real Audio, Shockwave, and QuickTime. A helper application is when the Web browser encounters a file type such as a sound, image, or video, and it hands off the data to other programs, called helper applications, to run or display the file. Most helper applications are shareware or freeware that can be found on the World Wide Web. One can configure these in the helper application dialogue box found under *preferences* in the options menu of Netscape Web browser. A helper application, unlike a plug-in, runs independently outside the Web browser. Examples include Microsoft Excel.

Ports

In a Web server address, users sometimes use a suffix that ends with a number. These numbers are used to address software or services running on a computer. Using port numbers, a host computer can run several software applications or processes including ftp, telnet, http, and pop all on the same machine, but each on a different port number. The typical port number assignment schema is as follows:

- ftp: 20
- telnet: 23
- smtp: 25
- http: 80
- pop3: 110

To find more about Web browsers, check the URL http://www.upsdell.com/browsernews.

Search Engines

An in-depth discussion of search engines can be found in a separate article within the encyclopedia.

Zipping and Unzipping Files

To save storage space on the medium (floppy or hard disk), one can use an algorithm and substitute a simple code for the repeated words of the text. For a simplistic illustration, let's say we have the text "when it rains, in the Amazon, it rarely rains, but pours." Here, a simple letter can substitute for all the repeated words or numbers and thus save the storage space. Obviously, when one is transmitting this file on the Internet, it takes less time because it is smaller in size. At the receiving end, however, the data has to be restored by filling in the appropriate words in place of the codes used before. Programs such as WinZip or PK Zip accomplish this process of stripping the words and inserting the words before and after.

Users may download this shareware from http://www.winzip.com.

What Is a Cookie?

When a user visits a Web server and performs a transaction, for example, buying an item online, the Web server leaves an identification number on the user's Web client (the Web browser). This information is stored in a place called a *cookie* and stored on the user's computer on the hard drive. The next time the user makes a visit to this server, the browser will send the cookie information to the server so it can recognize the user and his or her past activities. Then, for example, it might greet the user by name and show the pages that were last visited. The user has the option of accepting or rejecting the cookie information by configuring the Web browser appropriately.

Security Risks on the Internet

Users of the Internet should be knowledgeable about the security risks associated with it. These can originate from servers, Trojans, and viruses. A hacker or cracker, a computer miscreant, targets a list of known servers and gains unauthorized access to others' computers or causes computer resources to become unavailable resulting in what is known as *denial-of-service* (DOS) attack.

A *Trojan horse* or *Trojan* is a computer program used by hackers/crackers to gain entry to other computers. The cracker uploads such a program to a popular download site. When a user downloads this program, it gets installed on the user's computer and can delete files on the computer or can function as an unauthorized server whereby the cracker can telnet to this server and access it illegally.

Internet users can take the following measures to reduce security risks: (a) Use antivirus software to detect and warn the user of the presence of Trojan horse; (b) restrict downloads from only trusted sites; (c) configure firewalls to block certain classes of outgoing traffic, for instance, traffic directed to ports associated with protocols that one does not use. A firewall is a dedicated router or computer that protects a network from outside attack and prevents insiders from attacking outside systems. A more detailed discussion of Internet security, firewalls, and viruses can be found in other chapters of this encyclopedia.

Web Portals

What Is a Portal?

Winston Dictionary defines a portal as “a gate, door, or entrance; especially one that is stately and imposing, as of a cathedral.”

Jenny Rickard (2000, p. 3) noted that “portal technology provides the capability to aggregate content from multiple sources, integrate [Enterprise Resource Planning] backbone systems into a role-based self-service transactions . . . access role-based analytic information and, if desired, facilitate commercial transactions.”

The Evolution of Portals

The word portal derives from the meaning of the word *port*, which denotes an entrance. The concept of a Web portal has evolved from the domain of the intranet, which leads to the Internet. Discussing the distinction between an intranet and a portal, Paul Krill (2001) observed that the portals lead the user to a structure on data and a navigation scheme around massive quantities of information. The intranet on the, other hand, operates as a data silo, providing unstructured information and a conglomeration of services. They work together. Whereas the intranet holds the data repository, portals provide the seamless infrastructure to access the information from the intranet.

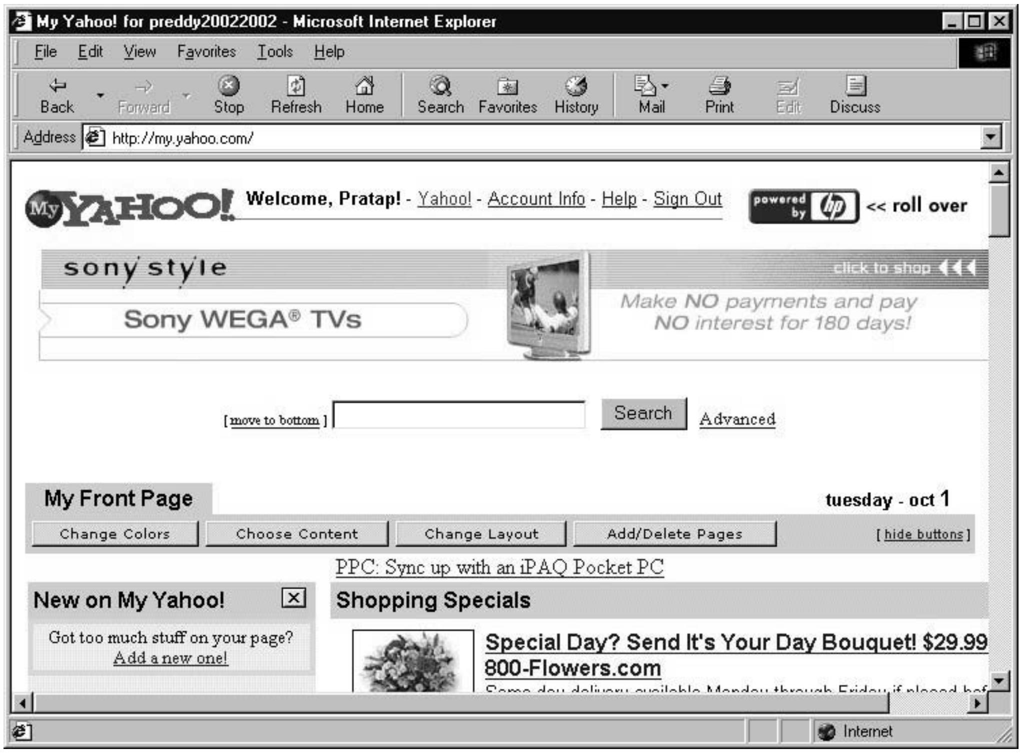
So a portal is a Web site that offers a broad array of resources and services such as e-mail, search engines, online shopping and auctions, chat, instant messaging, personalized calendars, and more. This definition includes the popular community portals such Yahoo, MSN, Excite, and Lycos. The definition of a corporate portal is a bit more elaborate.

Although search engines and portals have similarities, they are different. A key distinction between a search engine such as AltaVista and a portal such as Yahoo is that whereas a portal provides a structure of categorized links to topics such as business, finance, leisure, news, sports, and travel, a search engine provides search results containing the key words that you have entered for the search. See Figures 4–6 for a listing of Yahoo (http://my.yahoo.com) and Google (http:/news.google.com) portals and AltaVista (http://www.altavista.com) search engine results.

Portal Architecture

Whereas a client–server employs a problematic two-tiered architecture, portals employ three-tiered architecture. The weaknesses of two-tiered architecture are many (3- and n-Tier Architectures, http://www.d-tec.ch/e/3tier.html):

- The client or the personal computer processes and presents information that leads to monolithic


My Yahoo! for preddy20022002 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address
My YAHOO!
powered by hp << roll over
sony style
click to shop
Sony WEGA® TVs
[move to bottom]
Search
Advanced
My Front Page
tuesday - oct 1
Change Colors
Choose Content
Change Layout
Add/Delete Pages
[hide buttons]
Internet

http://my.yahoo.com/

Welcome, Pratap! - Yahoo! - Account Info - Help - Sign Out

Make NO payments and pay NO interest for 180 days!

PPC: Sync up with an iPAQ Pocket PC

New on My Yahoo!

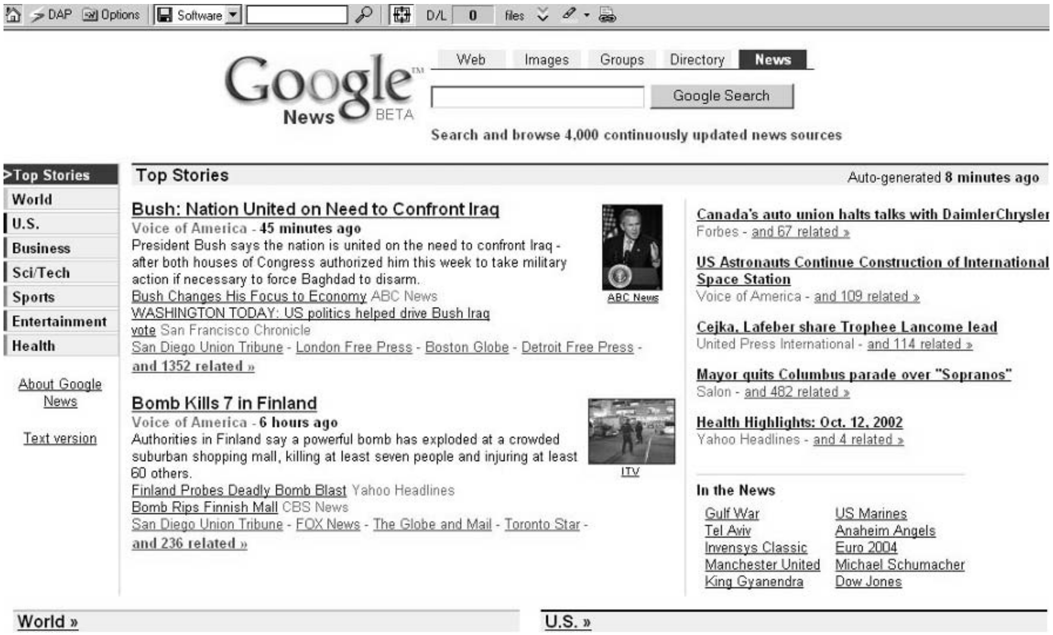
Got too much stuff on your page? Add a new one!

Shopping Specials

Special Day? Send It's Your Day Bouquet! $29.99

800-Flowers.com

Figure 4: Yahoo portal.



Search and browse 4,000 continuously updated news sources

Top Stories Auto-generated **8 minutes ago**

Bush: Nation United on Need to Confront Iraq
Voice of America - **45 minutes ago**
President Bush says the nation is united on the need to confront Iraq - after both houses of Congress authorized him this week to take military action if necessary to force Baghdad to disarm.
Bush Changes His Focus to Economy ABC News
WASHINGTON TODAY: US politics helped drive Bush Iraq vote San Francisco Chronicle
San Diego Union Tribune - London Free Press - Boston Globe - Detroit Free Press -
and 1352 related »

Bomb Kills 7 in Finland
Voice of America - **6 hours ago**
Authorities in Finland say a powerful bomb has exploded at a crowded suburban shopping mall, killing at least seven people and injuring at least 60 others.
Finland Probes Deadly Bomb Blast Yahoo Headlines
Bomb Rips Finnish Mall CBS News
San Diego Union Tribune - FOX News - The Globe and Mail - Toronto Star -
and 236 related »

Canada's auto union halts talks with DaimlerChrysler
Forbes - and 67 related »

US Astronauts Continue Construction of International Space Station
Voice of America - and 109 related »

Cejka, Lafeber share Trophee Lancome lead
United Press International - and 114 related »

Mayor quits Columbus parade over "Sopranos"
Salon - and 482 related »

Health Highlights: Oct. 12, 2002
Yahoo Headlines - and 4 related »

In the News

Gulf War	US Marines
Tel Aviv	Anaheim Angels
Invensys Classic	Euro 2004
Manchester United	Michael Schumacher
King Gyanendra	Dow Jones

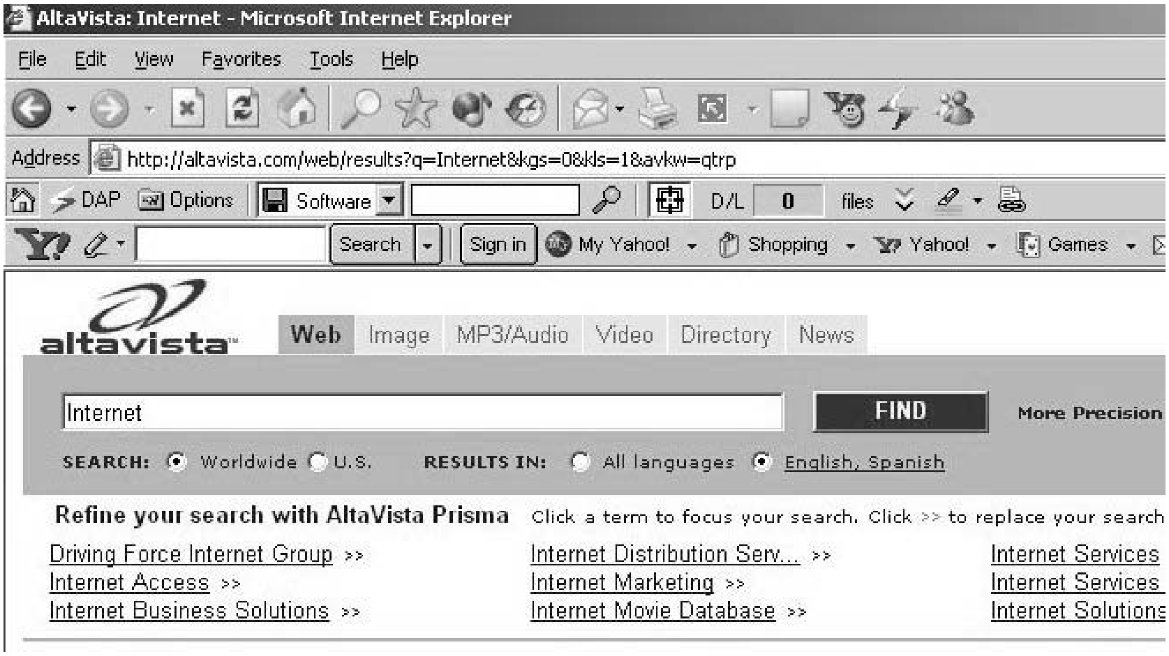
Figure 5: Google portal.

applications that are expensive to maintain which is why these systems are referred to as "fat clients."

- Because most of the work is being done on the personal computer, there would be an increased load on the network to transport data back and forth from the server to the computer.
- Computers are less trustworthy in terms of security because they can be easily cracked.

What Is Three-Tiered Architecture?

Three-tier architecture, as shown in Figure 7, includes a client tier, an application tier, and a data storage tier. In



http://altavista.com/web/results?q=Internet&kgs=0&kls=1&avkw=qtrp

Internet

SEARCH: ● Worldwide ○ U.S. RESULTS IN: ○ All languages ● English, Spanish

Refine your search with AltaVista Prisma Click a term to focus your search. Click >> to replace your search

Driving Force Internet Group >>
Internet Access >>
Internet Business Solutions >>
Internet Distribution Serv... >>
Internet Marketing >>
Internet Movie Database >>
Internet Services
Internet Services
Internet Solutions

Figure 6: AltaVista portal.

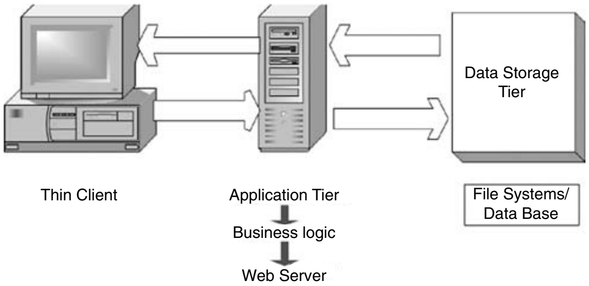


Figure 7: Multi-tier architecture.

this paradigm, the client tier is responsible for the presentation of data, receiving user events, and controlling the user events. This makes it a "thin-client." The application tier hosts the business logic, whereas the data storage tier is responsible for data storage. The benefits of three-tier architecture include the following:

- Development time is quicker because of the reuse prebuilt business logic components.
- Servers are trusted systems; they have fewer security threats compared with clients.
- The business logic and the data storage are closer. The client only receives the processed data, so that network traffic is reduced.

Types of Portals

Based on their features, portals are divided into horizontal and vertical categories. Horizontal portals are also called community or mega portals. They are public Web sites. Examples include Yahoo and Excite. An enterprise portal that caters to a particular market segment is known as a vertical portal (or vortal). In other words, a vortal offers a gateway to information related to a specific industry such as automotive, banking, health care, insurance, law, and others. Examples include esteel.com, a site for buying and selling steel products. A vertical portal used by a university, for example, can deliver the information specific to the college, and in addition it should also function like a public Web portal. Unlike a horizontal portal, it requires authentication information when one tries to login (Katz, 2002). For a detailed list of vertical portals see the link at http://www.verticalportal.com.

The list of portal categories includes animals, arts, beauty, business, computers, entertainment, education, food, home and garden, government, health, living and society, money and investing, news, real estate, reference, shopping, sports, people and family, travel, and vehicles. Each category includes many portals.

Community Portals

The dominant players in the community portal arena include Yahoo, Excite, MSN, Lycos, and Google. In a 1999 report, *The Power of Portals: Who Is Using Them and How,* the International Data Corporation (IDC) and Relevant Knowledge reported that Yahoo and Excite are the two community portals in the *visitor loyalty* and the *duration of visitor stay;* 161 minutes for Yahoo and 160 minutes for Excite. Yahoo is one of the speediest portals on the Web. Its competitors—excite, geocities, and tripod—do a much better job of allowing the user to set up a Web page with built-in message boards and chat. Yahoo's message boards and the chat are the most populated on the Web. "Simplicity and familiarity breed loyalty in the Web-portal market place," observes Glen McDonald (1999), in his article *Portal Profiles: Why Yahoo Is #1.* Community portals make money by selling ads, e-commerce transactions, and sponsorships. For example, the portal http://GoTo.com offers top placement on its search site to the company willing to pay the most (McDonald, 1999).

My.Yahoo. My.yahoo has an excellent directory that will guide the user to other sites if the user decides to leave Yahoo. Notable services include online photo albums and an experts channel (for quick answers to questions on a host of topics). The my.yahoo site is probably the most highly customizable of all the portals. For example, when one configures an http://my.yahoo.com portal, it adds information about movies playing in that locality.

MSN. Microsoft Network, or MSN, is highly configurable, including the layout and colors, even letting the user choose which components to show. One can also add local weather and favorite links and gain quick access to a hotmail account. It can be accessed from http://www.msn.com.

My Netscape. Mynetscape is also highly configurable, including any favorite features (mail, stocks, bookmarks) in one place. It is accessible from http://www.mynetscape.com.

Excite.com. This portal (http://www.excite.com) also has many useful features, as well as the flexibility to customize it to the user's taste.

Google. The news portal (http://news.google.com) recently launched by google.com uses complex mathematical algorithms to cull news from all over the world and display these hundreds of news stories with automatic updates every 15 minutes.

GeoPortals. This portal (http://www.GeoPortals.com) includes groups of sites and search engines organized by subject and subcategories.

Corporate Portals

Sagemaker (n.d.) in its white paper *The Benefits of an Enterprise Information Portal* presents a generally accepted view of the corporate portal as a browser-based environment that

- securely integrates and personalizes internal and external content and applications into a scalable user interface for communities of e-business users across the enterprise,
- provides scalable infrastructure to seamlessly integrate disparate data sources and applications, presenting a unified view across a fully configurable user interface, and

- enables an individual member of staff to manage the flow of information in order to make informed business decisions, to implement these decisions and to be able to communicate with others in making similar decisions (http://www.itpapers.com).

Wayne Eckerson (1999), a senior consultant with the Patricia Seybold Group, characterizes the ideal characteristics of an enterprise portal in this way:

- The enterprise portal must connect to multiple, heterogeneous data stores, including relational databases, multidimensional databases, document-management systems, e-mail systems, Web servers, news feeds, and various file systems and servers.
- An enterprise portal needs to provide a full range of query, report, and analysis capabilities in a highly integrated fashion.
- The enterprise portal must support network-level security, encryption, session-management, and authentication services to safeguard sensitive corporate information and prevent unauthorized access.
- Developers and administrators should be able to customize the portal to create a distinct corporate look and feel, including graphics, banners, and channels.
- You want the professionals in your organization to have easy, quick, intelligent access to all the information they need in order to make good business decisions.

Examples of Corporate Portals

Before they adopted corporate portals, administrators at the Ames Department Store, with 479 store locations and 32,000 sales associates in 19 states, had to make 479 phone calls to communicate an urgent massage like a product recall. Similarly, they had to mail 32,000 mailers to its sales associate to inform them of new benefits. Likewise, Colliers, a Boston-based commercial real estate company, with offices in 52 countries and more than 250 offices, used to disseminate news, training, and research information to employees via push technologies such as fax and e-mail. Even a high-tech company like Hewlett Packard (HP) with more than 10,000 personnel used to rely on 4,700 intranet URLs, many containing duplicate and out-of-date information.

The solution to the problems experienced at Ames, Colliers, and HP was to use enterprise information or corporate portals (EIP), software developed by companies such as Plumtree and Hummingbird. Such software transforms the intranets and the paper-based information systems into the equivalent of a personalized start page. As observed by Sarah Roberts-Witt in her article *Digital Doorways* (2001), corporate portals are a doorway to all the information a company needs to function. A simple Web browser is used to access the knowledge management, collaboration tools, and other applications that the company makes available to the appropriate people to make intelligent business decisions.

Richard Dragan (2001) noted, in his article *Corporate Portals: More Than Just a Pretty Face*, that the major players in the corporate portal software design are Lotus (K-Station), Microsoft (Share Point Portal), Plumtree (Corporate Portal), Viador (E-Portal), Hummingbird, Epicentric (Foundation Sever), Oracle (9/As 2 portal), Sybase (Sybase Enterprise portal 2.5), IBM (WebSphere portal), CA (Clever Path), Sun (Sun One portal), and Yahoo (Portal Builder 4.0).

Obstacles

One of the major hurdles in the implementation of corporate portals is the problem associated with the integration of numerous data sources and applications. Using software connectors, gadgets, and portlets (small applications), however, one could include the popular software applications like outlook e-mail or other company applications.

Studies reveal that the portal development and implementation process can be expensive, ranging from $300,000 to $1 million. The benefits are worth every bit, however. These include increased productivity and low costs over time. It was revealed in a study that Johnson Controls, a $17-billion component manufacturer with 22,000 employees who use the portal, saved as much as an hour a day on information searches. This translates to significant cost savings. Corporations should encourage the use of their portals, however, because the benefits are not automatic. Furthermore, end users should be asked their opinion as to what they would like to see in the portal design. With this “buy in” should be more widespread.

For someone on a limited budget, hosted portals may be an option. For example, http://www.onepage.com is a software company that offers both hosted and licensed versions of the portal software for under $100,000. There are other less expensive options, such as PHP-Nuke, an open source software that can be used to develop a small business or personal portal. This software and other information can be obtained free of cost, from http://phpnuke.org and http://www.postnuke.org

Financial Portals

Writing in *ABA Banking Online*, Bill Orr (2000) predicted that, in the near future, the portal business will be restructured from the present general-purpose sites to a small number of generic portals such as shopping, financial services, sports, and entertainment. The banking industry plans to offer aggregated accounts to its customers. For example, consumers might prefer to shift their accounts to one primary service provider from two banks, say, a brokerage and a mutual fund firm.

Healthcare Portals

CGI Group (2001), in its white paper report titled *Corporate Portals for Healthcare Providers—A Strategic Solution to HIPPA Compliance*, writes that to comply with the Health Insurance Portability and Accountability Act of 1996 (HIPPA), many health care providers are integrating their patient accounting systems with multiple clinical systems and the user community, resulting in portal-based platform systems.

Voice Portals

NMS Communications (2002; http://www.nmss.com), in its white paper *Voice Portals: The Heat of the Voice Web* mentions that the confluence of the Internet and the telephone has resulted in a new technology named voice portal. Voice portals provide Internet content and electronic commerce services—from any phone, anywhere—using spoken commands. Voice portals allow callers to access e-mail, traffic information, local, national, international news, latest sports scores, stock prices, weather reports, and much more.

Advances in speech technology and the popularity of the mobile phones have helped the growth of voice portals tremendously. The Kesley Group predicts that by the year 2005, there will be more than 128 million voice portal users and revenues will reach $12.3 billion. The key players in the voice portal business include Tellme (http://www.Tellme.com), Bevocal (http://Bevocal.com), and Yahoo (http://phone.yahoo.com). Speech software vendors include Nuance (http://www.nuance.com) and SpeechWorks (http://www.speechworks.com).

Education Portals

Writing in the *Chronicle of Higher Education,* Florence Olsen stated that at many colleges and universities, particularly the larger ones, it is not uncommon to have hundreds of thousands of Web pages and other applications scattered on scores of Web servers. If all this information could be consolidated into a Web portal, then the organization, storage, and retrieval of information would be more efficient. For example, instead of sending an e-mail message to every college member about the college closing due to an inclement weather, the institution could just publish this information on the Web portal. Educational portals have many capabilities. For example, a student can log in to one place to do a variety of things—enroll for a class, send e-mail to instructors, look at grades, pay tuition—all in one place. In the preportal days, the student was required to log in to many servers.

Some of the education portals in use today allow users to access information on financial aid, student loans, transcripts, paying bills, and billing statements. In the world of higher education, Campus Pipeline, Blackboard, and PeopleSoft have been in use for the last 5 years or so. Some colleges are building their own portal software. Campus Pipeline, Plum Tree, and Epicentric are the major players in the development of education portals. Software tools from these companies integrate the information from various databases, file servers, and the Web pages that the colleges already have in place.

In addition to the commercial activities listed here, the open source approach is also in play for the development of education portals. Campus Pipeline, in collaboration with 10 universities, has designed new software called *luminis,* by starting out with open source software called uportal (http://www.uportal.com), developed by several hundred universities.

Other Resources

For additional resources for Internet navigation, review the following Web sites.

General

- http://www.learnthenet.com
- http://www.w3.org
- http://www.ipl.org (Internet Public Library)
- http://www.lii.org (Librarian's Index)

Encyclopedias

- http://techWeb.com
- http://whatis.com
- http://Webopedia.com

Directory Search

- http://www.yahoo.com
- http://www.looksmart.com
- http://www.galaxy.com

For more information on directory searches, visit http://www.notess.com/search/dir.

CONCLUSION

As discussed in the aforementioned pages, the Internet has many services—from simple e-mail to the powerful Portal—that can be harnessed to make our lives more productive. As Smith (1994, p. 2) observed, "Change in technology exerts a greater influence on societies and their processes than any other factor." The principle behind this school of thought, known as technological determinism is certainly true in the case of the Internet. It has changed the way of life in many in the industrialized countries. In the best circumstances, we will continue to seize the opportunities that the Internet can provide to enhance the quality of life for people around the world.

GLOSSARY

Bandwidth The range of frequencies that move along a communication link, such as a telephone line or cable television wire. The higher the bandwidth, the more the information—whether voice, video, or data—that can travel to a computer.

Bookmark A system used to keep track of uniform resource locator (URL) locations on the Web.

Browser A software program used to navigate the World Wide Web. Netscape Communicator, Microsoft Internet Explorer, Opera, and Lynx are examples of browsers.

Client–server A program with two components: software that runs on a personal computer (the client) and software that runs on a host computer (the server). A client software is Netscape Communicator; a server software is either Netscape Enterprise Web Server, Apache, or the Internet information server (IIS).

Domain name system (DNS) server A computer that identifies host computers on the Internet.

File transfer protocol (FTP) An Internet protocol for transferring files from one computer to another.

Helper application On the Web, a program that provides a function that a browser cannot perform; such as processing pictures, video, or other types of data or initiating a telnet session.

Hypertext markup language (HTML) A set of formatting commands or tags placed around text or pictures to create a Web page.

Hyperlink A button or highlighted text that activates a jump to another Web page.

Internet A network of computer networks that allows computers to exchange data, messages, and files with other computers.

Internet service provider (ISP) A business that offers Internet accounts that connect directly to the Internet. AOL is an example of an ISP.

Local area network (LAN) A network of computers set up in a limited geographic area such a building or a college campus.

Listserv A program that maintains lists of e-mail addresses. A listserv can be used as a bulletin board; messages mailed to it are sent to all subscribers.

Modem A device used to convert digital data into analog and vice versa.

Netiquette Rules to maintain civility in online discussions as well as special guidelines unique to the electronic nature of forum messages on the Internet.

Plug-in Software that extends the functionality of a Web browser. Examples include Real Audio, and Shockwave.

Portal A portal is a Web site that provides the entrance and access to structured information.

Protocol A set of rules used for computer communication. Examples include TCP/IP, FTP, and Telnet.

Router A device that connects any number of LANs.

Spamming Electronic junk mail or junk newsgroup postings. Some people define spam even more generally as any unsolicited messages.

Transmission control protocol (TCP) A part of the TCP/IP set of rules for sending data over a network.

Telnet A protocol that allows a user to connect one computer with another remote computer.

Uniform resource locator (URL) A host address on the Internet, for example, http://www.raritanval.edu

USENET The term that applies to a group of computers that exchange network news.

Wide area network (WAN) A computer network covering a wider geographic area than a LAN. For example, a corporate network covers many locations across the country.

WinZip A Windows-based program to zip (compress) and unzip (uncompress) files.

World Wide Web (WWW) A collection of HTML documents on the Web.

CROSS REFERENCES

See *Downloading from the Internet; E-mail and Instant Messaging; File Types; Internet Etiquette (Netiquette); Internet Literacy.*

REFERENCES

Ackermann, E., & Hartman, K. (2003). *Learning to use the Internet and World Wide Web.* Wilsonville, OR: Franklin, Beidele & Associates.

Andrews, J. (2001). *i-Net ± guide to Internet technologies.* Cambridge, MA: Thompson Learning.

CGI Group (n.d.). *Corporate portals for healthcare providers—a strategic solution to HIPPA compliance*, Retrieved April 20, 2003, from http://www.itpapers.com/cgi/PSummaxryIT.pl?paperid=38099&scid=189

Dragan, R. (2002, August 1). Corporate portals: More than just a pretty face. *PC Magazine.* Retrieved on April 20, 2003, from http://www.pcmag.com/article2/0,4149,369842,00.asp

Eckerson, W. (1999, July). Retrieved April 20, 2003, from http://www.oracle.com/oramag/oracle/99-Jul/index.html?49ind.html

Katz, R. N., & Associates (2002). *Web portals and higher education.* San Francisco, CA: Jossey-Bass.

Krill, P. (2001, April 30). Portals at forefront of Internet revolution. *InfoWorld,* 1–2.

Lsoft Web Site (n.d.). Retrieved April 28, 2002, from http://www.lsoft.com/manuals/1.8d/user/user.html#QS.1

McDonald, G. Portal profiles: Why Yahoo is #1. *PC World.* Retrieved April 20, 2003, from http://www.pcworld.com/news/article/0,aid,9902,00.asp

NMS Communications (n.d.). *Voice portals: The heat of the voice web.* Retrieved on April 20, 2003, from http://www.itpapers.com/cgi/PSummaryIT.pl?paperid=25957&scid=189

Olsen, F. (2002, August 9). The power of portals. *Chronicle of Higher Education, 48,* A32.

Orr, B. (n.d.). *ABA Banking online.* Retrieved September 24, 2002, from http://www.itpapers.com/cgi/PSummaryIT.pl?paperid=14297&scid=189

Rickard, J. (2000, October). Portals: Creating lifelong campus citizens. Portal Technology. 2000 Symposium: Portals in higher education. *Converge Magazine* (Suppl.), 3.

Roberts-Witt, S. (2001, June). Digital doorways. *PC Magazine.* Retrieved on April 20, 2003, from http://www.pcmag.com/article2/0,4149,98032,00.asp

Sagemaker (n.d.) *The benefits of an enterprise information portal* (White paper). Retrieved on April 20, 2003, from http://dis.shef.ac.uk/inf201/sagemaker.pdf,

Smith, M. R. (1994). Technological determinism in American culture. In M. R. Smith and L. Marx (Eds.), *Does technology drive history?* Cambridge, MA: MIT Press.

Smith, R. (2002). *Broadband Internet connection.* New York: Addison-Wesley.

3- and n-Tier Architectures (n.d.). From the Distributed Technologies Web site (Germany). Retrieved on April 20, 2003, from http://www.d-tec.ch/e/3tier.html

FURTHER READING

Search Engine showdown. Retrieved May 20, 2002, from http://www.searchengineshowdown.com/reviews

Franklin, C. (n.d.). How DSL works. How Stuff Works Web site. Retrieved May 20, 2002, http://www.howstuffworks.com/dsl.htm, Date of Access: May 20, 2002.

Collins, H. (2000). *Corporate portals: Revolutionizing information access to increase productivity and drive the bottom line.* New York: AMACOM.

Firestone, J. (2002). *Enterprise information portals: Business goals and information technology.* New York: Wiley.

Kirchoff, S., & Mendonca, S. (2000). *Instant advantage.com.* Englewood Cliffs, NJ: Prentice Hall.

Schneider, P., & Evans, J. (2002). *The Internet* (3rd ed.). Cambridge, MA: Thompson Learning.

Internet Relay Chat (IRC)

Paul L. Witt, *University of Texas at Arlington*

ONLINE COMMUNICATION USING INTERNET RELAY CHAT

Internet relay chat (IRC) is a worldwide system of real-time, text-based conferencing on the Internet. IRC was the first online chat system to achieve widespread adoption, and despite the increasing popularity of Web-based chat rooms and instant messaging systems, it continues to be the synchronous medium of choice for many thousands of Internet users. Numerous sites on the World Wide Web contain IRC tutorials, network information, and software downloads. Because IRC is based on a client-server networking model, users simply install client software on their local computer and access the Internet to create a new channel (chat group) or join an existing channel. Using established network protocols, messages travel from an individual's computer to an IRC server, then are relayed through other servers if necessary before being delivered to the designated recipient(s). IRC supports private, one-on-one conversations (analogous to real-time e-mail) as well as the simultaneous distribution of messages to all users in the channel (analogous to a real-time listserv or newsgroup). Interactions on IRC channels involve users from all over the world, take place in many different languages, and focus on a wide range of topics. Most interactions on IRC are casual and "chatty" in nature, but sometimes they involve specific personal, academic, or organizational communication objectives. At any given moment, tens of thousands of IRC users are conversing about sports, science, technology, academics, news, recreation, and entertainment, as well as engaging in spontaneous personal and social interaction.

Historical Development

Following the earliest experiments in real-time chat systems in the 1960s and 1970s, IRC was created in 1988 by Jarkko Oikarinen, an engineering student at the University of Oulu, Finland. By extending the features of an early Unix-based function called Talk, Oikarinen's innovative technology allowed multiple users at different geographical locations to enter designated channels and engage in synchronous online discussion. By 1990, the system had expanded to include 38 IRC servers located at universities in Scandinavia and the United States (Hamman, 1999). IRC gained notoriety as a strategic global communication system in 1991, when real-time messages were received from IRC users inside Kuwait during Iraqi occupation. In 1993, IRC enabled computer users in Moscow to communicate with outsiders during the uprising against President Yeltsin. Following the terrorist attacks in New York and Washington in September 2001, IRC networks became extremely active, initially to support communication with survivors and rescue workers, and later to allow users from around the globe to vent their feelings, rally support for victims, and engage in political and ideological debate. When other communication systems failed or became congested during these international crises, computer users with access to the Internet used IRC for up-to-the-minute interaction with external parties.

In recent years, rapid growth in the number of individuals with Internet access has contributed to rapid growth in the use of IRC. At the beginning of the 21st century, hundreds of thousands of users were communicating via IRC networks. Charalabidis (2000) reported that the total number of IRC users was increasing as much as 30% annually, and that the largest IRC networks had tens of thousands of channels and sometimes experienced more than 50,000 simultaneous connections. Although many Web users gravitate to popular, easy-to-use messaging systems, such as ICQ ("I Seek You") and Instant Messenger, IRC continues to appeal to the more technology-oriented specialist. In fact, "newbies" (newcomers) may find the IRC communication climate to be unfriendly or hard to understand at first. IRC commands and interfaces are not especially complicated, but casual Web users are often more accustomed to the very simple interfaces of browser-based Web chat systems. Furthermore, experienced users on some networks adopt an exclusivist attitude by refusing to communicate with newbies, or by using vague or highly specialized jargon that is unintelligible to outsiders. On these networks, newcomers are more likely to obtain advice and direction from channels devoted to IRC help.

STRUCTURE AND OPERATION

Servers, Networks, and Channels

As IRC works on the client-server networking model, the IRC server is the central point of communication through which all messages pass en route to individual client computers. In the simplest IRC network structure, the computer of each user runs a client program to connect to a common server. The server runs an application called an IRC daemon (ircd), which allows connections from individual users and relays messages from one client to another by passing them through the server. Some IRC chat systems consist of a simple network in which a single server manages connections and communications among all its clients. However, most IRC servers are connected to other IRC servers, thus opening the IRC system to potential communication among all users with connections to all the servers on the network. Originally, there was only one IRC network consisting of a few servers, but as IRC grew in popularity and complexity, it splintered into multiple networks. Today there are hundreds of IRC networks that are constantly evolving through expansion and redefinition. They range in size from very small, family-oriented chat networks, to such global networks as EFnet or the expansive IRCnet, which links well over 100 servers and supports thousands of chat groups, or channels (see Table 1). Channel names are fairly stable, but channel topics change regularly, and the announced topic frequently bears no relation to the conversation actually occurring on the channel.

To engage in Internet relay chat, users select an IRC network and connect to one of the available servers, using a nickname (e.g., SuperTig) rather than their real name. Using the /list command, users can consult a list of public channels currently available on that server. The channels listed as *#channelname* are global and can be accessed through any server on the network; channels listed as *&channelname* are available through that server alone. Some channels are designated "secret" and are not open to others. By entering the command /join #channelname, users join a channel and usually receive the MOTD (message of the day), containing some basic information from the server's database about the channel's current activity (topic, nicknames of current users, etc.). The new user is then ready to communicate with others by sending a message that will appear more-or-less synchronously on the screens of all channel participants (see Figure 1). Newbies can usually find help on the channel *#irchelp*, and other help channels may be listed in the MOTD. The command /leave or /part will terminate the connection to the channel, and the command /quit will terminate the connection to the network. In reality, connecting to an IRC server is not always simple, especially for newcomers. Connection may be prevented if the user enters an incorrect server name, chooses an unacceptable nickname, or has been *K-lined* (specifically banned from the server for some reason). Users who connect to the Internet through large, popular ISPs (Internet service providers) may find some channels closed to them, based on the stereotype that such subscribers are probably newbies and therefore unwelcome on certain channels.

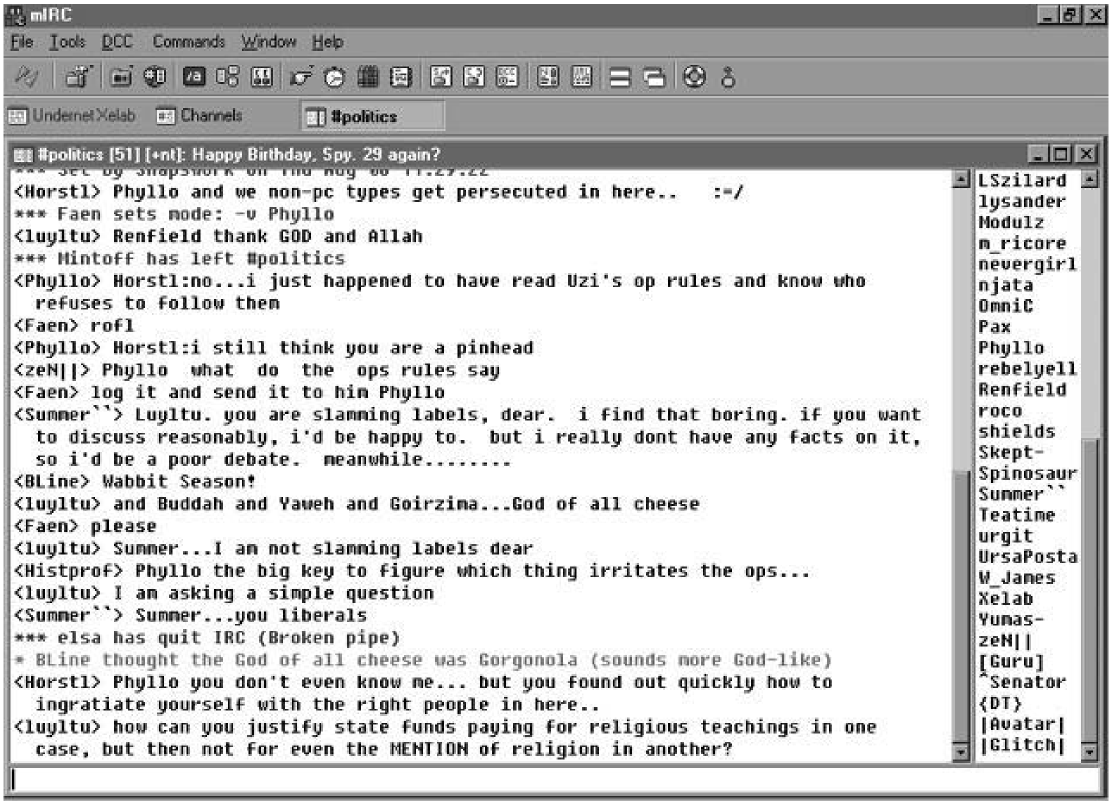
Table 1 Some of the Largest IRC Networks

Network	Number of Servers	Help Site and Server List
Efnet	40	http://efnet.org
IRCnet	>120	http://www.ircnet.com
DALnet	30	http://www.dal.net
Undernet	40	http://www.undernet.org
GalaxyNet	30	http://www.galaxynet.org
WebNet	15	http://www.webchat.org
NewNet	15	http://www.newnet.net
ChatNet	15	http://www.chatnet.com

With text-based IRC client software, such as ircII, commands are entered manually by the user in typical command-line format familiar to Unix users. With graphical interface client software, such as mIRC and Ircle, a click of the mouse on an icon generates the same result. Because the original IRC source code was not copyrighted, some networks have developed proprietary functions and features unique to their IRC system. The result is that all commands are not standardized across all IRC networks. Most of the basic commands are recognized by all IRC servers, however, and users who frequent a network typically learn specialized commands quickly through observation and by inquiring on a channel devoted to IRC help.

Although most users connect through IRC servers, the major IRC client programs also allow a direct client connection (DCC). Two clients who are connected to the same IRC network can initiate a direct connection between themselves. IP (Internet protocol) addresses are typically used to establish the DCC connection, and messages or shared files pass directly from one client to the other without being relayed by the IRC server. Many of the more experienced users prefer to use DCC for their one-on-one interactions and file transfers, because DCC frequently results in faster message exchange, a higher degree of privacy, and avoidance of network problems such as lags and netsplits.

IRC networks are not immune to problems of congestion or breakdown. Like all other packet-switching Internet communication systems, users sometimes experience delays in the transmission of electronic messages. This delay (known to IRC users as "lag") is irritating to all Internet users, but it is especially problematic in synchronous communication systems characterized by real-time interaction. A 10-s lag can dramatically change the timing and effectiveness of some types of online interaction, as threads of conversation become entangled and users appear to be "talking over" one another. Increased lag may be caused by heavy traffic on the Internet or IRC network, a malfunctioning server, flooding, or a denial-of-service attack (see below). Lag is sometimes related to a netsplit, which occurs when one or more IRC servers lose their connection to the network. Netsplits fracture the network and effectively create separate network fragments



#politics [51] [+nt]: Happy Birthday, Spy. 29 again?

```
*** Set by Snapswork on Thu Aug 08 11:27:22
<Horstl> Phyllo and we non-pc types get persecuted in here..   :=/
*** Faen sets mode: -v Phyllo
<luyltu> Renfield thank GOD and Allah
*** Mintoff has left #politics
<Phyllo> Horstl:no...i just happened to have read Uzi's op rules and know who
  refuses to follow them
<Faen> rofl
<Phyllo> Horstl:i still think you are a pinhead
<zeN||> Phyllo  what  do  the  ops rules say
<Faen> log it and send it to him Phyllo
<Summer``> Luyltu. you are slamming labels, dear.  i find that boring. if you want
  to discuss reasonably, i'd be happy to.  but i really dont have any facts on it,
  so i'd be a poor debate.  meanwhile........
<BLine> Wabbit Season!
<luyltu> and Buddah and Yaweh and Goirzima...God of all cheese
<Faen> please
<luyltu> Summer...I am not slamming labels dear
<Histprof> Phyllo the big key to figure which thing irritates the ops...
<luyltu> I am asking a simple question
<Summer``> Summer...you liberals
*** elsa has quit IRC (Broken pipe)
* BLine thought the God of all cheese was Gorgonola (sounds more God-like)
<Horstl> Phyllo you don't even know me... but you found out quickly how to
  ingratiate yourself with the right people in here..
<luyltu> how can you justify state funds paying for religious teachings in one
  case, but then not for even the MENTION of religion in another?
```

```
LSzilard
lysander
Modulz
m_ricore
nevergirl
njata
OmniC
Pax
Phyllo
rebelyell
Renfield
roco
shields
Skept-
Spinosaur
Summer``
Teatime
urgit
UrsaPosta
W_James
Xelab
Yumas-
zeN||
[Guru]
^Senator
{DT}
|Avatar|
|Glitch|
```

Figure 1: Screenshot of online interaction using mIRC.

that may or may not continue to operate separately. Netsplits often result in a buildup of packets queued for delivery to servers that are no longer available. After a netsplit, servers sometimes reconnect to the network and are able to update the server with activity that occurred during the netsplit. Sometimes a netsplit results when an IRC operator revises the network's connection to other servers in order to achieve a more efficient or more desirable network configuration.

IRC Client Software

In order to engage in online communication using IRC, users need access to the Internet through an Internet service provider (ISP) or direct network connection. They must also install a communications software program, called an IRC client. The client software serves as the interface between the individual computer (the client) and the IRC server (serving the needs of multiple clients). The IRC client processes the commands that are typed or clicked by the user and then employs network protocols, such as TCP/IP (transfer control protocol/Internet protocol) to transmit them to the server.

IRC supports communication across different operating platforms, enabling the flow of messages between PC and Macintosh computers, as well as UNIX-based computing systems. IRC clients for these and other operating systems are readily available, and free downloads or demonstration copies can be found on most of the major FTP (file transfer protocol) sites on the Internet. Some IRC clients are available as shareware, meaning that after a free trial period, if users wish to continue using it, they are asked to purchase the software or pay a modest registration fee. Generally, the cost of IRC client software is quite reasonable.

Text-Based IRC Clients

Early IRC users were almost exclusively Unix based, and the majority of today's IRC servers are Unix servers. Individual computer users who use a terminal emulation program to establish a dial-in connection to an organization for Internet access effectively turn their computer into a terminal, making use of a shell account on the Unix system. The most common IRC client for Unix and related systems (such as shell accounts and Linux) is ircII. This full-featured, text-based client was the first widely used IRC client and still serves as a reference standard for essential commands and functions. More advanced users sometimes use ircII's scripting language to create scripts that streamline complex processes, generate customized output, or perform certain functions automatically when prescribed conditions occur.

Graphical Interface IRC Clients

The majority of online communicators connect to the Internet through Windows-based PCs or Macintosh computers. IRC clients have been developed for both platforms, offering the point-and-click ease of use characteristic of graphical user interfaces (GUIs). The major GUI clients receive ongoing attention from their developers, and excellent support is available for learners. The client of choice for many PC users employing Windows operating systems is mIRC, a powerful, intuitive, and reliable application. Because PC users outnumber those on Unix and Macintosh platforms, mIRC is probably the most widely used IRC client on the Internet today. Easy-to-use menus and buttons, customizable functions, and extensive user support contribute to the popularity of mIRC (see Figure 1). The most widely used IRC client for Macintosh computers is Ircle. A stable, full-featured IRC client, Ircle provides buttons, menus, and icons to facilitate common IRC functions. Ircle users may also employ AppleScript to modify the client to their custom specifications. Good user support and innovative features place IRC interactions within easy reach, even for beginning Ircle users.

IRCops, Chanops, and Users

The IRC system hierarchy of networks, servers, channels, and clients is reflected in an administrative hierarchy of people who oversee each level of the system. End users who direct their IRC client to connect to a network should be aware of the scope of authority and control exerted by IRC operators and channel operators. Newbies sometimes embarrass themselves and irritate others by inappropriately posing questions or filing complaints with the person they perceive to be in charge of a channel or server.

IRCops

IRC operators generally oversee the operation of an IRC server, maintain routing functions in relation to other servers on the network, modify server settings, and monitor general activity on the server. Sometimes the original server administrator fulfills the role of IRCop, and sometimes the administrator selects one or more experienced and qualified users to serve as IRCop. Servers typically have more than one operator, in order to ensure constant observation of activity on the server. In the event of a server failure or network fault, IRCops work to restore network connections. They also enforce the policies and procedures of the server they oversee. If a user violates policy or misbehaves in some way, the IRCop usually issues a warning. Users who are reprimanded by an IRCop are sometimes unaware of the power the operator has to control activity on the server. For example, an IRCop can issue a /kill command that effectively bans the offending user from further access to the server and, in some cases, the entire network. Determining the identity of a server's IRCops is not always easy. When they are listed in the MOTD (message of the day) or active user lists, their "user mask" (nickname plus host name) is usually marked with an asterisk (*Stinkbug!skb220@rhl.net). The command /stats o may be used to identify active operators, and some IRC clients support the command /who -oper. Many IRC users never have direct communication with an IRC operator unless their activity on the server violates accepted practice and attracts the operator's attention. IRC operators do not exert control over channel activities; they leave those functions to the channel operators.

Chanops

Each IRC channel (chat group) is initiated or overseen by a channel operator, or chanop. Channel operators exert management and control over the channel's topic, mode, and settings, and they regulate how many users and specifically which users are allowed to interact on the channel. If users misbehave, become obnoxious, or irritate the channel operator, most chanops do not hesitate to ban them from the channel. Any user may become a chanop by creating a new channel or by being given operator status by a chanop on an existing channel. This status is valid only for the duration of the current IRC session, and when chanops leave the channel, they ordinarily lose their operator status. However, if operators are registered as super-ops with an IRC channel service, they can effectively maintain control over a channel even during their absence. Many IRC channels have several operators active at any given time, but occasionally a channel may function without an operator at all. Channel operators are usually identified with an @ sign before their nickname (@SuperTig). Most IRC clients support the command /who -chops as a means of identifying channel operators.

Chanops can remove users from a channel and ban them from future access. On password-protected channels, they can invite users to join and can issue them the key to enter. On moderated channels, chanops can give selected users a voice that allows them to post messages to the channel while other users can only receive messages. Channel operators can grant chanop status to other users, and they can remove chanop status from other operators. Any IRC user who creates a new channel with the /join #channelname command automatically becomes the operator for that channel. When a new channel is created, some basic channel modes and settings should immediately be set to establish guidelines for the activity allowed by users and operators on the channel.

Common User Commands

Table 2 contains a sampling of user commands commonly employed across different IRC clients and servers. Users of text-based clients enter these commands manually, whereas users of GUI clients usually click on icons or drop-down menus to engage the function, and the client automatically generates the command. Some commands contain specific parameters or are qualified by "flags" that set the limits of the command output. Table 3 contains the most common commands used by channel operators.

SOCIAL FACTORS

IRC and Online Community

As with all online communication systems, IRC is more than a network of computers relaying electronic files among themselves. Fundamentally, IRC systems are composed of humans who utilize the computer network to exchange information or conversation with others. The people who employ IRC for human communication reflect the demographic diversity of general Internet users, with a few notable exceptions. As a somewhat specialized online community, IRC has drawn those who are more technically proficient, whose computer skills enable them

Table 2 Basic User Commands

Command	What Command Does
/join #channelname	Joins the user to a global channel on the network, unless the user is banned or the channel is restricted (invite-only, secret, etc.)
/join#channel1,#channel2	Allows the user to simultaneously participate in both of the channels listed
/mode MyNickname +i	Sets the mode of the user "MyNickname" to invisible, enhancing privacy and security
/list	Produces a list of channels on the server; often used with various flags and parameters
/set hold_mode on	Prevents a long list from scrolling beyond the monitor's limits
/names	Shows a list of nicknames of users currently on the network (potentially a huge listing, so always use parameters)
/names #channelname	Shows a list of nicknames of users currently on the channel (also very large on some channels, so use parameters)
/who #channelname	Lists nicknames, usermasks, and other information about current users on a channel
/whois NewName	Produces basic information from the server's database about the user "NewName"
/msg NewName	Sends a private message to the user NewName; messages to all users are entered without the/msg command
/dcc chat NewName	Requests Direct Client Connection with the user "NewName;" if NewName accepts, a direct private chat will bypass the IRC server (may be more private, less lag)
/nick othername	Changes a user's nickname to the one specified
/ignore badguy	Blocks or ignores messages from user called "badguy"
/part or/leave	Signs a user off of a channel; may be followed by a parting comment
/quit	Terminates connection to the network; may include a brief comment

to easily perform such tasks as the following: downloading and installing software from an FTP site; learning intuitively the features and functions of new software applications; using line commands, scripting, or hand coding to accomplish tasks; adapting to the specific styles and conventions of online communication forums; and maintaining sufficient safeguards to protect his or her personal privacy and system security.

Many individuals "meet" regularly on IRC channels, where they develop close relationships and a true sense of community. On some channels, the basis for commonality is the agreed-upon topic (such as advanced applications of certain software), and participants willingly share techniques and resources in mutual support of one another. Other channels are predominantly social in nature, marked by spontaneous casual conversation. In both types of IRC communities, significant one-on-one relationships often develop. There are many reports of IRC users who decided to move their relationships offline and meet face to face, and not a few have become business partners or marriage partners (e.g., Hoff & Hoff, n.d.). Of course, the majority of IRC relationships are restricted to the online context, but they constitute legitimate relationships that fill meaningful social roles for many users. Friends made on IRC are not only "virtual" friends but also "real" friends.

Table 3 Basic Channel Operator Commands

Command	What Command Does
/join #NewChannel	Creates a new channel with the specified name
/mode #NewChannel +o|l|m|v|t	Sets the channel modes to Operator (chanop status can be assigned or removed), Limit (limits the number of users allowed at one time), Moderated (all users receive, but only some can post messages), Voice (grants selected users the ability to post messages), and Topic (set by operators only); several other modes are available
/mode #NewChannel	Reveals all the current mode settings for the channel
/mode #NewChannel +o mybuddy	Gives the user "mybuddy" operator status
/mode #NewChannel -o mybuddy	Removes chanop status from the user "mybuddy"
/mode #NewChannel +l 20	Limits the number of users on NewChannel to 20
/mode #NewChannel -l	Removes the limit on number of users
/mode #NewChannel +v angelbaby	Allows the user "angelbaby" to post messages to NewChannel
/topic #NewChannel IRC clients	Changes channel topic to "IRC clients"
/kick #NewChannel satanson	Kicks the user "satanson" off the channel
/invite mybuddy #NewChannel	Invites the user ″mybuddy" to join NewChannel

Communication researchers study the social, relational, and communicative dynamics of online communication forums such as IRC. Overall, these studies indicate that computer-mediated communication systems do support the development of meaningful relationships, effective decision-making groups, and a legitimate sense of community (Finholt, Sproull, & Kiesler, 1990; Walther, 1996, 1997). Nevertheless, research has also identified some important differences between face-to-face and online communication. For example, in face-to-face interactions, people normally communicate interpersonal liking through such nonverbal cues as eye contact, smiles, proximity, touch, open body positioning, and vocal expressiveness (Richmond & McCroskey, 1999). Because the usual nonverbal behaviors are absent in the online context, a greater responsibility is placed on verbal messages for feedback, affect, and nuance of meaning. Some IRC users attempt to replace nonverbal messages by inserting affect cues <grin>, demonstrative punctuation (no!!!!!!!!!), or emoticons, such as the smiley face :-) and its many variant forms. The absence of nonverbal communication cues led some early researchers to judge online communication as incapable of conveying subtle relational messages and therefore an unsatisfactory medium for meaningful interpersonal exchange (Daft & Lengel, 1984). However, in the intervening years, online skills have improved and online communication norms have evolved, with the result that 21st-century communicators routinely use IRC to engage in effective and meaningful interaction of all types.

IRC and Relationship Development

As IRC users employ new communication strategies, they sometimes find relationship opportunities in the online community that do not exist for them face to face. For example, research has shown that the decision to initiate face-to-face interaction with someone may be directly or indirectly affected by the other person's physical appearance, nonverbal behaviors, vocal attributes, age, gender, or ethnicity (Berscheid & Walster, 1978). It may seem unkind or unfair, but people often choose whether or not to interact with others on the basis of these traits. When people go online to engage in conversation on IRC, however, these physical characteristics are obviously not visible. Many users choose a nickname that is not gender specific, and many avoid asking or revealing details of personal identity. Even when personal characteristics are discussed online, they are virtually impossible to verify and seemingly have little or no impact on IRC interactions anyway. The result is that on IRC, unattractive people can relate as equals with the beautiful and handsome; those with vocal or speech problems can engage in fluent expression; those who painfully experience bias or discrimination in face-to-face encounters may confidently enter chat groups as social equals, perhaps for the first time in their lives.

Therefore, this new communication environment may actually enhance the potential for long-term relationship development by expanding the ways interpersonal relationships develop. An example is seen in the widely researched "matching hypothesis," which theorizes that people usually date and marry someone with approximately the same physical attributes as themselves (e.g., both are attractive or both are plain in appearance). By contrast, couples who first meet on IRC may vary considerably in physical attractiveness. Online relationships can deepen over time through self-disclosure, empathic "listening," and progressively more intimate communication. When the relationships are moved offline for face-to-face meetings, differences in physical appearance seem relatively unimportant in comparison to the many points of commonality already discovered. In such cases, IRC may provide a more just and balanced communication context than traditional meeting places.

Anonymity and Virtual Identity

The relative anonymity that enhances relationship development on IRC might also introduce an element of deception. Obviously, users may project a virtual identity very different from their true identity. Their intent in doing so might be harmless and recreational, or it might be deceptive and dangerous. IRC, as do all online communication systems, occasionally draws users who are malicious, misleading, or mentally unstable. For example, pedophiles frequently pose as youngsters in forums frequented by children, hoping that the child will trust them with information or engage in conversation with them. According to accounts in the popular press, some long-term relationships that seemed healthy and secure online turned out to be dangerous when the couple met face to face, because one of the parties had been projecting a virtual identity to conceal unwholesome motives.

SECURITY AND LEGAL ISSUES

Hate and Harassment on IRC

The anonymous nature of online communication systems such as IRC may contribute to the use of aggressive or inflammatory verbal abuse in the form of harassment or hate speech. Most IRC interactions involve the use of nicknames only, and users rarely reveal their true identities. This anonymity may contribute to a perception of low accountability, with the result that some IRC users may verbally lash out against individuals or groups in ways that are offensive and illegal. For American citizens, First Amendment (free speech) privileges extend to IRC interactions, but so do criminal laws preventing hate speech, inciting violence, and stalking. Cyber-stalking, which is expressly forbidden by U.S. federal law, occurs when a user annoys, abuses, threatens, or repeatedly initiates unwanted communication with another. Responsible channel operators seek to eliminate such activity on IRC by banning the offending user and, if necessary, reporting them to legal authorities. Of course, many IRC participants are not U.S. citizens and not constrained by American laws, pointing up the ethical and legal complexities of global communication networks.

Children and IRC

Although a significant percentage of overall Internet users are under the age of 18, a relatively small percentage of IRC users are minors. In truth, IRC is not especially suited to the needs and concerns of children. The more widely known Web-based chat systems maintain terms of service that restrict certain activities and topics of discussion. Offenders in these chat rooms can be quickly and effectively removed by administrators. IRC channels, on the other hand, are generally not restricted, and people may talk about anything they choose. It is not infrequent to encounter adult conversations and offensive language on IRC that would be unsuitable for children. On moderated channels, chanops can always remove a user's "voice" and disable the offender's ability to contribute to further discourse, but such censorial functions are not characteristic of most IRC channels. IRC is perceived by many to be spontaneous, freewheeling, unrestricted, and unpredictable—a kind of "wild west" communication context.

In carefully controlled and appropriate IRC channels, children may benefit from online conversation with other children from around the world, with teachers and classmates, and with extended family members or parents who are traveling. Unsupervised, however, children on IRC may encounter such risks as cyber-stalking, possibly leading to physical stalking, if the child reveals personal information; offensive and harmful conversations with unwholesome characters such as pedophiles; and exposure to sexual content, viruses, or other harmful material, if the child accepts a file transfer from an unknown IRC user. Numerous laws and regulations exist to protect minors on the Internet, and responsible IRC users should immediately report any suspicious activity involving a minor on IRC.

Scripts and Viruses

Informed IRC users are aware of the many risks involved with file transfers, through both hard media and the Internet. When users install new software, download Web pages, or receive files such as disks or CDs from external sources through networking or hard media, some degree of security risk is introduced. It is a rare occurrence, but even respected commercial software has been distributed with inadvertent flaws that damage systems. More common is the intentional or unintentional passing of a virus or harmful script from user to user through various Internet connections, including IRC. Responsible users minimize these risks by maintaining awareness of current security threats, practicing discretion when networking, and using updated virus detection software.

Most viruses are in .exe or .ini files, but they can also be concealed in zipped files or embedded in completely unrelated file types, such as .jpg or .gif files. When launched by an unsuspecting user, viruses can cause irreparable harm to individual computers and potentially spread throughout entire systems. Some viruses are not destructive but merely irritating, such as those with pop-up messages of a humorous or salacious nature. Others may disable or destroy hard drives, data files, operating systems, and peripheral components or create havoc throughout entire networks.

People who share word processing files, script files, graphics, or other file types on IRC may potentially be spreading viruses, trojans, or worms. Trojans (or Trojan horses) are spread by users who transfer seemingly benign files containing harmful concealed code. Worms are viruses that propagate themselves and migrate from one application to another within a computer, and from one computer to another within a network. Fortunately, virus protection software is inexpensive and readily available on all operating platforms. IRC users should load virus protection software, keep it current through frequent data updates as new viruses appear, and carefully scan every file that is received or shared by IRC transfer.

Scripts represent another source of external control and increased security risk. Scripts are files containing macros (or miniprograms) that enable an application to combine or automate various functions. IRC users have created scripts for such tasks as completing nicknames from typed initials, automatically performing ping operations, generating automated "away" responses, and customizing the output of IRC interactions. Some early IRC users enhanced the function of their IRC clients by writing scripts that have subsequently been incorporated into more recent versions of the software. Within the IRC community, techniques of script writing constitute a frequent topic of conversation, and the sharing of scripts among users is relatively common. A number of Web sites or FTP download sites serve as archives or clearing houses for the open sharing of scripts and script paks (sets of coordinated scripts) for various IRC clients. When an IRC user downloads a script file from an archive or another user, however, there is always risk that the file may contain concealed or deceptive code that may launch unexpected or harmful functions in the receiver's system.

Most IRC clients include options to limit or disable the automatic or unintentional running of scripts, just as the major Web browsers include settings to control scripting functions encountered on the World Wide Web. Because harmful functions can be inserted into otherwise legitimate scripts, cautious IRC users run scripts only after proofreading every line of the code. Safe practice also calls for the careful screening of IRC users before accepting from them any files of any type. Common practice states that files should never be accepted from an unknown person, and file transfers from known parties should always be scanned for viruses before being opened.

Mischief and Sabotage on IRC

Many people use IRC safely and never experience viruses, takeovers, or threats to their security. It is not uncommon, however, to encounter less-serious offenses allegedly launched by mischievous youngsters known as "script kiddies." Anyone who has been on IRC very long has probably observed attempts by an aggressive user to take charge of a channel or disrupt the flow of messages on a channel or network. "Channel taking" may be accomplished if a user is granted chanop status or deceptively uses an operator's nickname, then "de-ops" the other operators and indiscriminately bans users, kicks them off the channel, or silences their voices by turning the channel into a moderated one. Sometimes malicious users simultaneously send multiple files to another user, known as "flooding." If practiced on a large scale, flooding can significantly increase lag on a channel or entire network. Most IRC clients include optional settings that limit flooding to that client.

A more serious form of sabotage exists in denial of service (DoS) attacks, harmful data packets that sneak through loopholes in an operating system's security and launch erroneous messages from or to the client. DoS attacks, sometimes called "nukes," are against the law and can cause the host to disconnect from the Internet, freeze, or crash entirely. IRC can be exploited to facilitate a DoS attack by sending scripts called "bots" to unsuspecting users. Bots use vulnerabilities in the operating system to initiate the attack simultaneously from multiple computers.

Because internal security is especially critical to certain types of organizations, some companies or institutions have restricted the use of IRC and other Internet functions through the use of firewalls, prohibiting any computer in their network from connecting to an IRC server. For any IRC user, whether institutional or private, the best defense against mischief and sabotage is to refuse to accept file transfers or DCC chat requests from unknown persons, to carefully peruse every line of script code before launching it, and to maintain updated security files for operating systems, IRC clients, and virus protection software.

THE FUTURE OF IRC

To date, many people view real-time online communication systems such as Internet relay chat as little more than a novelty. In fact, use of the word "chat" in serious organizational contexts may cause some institutional leaders or educators to grimace and assume that the medium only promotes idle chitchat. As a generic communication medium, however, IRC and other chat systems possess great potential to assist modern communicators in reaching important organizational and personal communication goals. Future developments in IRC software and the Internet infrastructure may enable everyday use of video and voice chat, for example, that could dramatically alter the nature of the IRC communication medium.

Organizational Applications

Businesses, institutions, and organizations of all kinds place a high value on frequent communication among employees, suppliers, clients, customers, and/or members. IRC can be employed within an organization to support real-time collaboration among employees at different geographic locations, increasing productivity while keeping down long-distance communication costs. Customer service systems that now offer e-mail support can be expanded to include real-time exchanges via IRC. Employee teams, whether large or small, can use IRC for group planning, brainstorming, or coordinating joint projects. The ability to use IRC for file transfers can support the sharing of internal documents, or the distribution of such external information as product brochures, technical specifications, and press releases. As mentioned above, however, a growing number of organizations have responded to concerns about security or privacy issues by installing firewalls that prevent the use of IRC and other external file-sharing applications on their networks. Some universities, for example, have discontinued the use of IRC or created tighter accessibility restrictions to protect their systems from outside interference. Of course, an organization could choose to maintain an internal IRC server with no connections to outside networks, or to use moderated or secret channels on existing networks.

Educational Applications

The same collaborative benefits available to organizations apply to the use of IRC in educational settings. With the rise of distance learning, distributed education, Web-based training, and e-learning, teachers and students are enthusiastically embracing new communication technologies to support interaction across distances. Furthermore, teachers and trainers who provide classroom courses are finding that chat systems can serve to link the course participants in dialog outside of class. Some college professors, for example, require students to engage in a certain number of exchanges on the class listserv, or to participate in a certain amount of time in class chat groups, as a means of increasing student dialog or gaining facility in the use of new technologies. As the number of World Wide Web users continues to grow, so also does the demand for visual communication in online contexts. The introduction of new Java-based IRC clients will presumably open the way for IRC networks to develop multimedia or highly interactive capabilities, which may very well signal a new era for IRC.

International Applications

A specific educational and practical application that is enhanced by IRC is the learning of different languages. As a global communications medium, IRC has drawn users from virtually every country of the world. By connecting to an IRC channel in a second language, learners can observe native speakers using everyday language, a valuable learning tool. Because IRC easily supports private, one-on-one conversations, communicators from two different language groups can agree to engage in basic conversation, adjusting the topic and/or linguistic level to match the abilities of the learner as they would in face-to-face conversation. A great advantage to the learner is the near-synchronous nature of IRC—the ability to pause and reflect, or even quickly look up an unknown word in a dictionary—before responding. IRC has great potential to enable multilingual and multicultural interaction between individuals or groups of individuals who wish to engage in global communication. For example, children in two classrooms in two different countries can interact as individuals or groups on an IRC channel created for a joint class project. Anyone who values global communication as a means of enhancing tolerance and understanding can embrace IRC as a convenient and readily available global communication system. For those who travel, study abroad, or live far from their home country, IRC provides a reliable and inexpensive means of keeping in touch with home or office.

Emergency Applications

In times of local, national, or international crisis, public communication systems are often overloaded or out of service. In September 2001, for example, in the hours following the terrorist attacks on New York and Washington, telephone switching systems became clogged as tens of thousands of people tried to connect with friends and family in the area. In such times of crisis, computer users with an Internet connection can use IRC to provide real-time contact when other communication systems are unavailable.

In the years ahead, new communication technologies will emerge to support increasing demands for interpersonal and mass communication. Tomorrow's global society will depend to an increasing degree on terrestrial wireless, satellite, and fiberoptic communication systems. Internet relay chat, as a reliable means of real-time online conferencing, will continue to evolve into a medium capable of serving the communication needs of the 21st century.

GLOSSARY

Ban Status of a user forbidden by a chanop to enter a channel.

Channel A virtual chat group in which users interact, discussing specific topics or engaging in general conversation.

Chanops Channel operators; they monitor and maintain a channel by setting the parameters and inviting or expelling users.

Client Software run by the local computer of an IRC participant, converting the user's actions into commands recognized by an IRC server.

Daemon (IRC daemon, or ircd) Software run by a server connected to the Internet, accepting connection requests from clients and relaying messages among users.

Direct client connection (DCC) Connection established between two users that allows them to bypass the network for increased privacy and faster connection.

Flooding Simultaneously sending many messages to a single user, thus "flooding" a channel or an entire network with unwanted messages, probably increasing lag.

Internet Relay Chat (IRC) Real-time, text-based communication medium that employs a network of servers to connect geographically separated users.

IRCops IRC operators who monitor and maintain IRC servers; they generally have very little to do with the operation of channels.

Lag A delay in the transmission of messages between users or between a user and a server.

MOTD (Message of the day) An automated confirmation message containing certain information and practices relating to a channel, its activities, and its current participants.

Netsplit Disruption in the connection between servers that results in the fracturing of the network into subdivided parts and in the disconnection of some users.

Newbie A newcomer to IRC.

Nickname An invented name that uniquely identifies the user on the network.

Scripts Files containing code that enables an application to combine or automate various functions.

CROSS REFERENCES

See *Distance Learning (Virtual Learning); Internet Etiquette (Netiquette); Legal, Social and Ethical Issues; Online Communities; Privacy Law.*

REFERENCES

Berscheid, E., & Walster, E. H. (1978). *Interpersonal attraction*. Reading, MA: Addison-Wesley.

Caraballo, D., & Lo, J. (2000). *The IRC prelude*. Retrieved August 12, 2002, from http://www.irchelp.org/irchelp/new2irc.html

Charalabidis, A. (2000). *The book of IRC*. San Francisco: No Starch Press.

Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organization design. In B. M. Staw &L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 6, pp. 191–233). Greenwich, CT: JAI.

Finholt, T., Sproull, L., & Kiesler, S. (1990). Communication and performance in ad hoc task groups. In J. Galagher, R. E. Kraut, & C. Edigo (Eds.), *Intellectual teamwork: Social and technological foundations of cooperative work* (pp. 291–325). Hillsdale, NJ: Erlbaum.

Hamman, R. (1999). *History of the Internet*. Retrieved March 29, 2002, from http://www.socio.demon.co.uk/history. html

Hoff, M., & Hoff, J. *Chatting on the net*. Retrieved August 12, 2002, from http://www.newircusers.com

Stewart, W. (1997). *Internet Relay Chat.* Retrieved August 12, 2002, from http://www.livinginternet.com/r/r.htm

Lo, J. (Ed.). (1994). *Internet Relay Chat help archive*. Retrieved August 12, 2002, from http://www.irchelp.org

Reed, D., & Oikarinen, J. (1993). *Internet Relay Chat protocol*. Retrieved August 12, 2002, from ftp://nic.merit.edu/documents/rfc/rfc1459.txt

Richmond, V. P., & McCroskey, J. C. (1999). *Nonverbal behavior in interpersonal relations* (4th ed.). Boston: Allyn and Bacon.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23,* 3–43.

Walther, J. B. (1997). Group and interpersonal effects in international computer-mediated collaboration. *Human Communication Research, 23,* 342–369.

Internet Security Standards

Raymond R. Panko, *University of Hawaii at Manoa*

INTRODUCTION

When the Internet was created, security was left out of its Transmission Control Protocol/Internet Protocol (TCP/IP) standards. At the time, the crude state of security knowledge may have made this omission necessary. Today, however, security expertise is more mature. In addition, the broad presence of security threats on the Internet means that security must be addressed deliberately and aggressively in Internet standards. This chapter describes three ways to add standards-based security to the Internet. The first is for users to add standards-based security to specific dialogues, that is, to two-way conversations between pairs of communicating processes. Today this can be done largely through the use of virtual private networks (VPNs) using Point-to-Point Tunneling Protocol, IPsec (Internet Protocol security), secure socket layer, and other "add-on" security protocols. This "security overlay" approach is not as desirable as good general Internet security standards, but it solves some specific current needs of users. The second approach is to add more security to central Internet standards. In addition to the "core four" standards (IP, TCP, User Datagram Protocol, and Internet Control Message Protocol), these central Internet standards include supervisory standards (such as the Domain Name System, Dynamic Host Communication Protocol, Simple Network Management Protocol, Lightweight Directory Access Protocol, and routing protocols), and application standards (such as Simple Mail Transfer Protocol, Hypertext Transfer Protocol, and File Transfer Protocol). As this chapter describes, a strong effort to retrofitting existing Internet standards to add security is underway at the Internet Engineering Task Force (IETF) but is far from complete. The third way to add security is to take a comprehensive look at the Internet to decide how security *should* work on the Internet. This could lead to anything from the creation of *multiple Internets* with different classes of security for different types of users to the creation of international policies to make forensics more effective when there are attacks.

This chapter looks at these three broad approaches in the order just presented. It does not include a discussion of security for single network standards, such as Ethernet, Asynchronous Transfer Mode (ATM), Frame Relay, and wireless local area networks. Rather, it focuses on internet standards, which link together multiple networks. The chapter focuses more specifically on TCP/IP internetworking, which is used on the Internet, rather than on IPX/SPX and other forms of internetworking. More formally, it focuses on security in TCP/IP standards at the Internet, transport, and application layers.

SECURITY THREATS AND DEFENSES

Before looking at security in Internet standards, it is necessary to have an understanding of the major security attacks and defenses.

Penetration Attacks

One set of threats against which networks must guard is attackers attempting to hack systems (access them intentionally without authorization or in excess of authorization) or attempting to conduct denial-of-service attacks against them by crashing or slowing them to the point of being useless to their legitimate users.

Attacks on Dialogues

Second, when two parties communicate, they normally engage in a dialogue in which multiple messages are sent in each direction. These dialogues must be secure against attackers who may wish to read, delete, alter, add, or replay messages. In general, dialogue security has been the main focus of IETF efforts to add security to Internet (TCP/IP) standards. Hacking and denial-of-service are sometimes addressed, but this is comparatively uncommon. Consequently, I focus on dialogue security here.

Confidentiality

Confidentiality means that eavesdroppers who intercept messages en route over the Internet cannot read them. Encryption is the chief tool against unauthorized reading en route.

Authentication and Integrity

Authentication means that the receiver of messages can verify the identity of the sender to ensure that the sender is not an impostor. There are two forms of authentication. In *initial authentication*, one side proves its identity to the other side at or near the beginning of a dialogue. In *message-by-message authentication*, the sender continues to identify itself with every message, much as parties exchanging letters sign each letter. Message-by-message authentication is crucial to good security.

Integrity means that if a message has been captured and changed en route, then the receiver will be able to detect that a change has occurred. All forms of message-by-message authentication provide message integrity as a by-product. If a message is changed during transmission, either deliberately or through transmission errors, authentication will fail, and the receiver will discard the message.

Message-by-Message Authentication

Impersonation is a serious problem because many Internet standards are supervisory standards that govern how devices on the Internet function. If an attacker can trick a network management program system or other supervisory system into accepting false information, the resultant damage can be widespread. There are two common message-by-message authentication methods: HMACs and digital signatures. Both are described in more detail in the chapter Digital Signatures and Electronic Signatures.

Digital Signatures

In digital signatures, the sender hashes the message to be sent. This creates a message digest. The message digest is then signed (encrypted) with the sender's private key, which only the sender should know. This produces the digital signature, which the sender appends to the message before transmission.

The receiver recomputes the message digest in two ways. First, as the sender did, the receiver first hashes the message to produce the message digest. Second, it decrypts the digital signature with the true party's public key. This should also produce the message digest—if the sender is the true party. Otherwise, the message digests will not match, and the message will be rejected. In addition, digital signatures produce message integrity; if an attacker changes the message en route, the message digests will not match and the message will be rejected.

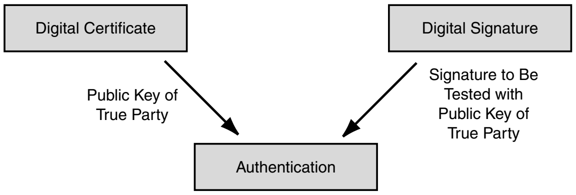


Figure 1: The need for digital certificates with digital signatures.

Digital Certificates

Note that digital signatures require the digital signature to be tested with the *true party's* public key—not the public key of the sender (who may be an impostor). The definitive way to get public keys of a true party is to get a digital certificate from a trusted certificate authority (CA). This digital certificate has the name of the true party and the true party's public key. It also has a digital signature signed by the CA's private key so that the certificate cannot be changed without this change being detected.

One problem with digital certificates is that certificate authorities are not regulated in most countries, leading to questions about their trustworthiness. Although European countries are planning to create regulated CAs, the United States and most other countries are trusting market forces to handle certificate authorities.

Another problem with digital certificates is that a CAs must create complex public key infrastructures to create public key–private key pairs, distribute private keys and digital certificates, and allow users to download certificate revocation lists (CRLs), which are lists of ID numbers for certificates that were revoked before their valid period ends.

Is it possible to distribute public keys without using certificate authorities? Pretty Good Privacy (PGP) offers one way to do so. Each user has a "key ring" of trusted public key–name pairs. If User A trusts User B, User A can trust User B's key ring, which may, of course, contain keys received through further distributed trust. This approach works, but if an impostor can dupe even one user, trust in the impostor can spread widely. This is not a good way to manage large public systems of routers, Domain Name System (DNS) hosts, and other sensitive devices.

It is also possible to distribute public keys manually. This works well in small systems, but it does not scale well to very large systems because of the labor involved.

HMACs

Another tool for message-by-message authentication is the key-hashed message authentication code (HMAC). For HMAC authentication, two parties share a secret key. To send a message, one of the parties appends the key to the message, hashes the combination, and adds this hash (the HMAC) to the outgoing message. To test an arriving message, the receiver also adds the key to the message and hashes the combination. This creates the HMAC. If the received and computed HMACs are the same, the sender knows the secret key, which only the other party should use.

HMAC processing requires little processing power because it only involves hashing, which is a fast process. In contrast, public key encryption and decryption, which are used in digital signatures, are extremely processing intensive.

A secret key needs to be created and distributed for each pair of communicating parties, however. In large systems of routers and hosts, this is highly problematic. One approach to reduce this problem is to use public key authentication for initial authentication, then use Diffie–Hellman key agreement or public key distribution to send the secret keys between pairs of communicating parties. This still requires a public key infrastructure.

Another solution is to use community keys, which are shared by all communicating parties in a community. The problem here, of course, is that if a single member of the community is compromised, attackers reading its community key will be able to authenticate themselves to all other members of the community.

Adding Security to Individual Dialogues

The first approach to creating standards-based security is for users to add security to individual existing dialogues. Effectively, this overlays security on the Internet for limited purposes without requiring the creation of completely new Internet standards.

Cryptographic Systems

Secure dialogues require that confidentiality, authentication, and integrity be protected. This is a somewhat complex process involving several phases.

Cryptographic System Phases

Establishing a secure dialogue typically involves four sequential phases (although the specific operations and the order of operations can vary among cryptographic systems). Figure 2 illustrates these phases. The first three are handshaking stages at the beginning of the dialogue. After the secure dialogue is established, the two parties engage in ongoing secure conversation.

First, the two parties must select standards options within the range offered by a particular cryptographic system. For instance, in encryption for confidentiality, the system may offer the choice of a half dozen encryption methods. The communicating parties must select one to use in their subsequent exchanges. The chosen methodology may offer several options; these, too, must be negotiated.

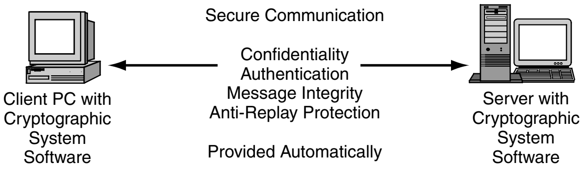


Figure 2: Cryptographic system.

Second, the two parties must authenticate themselves to each other. Although this might seem like it should be the first step instead of the second, the two parties need the first phase to negotiate which authentication method they will use to authenticate themselves.

Third, the two parties must exchange one or more secret keys securely. These keys will be used in the ongoing dialogue that will take place after the "handshaking" steps are finished.

Fourth, the security of the communication is now established. The two parties now engage in ongoing dialogue. Typically, nearly all communication takes place during this ongoing dialogue.

The User's Role

Few users have the training to select security options intelligently, much less handle authentication and other tasks. Consequently, cryptographic systems work automatically. The user selects a communication partner, and the systems of the two partners work through the four phases automatically.

At most, users may have to authenticate themselves to their own systems by typing a password, using an identification card, using biometrics, or by using some other approach. This user authentication phase tends to be the weak link in cryptographic systems because of poor security practices on the part of users, such as using weak passwords.

The Policy Role

Different security options have different implications for the strength of a dialogue's security. Users rarely are capable of selecting intelligently among options. Consequently, companies must be able to set policies for which methods and options will be acceptable, and they must be able to enforce these policies by promulgating them and enforcing their use.

Figure 3 shows how policy guides security in IPsec, which is discussed later in the chapter. In IPsec, security often is handled by IPsec gateways. When two IPsec gateways begin to communicate, they establish security associations, which are contracts for how security will be done. A policy server can tell the IPsec gateways what security association parameters they may or not select. This ensures that the IPsec gateways do not select security associations that are too weak for the traffic they will carry.

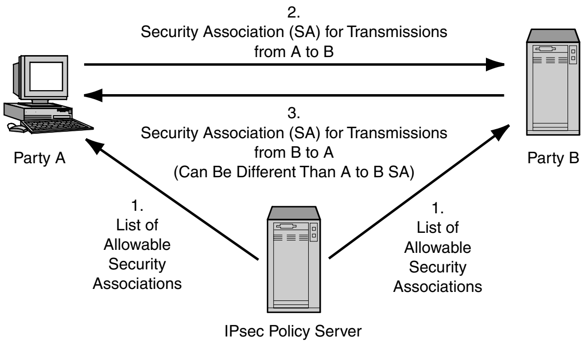

2.
Security Association (SA) for Transmissions from A to B
3.
Security Association (SA) for Transmissions from B to A
(Can Be Different Than A to B SA)
Party A
Party B
1.
List of Allowable Security Associations
1.
List of Allowable Security Associations
IPsec Policy Server

Figure 3: Policy-based security associations in IPsec.

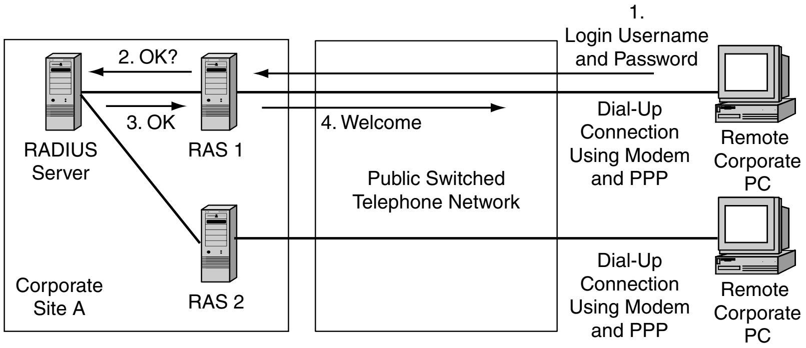

1.
Login Username and Password
2. OK?
3. OK
4. Welcome
RADIUS Server
RAS 1
RAS 2
Corporate Site A
Public Switched Telephone Network
Dial-Up Connection Using Modem and PPP
Remote Corporate PC
Dial-Up Connection Using Modem and PPP
Remote Corporate PC

Figure 4: RADIUS authentication.

This approach works with host computers as well as with IPsec gateways.

Adding Security at the Data Link Layer

Dial-Up Security and PPP

Early computer systems used dial-up security. As Figure 4 shows, the user dialed into a server, generally using the Point-to-Point Protocol (PPP) at the data link layer. Later, companies added security to this approach by creating remote access servers (RAS) for sites. The user dialed into the RAS, authenticated himself or herself (usually with a password), and then received access to all servers at a site or to selected servers.

If a company had several remote access servers, normally a RADIUS (Remote Authentication Dial-In User Service) server or some other authentication server is used to store authentication information remotely, as Figure 5 shows. This way, all RASs use the same authentication data. Otherwise, if a firm has several RASs, their authentication data may be inconsistent. Adversaries might try different RASs until they find one that is configured incorrectly.

PPP can be used without security, but PPP offers moderately good encryption and several options for authentication ranging from nothing to strong security options. Again, the dialogue partners must select the security options they wish to employ.

Tunneling

Although PPP is secure, it is limited to a single data link because it operates at the data link layer, as Figure 5 illustrates. When a connection is made over the Internet, however, each connection between a host and a router or between two routers is a separate data link. PPP cannot function over the Internet.

Consequently, Internet-based data link layer security approaches must use tunneling. In tunneling, a data link layer frame is placed within the data field of a packet (the opposite of the usual situation). The packet is sent from user computer to the RAS, across multiple links. The RAS reads the tunneled (encapsulated) frame. Although this effect seems needlessly complex, it allows us to use traditional PPP user-RAS security over the Internet.

Point-to-Point Tunneling Protocol (PPTP)

The first tunneling protocol was the Point-to-Point Tunneling Protocol (PPTP). As its name suggests, PPTP is a way to tunnel PPP frames over the Internet. PPTP uses PPP encryption and authentication mechanisms, using them over an entire Internet connection instead of over a single dial-up data link. PPTP has a number of moderate security weaknesses, but it is good for low-threat environments. It offers medium security for moderate threat environments.

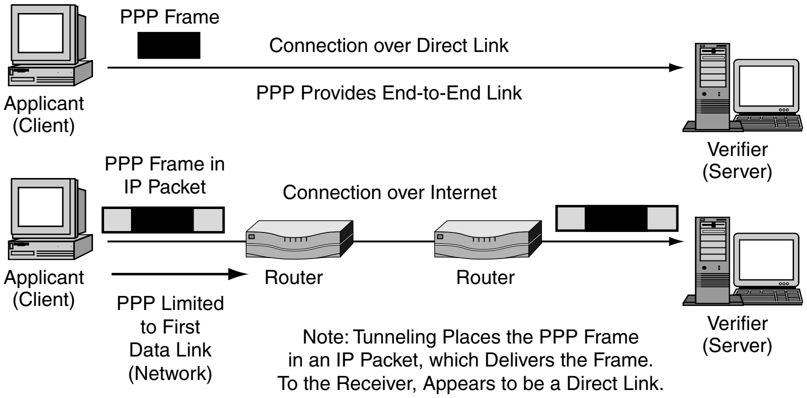

PPP Frame
Connection over Direct Link
PPP Provides End-to-End Link
Applicant (Client)
Verifier (Server)
PPP Frame in IP Packet
Connection over Internet
Applicant (Client)
Router
Router
PPP Limited to First Data Link (Network)
Note: Tunneling Places the PPP Frame in an IP Packet, which Delivers the Frame. To the Receiver, Appears to be a Direct Link.
Verifier (Server)

Figure 5: Point-to-Point Tunneling Protocol.

Layer 2 Tunneling Protocol (L2TP)

Although PPTP works and is attractive, it is limited to transmission over an IP network such as the Internet. The L2TP can work over a number of transmission mechanisms, including IP, frame relay, and ATM, to name just three. L2TP does not offer security by itself, however. It requires users to rely on the IPsec protocol, discussed in the next section, to provide security at the Internet layer during transit. L2TP is a pure tunneling protocol, not a security protocol.

Adding Security at the Internet Layer

The Internet layer (or network layer) is the core layer in TCP/IP internetworking. The Internet Protocol (IP) is the main packet standard at this layer. IP comes in two versions—IP Version 4 (IPv4) and IP Version 6 (IPv6). IPv4 is the dominant version in use on the Internet today, but IPv6 is beginning to grow, especially in Asia, where relatively few IPv4 addresses were allocated when the Internet was first created.

IPsec

The IETF has been working to retrofit IP to be more secure. Their effort has crystallized around a group of standards collectively called IPsec (IP security). Although IPsec was initially planned for IPv6, the IETF has developed it to work with IPv4 as well.

Encryption and Authentication

The dominant way of using IPsec is to employ the encapsulated security protocol (ESP) option, which offers both encryption and authentication. Here I focus on ESP rather than on the authentication header (AH) option, which only provides authentication. AH is useful primarily when encryption is illegal, which is a rare situation.

Tunnel Mode

As Figure 6 shows, IPsec can operate in two modes: tunnel mode and transport mode. In tunnel mode, IPsec is handled by IPsec gateways at the two sites of the communicating parties, not by the computers of the communicating parties themselves. The packet to be delivered securely is tunneled by encapsulating it in another packet, encrypting the packet to be delivered securely, adding authentication, and sending the encapsulating packet from one IPsec gateway to the other. Although attackers can read the IP header of the encapsulating packet, they cannot read the secured packet, nor can they change the secured packet without the change being obvious.

Tunnel mode is attractive because it does not require the individual users to have IPsec software on their computers or to know how to use the software. In fact, the users may not even be aware that IPsec is protecting their packets over the Internet. Windows, which dominates client operating systems today, did not get native

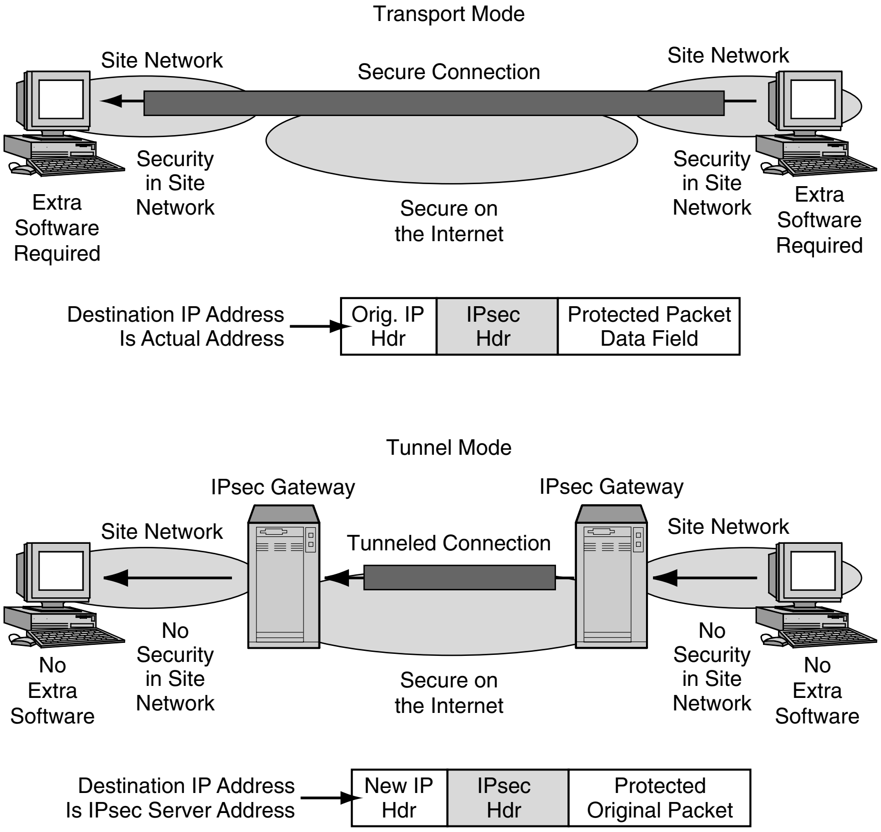


Figure 6: IPsec in tunnel mode and transport mode.

IPsec support until Windows 2000. Implementing IPsec software on the many older personal computers (PCs) in the firm would require a massive investment in large companies.

The disadvantage of tunnel mode is that it does not provide any protection for the packet as it travels within the two sites. The focus is entirely on security during Internet transmission.

Transport Mode

In contrast, as Figure 6 also shows, transport mode offers *end-to-end* encryption and authentication between the two computers. This approach provides security not only while packets travel through the Internet but also when the packets are passing through local site networks on their way to and from the Internet. Although the Internet exposes traffic to many attackers, there also are dangers within corporate networks. In fact, some of the worst attacks are made by corporate insiders working within corporate networks.

Although this end-to-end security is attractive, it comes at a price. Larger firms have thousands or even tens of thousands of PCs. Transport mode requires the presence of IPsec software on every PC to be protected. As noted previously, Windows only began shipping with native IPsec software with Windows 2000, and retrofitting older computers would be extremely expensive.

A more modest problem is that transport mode packets must have the IP address of the receiving computer in the destination address field of the packet header. If sniffers can be placed along the route of these packets, attackers will be able to learn the IP addresses of many corporate hosts. This is a first step in most types of Internet-based attacks. In contrast, tunnel mode packets only have the IP addresses of the receiving IPsec gateway in their destination address fields. This only tells attackers about the IP address of a single machine that usually is well hardened against attacks.

Combining Tunnel Mode and Transfer Mode

One solution to the relative weaknesses of the two modes is to use both. The user can employ transport mode end-to-end but also to use IPsec in tunnel mode between sites as a second layer of protection. This will provide end-to-end encryption while still hiding the IP address of communicating hosts. Of course, implementing IPsec protection twice adds to cost.

Adding Security at the Transport Layer: TLS

IPsec is a complex security mechanism, but it has the advantage of transparency. IPsec protects *all* higher layer traffic at the transport and application layers automatically, without requiring higher-layer protocols to do any work or even be aware of this protection.

In some applications, dialogues are limited to World Wide Web or at most the Web and e-mail. Under these conditions, the costly burden of implementing IPsec is not justified, and many firms turn to a simpler protocol, Transport Layer Security (TLS), which was originally created by Netscape as Secure Sockets Layer (SSL) and was then renamed by the IETF when it took over the standard's development. If your URL begins with "https," then you are using TLS.

Although IPsec provides a blanket of protection at the internet layer, TLS creates a secure connection at the transport layer. This secure connection potentially can protect all application layer traffic. Unfortunately, TLS requires that applications be *TLS-compliant* to benefit from this protection. Although all browsers and Web servers can use TLS, only some e-mail systems can use TLS, and few other applications can benefit from TLS protection at the transport layer.

TLS was created for electronic commerce, in which a residential client PC communicates with a merchant server. TLS requires the merchant to authenticate itself to the client using digital certificates and digital signatures. Few residential client PC owners have digital certificates. So although TLS allows client authentication, it makes it optional. This lack of mandatory client authentication is a major security vulnerability that is intolerable in many situations.

In corporate settings, however, the organization can require clients as well as servers to use digital certificates. This eliminates the main security limitation of TLS, although TLS in general offers only moderate security.

Adding Security at the Application Layer

For the application layer, IETF is adding security to a number of application layer standards (described in further detail later). In addition, many vendors are adding proprietary security to their applications, such as database applications.

Multilayer Security

As Table 1 shows, security can be applied at several standards layers. A major principle of security is defense in depth. Almost all security protections break down from time to time. Until they are fixed, the attacker has free access—unless the attacker must break through two or more lines of defense, the company still will be protected while it repairs a broken security countermeasure. Consequently, companies would be better protected if they added dialogue security at more than one layer, for example, by implementing both IPsec and application security. This is an expensive undertaking, however, so it is uncommon.

Table 1 Multilayer Security

LAYER	CRYPTOGRAPHIC SYSTEM
Application	Kerberos
Transport	Secure Sockets Layer/Transport Layer Security
Internet	IPsec (Internet Protocol security)
Data link	Point-to-Point Tunneling Protocol, Layer 2 Tunneling Protocol
Physical	Not applicable; no messages are sent at this layer, only individual bits

Added Dialogue Security and Firewalls

Although adding security to dialogues passing through the Internet is attractive, it creates problems for firewalls. Firewalls are designed to examine all packets coming into or leaving a firm. When firewalls find attack packets created by hackers, they drop them. Almost all cryptographic system packets are encrypted, however. Unless they are decrypted before passing through the firewall, the firewalls cannot scan their packets for attack signatures. Consequently, companies that buy added security by implementing cryptographic systems tend to lose some scanning security. This places a heavier burden on end stations to do scanning, and many end stations are client computers, the owners of which lack the willingness, much less the knowledge, to do packet scanning.

Adding Security to Individual Internet Standards

Although adding dialogue security to individual dialogues works, it would be better for Internet standards themselves to offer high security. When the Internet was first created, none of its standards offered any security. Today several TCP/IP standards offer security. Unfortunately, the more closely one examines their security features, the less adequate many appear. Fortunately, the IETF has a broad program to add security to a broad range of existing Internet standards individually.

A Broad IETF Commitment: The Danvers Doctrine

In 1995, a meeting of the IETF in Danvers, Massachusetts, reached a consensus that the IETF should develop strong security for all of its protocols, Request for Comments (RFC) 3365. Originally, the consensus was limited to a decision to use strong rather than weak encryption keys that met existing export restrictions. Soon, however, this Danvers Doctrine expanded into a consensus to develop strong security for all TCP/IP protocols. As a first step, all RFCs are now required to include a section on security considerations. Considerable progress has also been made in adding security to individual TCP/IP standards, as the following sections detail.

User Choice

Although all TCP/IP standards are to be given strong security, the IETF decided that it should be the option of individual organizations whether to implement security in individual protocols. In part, this decision reflects a desire not to force security on anyone. In part, it also reflects the fact that some networks are in protected environments that do not require security.

An important consequence of the decision to leave security up to organizations is that organizations must decide which security options to use. There are five broad layers of functionality in networking—physical, data link, internet, transport, and applications. Implementing security at all layers would be horrendously expensive and generally unnecessary.

Consequently, many organizations wish to implement security at only one layer. Although this provides security, individual security technologies often are found to have vulnerabilities. To maintain protection during these periods of vulnerability, organizations probably will wish to implement security in the protocols at two layers at least. While one layer's security is being repaired, the other or others will continue to provide protection.

SECURITY AND VULNERABILITIES

In discussing the security of Internet protocols, there are two issues. One is whether security has been built into the protocol at all and to what degree it has been placed in the protocol. This is the aspect of Internet protocol security I focus on in this chapter.

Another aspect of Internet protocol security is whether the protocol or its implementation has vulnerabilities that attackers can exploit. In some cases, the protocols themselves are exploitable because of design oversights. In more cases, vendor implementations cause problems. For instance the BIND program, which is dominant for Domain Name System servers, and the Sendmail program, which is dominant on UNIX Simple Mail Transfer Protocol (SMTP) servers, have both had long and troubled histories of security vulnerabilities.

In some cases, such as BIND and Sendmail, a single code base is used by most or all vendor implementations. In other cases, the same flaw is discovered in multiple vendor code bases. (For example, in February 2001, two major vulnerabilities were found in the Simple Network Management Protocol version 1 programs of multiple vendors.) In both of these situations, the discovery of a vulnerability can suddenly put a significant fraction of the Internet's servers or routers at risk until organizations can download patches.

In addition, many firms fail to install patches quickly or at all. This makes some vulnerabilities exploitable for weeks, months, or even years after they are discovered. All too often, firms only patch vulnerabilities in earnest when a virus or hackers create widespread damage.

Although vulnerabilities are important, they are situation-specific. This chapter must focus on security within Internet standards themselves.

The Core Four

There are four core standards at the heart of the Internet. These are IP, TCP, and UDP, and ICMP.

Internet Protocol (IP)

The main job of the Internet Protocol (IP) is to move packets from the source host to the destination host across the Internet, which consists of thousands of networks connected by routers. IP is a hop-by-hop protocol designed to govern how each router handles each IP packet it receives.

A packet may travel over one to two dozen routers as it passes from the source host to the destination host. Many core routers in the Internet backbone handle so much traffic that they can barely keep up with demand. To reduce the work done on each router, IP was designed as a simple, unreliable, and connectionless protocol. Although packet losses on the Internet are modest, IP is a "best effort" protocol that offers no guarantee that packets will arrive at all, much less arrive in order. Given this minimalist vision for IP, it is hardly surprising that security was completely left out of IP's core design. Given continuing router overload, it would be difficult to make IP secure throughout the Internet.

IPsec was created to add security to the Internet Protocol, making IPsec more than a security overlay method. IPsec cannot achieve its promise without a truly worldwide system of certificate authorities, however.

Transmission Control Protocol (TCP)

To compensate for IP's unreliability at the internet layer, TCP/IP was given a reliable sibling protocol, TCP at the transport layer. When TCP was created in the early 1980s, however, security technology was far too immature to be implemented. Today there are no plans to add security to TCP given the IETF's focus on IPsec and IPsec's ability to provide security for higher layer protocols.

User Datagram Protocol (UDP)

When transport layer error correction is not required or is not practical, applications specify the UDP at the transport layer. Like TCP, UDP was created without security and is also is unlikely to receive security extensions because of the IETF's reliance on IPsec.

Internet Control Message Protocol (ICMP)

IP merely delivers packets. It does not define any Internet layer supervisory messages. To compensate for this, the IETF created the ICMP to carry supervisory information. ICMP messages are carried in the data fields of IP packets. ICMP messages allow hosts to determine if other hosts are active (by "pinging" them). It also allows hosts to send error messages to other hosts and to tell other hosts or routers to act differently from how they have been acting. ICMP is a powerful tool for network managers. Unfortunately, this power also makes ICMP a popular tool for hackers. Making ICMP less useful to hackers would also tend to make it less useful to network administrators, so the IETF has done nothing to implement ICMP security. For this reason, most corporate firewalls block all ICMP traffic except for outgoing pings and returning pongs.

If ICMP were given good authentication, corporations might be more willing to allow it through firewalls. In addition, authenticated ICMP would protect against attacks generated within the firewalls at individual sites. There are no current IETF activities to add authentication and other security to ICMP, however. The reason may be that ICMP messages are carried in the data fields of IP packets, which can be protected by IPsec, including the authentication of the sending ICMP process.

Administrative Standards

The "core four" Internet standards do most of the work of the Internet, but several other administrative protocols are needed to keep the Internet functioning. The following sections discuss security in the most important of these standards.

Domain Name System (DNS)

The DNS allows a source host to find the IP addresses of a destination host if the source host only knows the destination host's host name. In a sense, DNS is like a telephone directory for the Internet.

Most client users only know the host names of the servers they use. Consequently, if the DNS were to fail for some period of time, clients could no longer reach servers. On the positive side, DNS is hierarchical and highly distributed, so that if one DNS server goes down, service continues. The top level of the DNS server hierarchy only has 13 root DNS hosts, however, and one of them periodically feeds changes to the others. Although these root DNS hosts are geographically distributed, most use the same software, leaving them open to common mode attacks if vulnerabilities are found. Also, geographical distribution is not much protection when there are only 13 targets to attack. In later 2002, a brief denial-of-service attack degraded the service given by 9 of the 13 root DNS hosts. Had this attack been more intense, and had it continued over many hours, service disruption across the Internet would result.

The threat of such concentrated resources (rare on the Internet) is real, but the root DNS servers are well protected with both technical and human protections. Below the root servers are a larger number of DNS servers for the various generic and national top-level domains, and there are many corporate DNS servers. At each level, there are more DNS servers to give less concentration, but the loss of nonroot DNS servers near the top could still be widely damaging to the resources within the domain they serve.

To address problems with DNS security, the IETF DNSSEC Working Group is developing an authentication approach based on public key encryption with digital signatures. There are a (relatively) limited number of high-level DNS servers, so giving each a public key and a private key is not too daunting a challenge. As long as DNS servers do not get taken over so that their private keys can be stolen, authentication should be good. The DNSSEC effort is just getting underway, however, so widespread DNS protection will not be realized for some time to come. Nonetheless, methods for the critical issue of authentication have already been created (RFC 3118).

Simple Network Management Protocol (SNMP)

The goal of the Simple Network Management Protocol (SNMP) is to allow a central administrative computer (the manager) to manage many individual managed devices. Figure 7 shows that the manager talks to an agent on each device. The manager does this by sending the managed devices Get messages (which ask for information on the device's status) or Set messages (which tell the managed device's configuration).

Obviously, SNMP is a powerful management tool. Unfortunately, it is also a golden opportunity for attackers. It potentially allows them to learn a great deal of information about a network through Get commands. It also allows them to do an endless amount of damage through malicious use of the Set command to create misconfigurations on large numbers of devices. These misconfigurations can make the devices unusable or can even make them interfere with the operation of other devices.

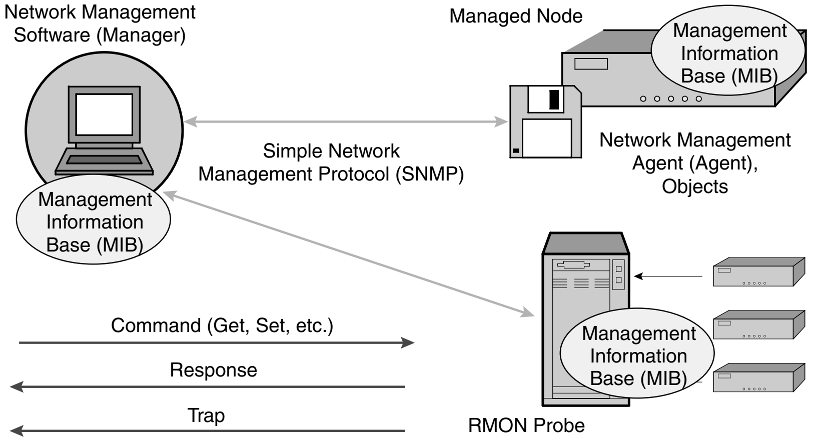


Figure 7: Simple Network Management Protocol (SNMP).

SNMP version 1 had no security at all, making the protocol extremely dangerous, given its power. SNMP Version 2 was supposed to add security, but an inability to settle differences within the IETF prevented full security from being built into it.

Version 2 did receive one authentication advance. To send SNMP messages to a version 2 SNMP agent, a manager would have to know the agent's "community string," which is like a password. In practice, many firms have all of their agents use the same community string. In fact, many do not change the vendor's default community string, which often is "public." Furthermore, community strings are sent in messages without encryption, so attackers with sniffers can quickly learn community strings.

Only in version 3 did security get built into SNMP extensively. Version 3 offered confidentiality (optional), authentication, message integrity, and time stamps to guard against replay attacks.

Unfortunately, SNMP version 3 security is based on HMAC authentication and integrity, which requires the network administrator to have a secret bit string and for each managed device to know that secret bit string. This means that any attacker learning the string will be able to impersonate that manager. As noted earlier, widely shared secrets tend to be easy for attackers to learn, and when many devices must know the shared secret, the company is not likely to change the shared secret frequently.

This severe problem could be solved if the standard specified the use of digital certificates for managers. Authentication and integrity only require that managed devices know the manager's public key, and there is no problem with shared public keys. In addition, public key encryption allows the secure exchange of symmetric keys for bulk encryption for confidentiality. Nonetheless, the large number of devices involved and the desire to keep processing loads on managed devices low has made public key authentication and integrity unattractive to standards makers.

In addition, SNMP is connectionless, and it uses connectionless UDP at the transport layer. Connectionless operation reduces the network load created by SNMP. It is difficult to add security to connectionless applications because, as noted earlier, security typically begins with handshaking phases that are difficult (although possible) to implement in a connectionless protocol.

Lightweight Directory Access Protocol (LDAP)

Increasingly, companies store security information and other critical corporate information in central repositories called directory servers. To query data in a directory server, devices commonly use the Lightweight Directory Access Protocol (LDAP), which governs search request and responses.

The IETF has long understood the importance of creating good security for LDAP in light of the extreme importance of information contained in the directory. The core security focus of LDAP must be authentication because successful impostors could learn large amounts of damaging information about a firm if they could get wide access to this data by claiming to be a party with broad authorization to retrieve data.

The first version of the standard, LDAP version 1, had no security and generally was rudimentary. LDAP version 2 provided authentication options for the first time. It permitted initial anonymous, simple, and Kerberos 4 authentication.

Anonymous authentication means that no authentication is needed for some users. These users would be given access to certain information on the server, much as anonymous File Transfer Protocol (FTP) users are allowed to download information from selected parts of an FTP server. Generally speaking, anonymous authentication should be turned off on directory servers with sensitive information.

Simple authentication requires the user to transmit a user name and password. This information is sent in the clear, without encryption. Anyone who reads the user name and password will be able to authenticate themselves later as the party.

Kerberos 4 authentication uses a central authentication server separate from the directory server. Unfortunately, Kerberos 4 had known weaknesses. Most notably, Kerberos 4 uses a long-term shared secret key to communicate with each device. If enough traffic uses this key, a cryptanalyst can determine the key and then impersonate the device. (Kerberos version 5 only uses the long-term key for initial authentication and then uses a session key for remaining communication during a time-limited session.)

The latest version of LDAP, version 3, also supports anonymous and simple authentication. Again, anonymous authentication should not be turned on for sensitive servers. In addition, sites that use simple authentication should protect application-level LDAP traffic with lower layer protection, most commonly TLS at the transport layer. This ensures that eavesdroppers will not be able to read the user name and password.

LDAP version 3 also supports the simple authentication security layer (SASL). This method allows the searcher and the directory server to authenticate each other in several ways, including Kerberos 5, HMACs, and "external authentication." External authentication means that the parties can perform authentication any way they wish, essentially taking authentication outside the LDAP process. For instance, if the searcher and directory server create a TLS connection at the transport layer, the authentication that takes place at that layer may be sufficient. SASL is extremely flexible, but companies that use it must ensure that sufficiently strong authentication is selected.

Although LDAP version 3 offers good authentication, it does not support public key authentication using digital signatures except through SASL or if the organization using it uses TLS and digital certificates for both parties to secure LDAP traffic.

LDAP version 3 makes encryption for confidentiality optional. This may seem odd, but many firms use TLS at the transport layer to protect LDAP traffic. TLS does encryption, and many firms feel that doing encrypting LDAP traffic at the application layer as well would be needlessly expensive.

Address Resolution Protocol (ARP)

When a router receives a packet, it looks at the packet's destination IP address. Based on this information, the router sends the packet back out another port, to the destination host or to a next-hop router that will handle the packet next. In this chapter, I use the term "target" for the destination host or next-hop router.

The destination IP address in the packet gives the target's IP address. To deliver the packet, however, the router must encapsulate the packet in a data link layer frame and deliver the frame to the target. This requires the router to know the data link layer address of the target, for instance, the target's Ethernet message authentication code address.

If the router does not know the target's data link layer address, the router must use the Address Resolution Protocol (ARP). The router first broadcasts an ARP request message to all hosts on the subnet connected to the port out which the packet is to be sent. If the subnet is an Ethernet network, the broadcast data link destination address is 48 ones. Switches will deliver frames with this address to all hosts on the subnet.

Within a broadcast frame, the router sends an ARP request message. The ARP request message contains the IP address of the target. All hosts except the target ignore this message. The target, recognizing its IP address and sends an ARP reply message back to the router. This message tells the router the target's data link layer address.

Now that the router knows the target's data link layer address, the router takes the IP packet it has been holding and places it in a frame with the target's data link layer address in the destination address field. The router sends this frame to the target.

The router also places the target's IP address and data link layer address in the router's ARP cache. When the next packet addressed to this IP address arrives, the router will not have to use ARP. It will simply look up the data link layer address in the ARP cache.

ARP assumes that all hosts are trustworthy, and thus an attacker host on the same subnet as the target can use this trust to make an attack. Most simply, the attacker host can flood the router with ARP reply messages associating its data link layer address with every IP address on the subnet. If the router accepts some of these messages, it will suffer ARP cache poisoning.

The next time a packet arrives at the router, the router may rely on the poisoned cache to send the packet to the attacker rather than to the correct target. This allows the target to read the contents of the packet. The attacker can then simply drop the packet, creating a denial-of-service attack against hosts on the subnet.

Alternatively, the attacker can listen to real ARP response messages to learn the true data link layer addresses of subnet hosts and send the packet on to the correct target. This attacker-in-the-middle attack permits the attacker to continue reading packets without alerting the victim hosts to this fact. This approach may also allow the attacker to hijack the communication session between the source and destination hosts, inserting its own packets into the dialogue. Currently, there are only preliminary efforts within the IETF to add authentication to ARP.

Dynamic Host Communication Protocol (DHCP)

It is possible to configure a client PC manually, typing in its IP address and other configuration parameters needed to communicate over the Internet or a corporate internet. Servers typically are configured manually. Client PCs, however, are configured automatically by getting these configuration parameters from a Dynamic Host Configuration Protocol (DHCP) host.

When a client PC boots up, it realizes that it has no IP address. It broadcasts a DHCPDISCOVER message to learn what DHCP servers are reachable. Each DHCP server that receives the request sends back a DHCPOFFER message that gives an offer specifying what IP address it will provide, a lease time (how long the client may keep the IP address), and other configuration information.

The client PC selects one offer and sends a DHCPREQUEST to that server accepting the offer. The DHCP server sends back a DHCPACK message acknowledging the acceptance. The client PC is now configured.

DHCP was note designed to authenticate the client or the DHCP server, and this allows several possible attacks. For an example, an attacker posing as a client PC can repeatedly request DHCP service, each time accepting an

IP address. Within a few seconds, the DHCP server will be drained of available IP addresses. If a legitimate client then requests configuration service, it will not be able to get an IP address and so will not be able to use the Internet or an internal corporate internet.

The attacker also can impersonate a DHCP server. The attacker will then respond to all DHCPDISCOVER messages with an attractive offer (long lease duration, and so forth). If a client accepts this configuration information, this information may cause the client to send packets that cannot be delivered or to act in other ways that harms itself or other hosts.

The IETF is now working on DHCP authentication. RFC 3118 (Authentication for DHCP Messages) defines two authentication options for DHCP. The first involves a key that is shared by the client PC and by the DHCP server. This key is transmitted in the clear. This option is only useful for protecting the client against a host that accidentally installed a DHCP server.

The second option requires each client to share a secret key with each DHCP server. The client and server use this shared secret to add an HMAC to each message. If vendors adopt this option, companies can have fairly strong authentication, although the authentication will not be as strong as it would be if digital signatures and digital certificates were used.

Application Layer Standards

To users, the Internet is attractive because of its application standards for the World Wide Web, e-mail, and other popular services. Although decent security is present for some application standards, many continue to have only weak security, if they have any at all.

Hypertext Transfer Protocol (HTTP)

To communicate with a Web server, a browser uses the Hypertext Transfer Protocol (HTTP), a simple protocol that offers no security by itself. All browser and Web server programs support TLS security, however. Although TLS has some theoretical weaknesses, it offers sufficiently strong Web server authentication and confidentiality for consumer e-commerce transactions. For high-volume business-to-business e-commerce, however, TLS security is marginal.

E-mail Security

Security for e-mail is perhaps the great scandal in IETF history. Competing cliques within the IETF have consistently refused to cooperate in selecting security standards for e-mail. Consequently, companies that wish to implement e-mail security have to use a nonstandard method to do so.

Not surprisingly, sending application layer e-mail traffic over a secure TLS transport connection is the most popular way to secure e-mail in organizations today. TLS is well understood, offers consumer-grade security, and has been widely used for HTTP security for several years. Fewer organizations use another approach, S/MIME (Secure Multipurpose Internet Mail Extensions).

A third approach, "Pretty Good Privacy" (PGP), is used primarily by individuals. Organizations have tended to stay away from PGP because it uses user-based transitive trust (if User A trusts User B, and if User B trusts User C, User A may trust User C). As noted earlier, this is not a good security policy. If a single user mistakenly trusts an impostor, others may unwittingly trust the impostor as well. Also, PGP has had a troubled development history.

Remote Access

The first ARPANET application was Telnet, which allows a user to log in to a remote computer and execute commands on it as if the user were local. This allows ordinary users to access remote services. It also allows system administrators to manage servers and routers remotely. This use of remote administration is attractive, but it must be done carefully or hackers will end up "managing" corporate servers and routers.

Unfortunately, Telnet has poor security. For example, Telnet does not encrypt host user names and passwords. Hackers can intercept user names and passwords and then log in as these users with all of their privileges.

Telnet should never be used for remote administration because hackers intercepting root passwords would be able to execute any commands on the supervised machine. Telnet is not alone in this respect. In the UNIX world, rlogin and rsh do not even require passwords to gain access to a computer, although other conditions must apply.

For remote administration, some organizations turn to the Secure Shell (SSH) protocol, which offers good authentication, integrity, and authentication. Unfortunately, SSH version 1 had security flaws, and although these have been fixed in SSH version 2, many version 2 implementations will also allow version 1 connections, thus leaving the system open to attack.

File Transfer Protocol (FTP) and Trivial File Transfer Protocol (TFTP)

Another early ARPANET service that continues to be popular on the Internet is the File Transfer Protocol (FTP), which allows users to download files from a remote computer to their local computers and sometimes to upload files from their local computers to the remote computer. Unfortunately, FTP also sends user names and passwords in the clear, making it dangerous. In addition, although the use of Telnet has declined to the point where quite a few companies simply stop all Telnet traffic at their firewalls, FTP is still widely used. The FTPEXT Working Group is considering security for FTP.

FTP has a simpler sibling, the Trivial File Transfer Protocol (TFTP). TFTP does not require user names or passwords, making it a darling of hackers who often use this protocol after taking over a computer to download their rootkits (collections of hacker programs) to automate their exploitation of the computer they now "own."

Other Applications

There are many other Internet applications, and their standards vary widely in their degree of security. In addition, database applications and many other applications are not standards based. Unfortunately, many operating system installations automatically turn on applications without the systems administrator's or user's knowledge. For example, when SNMP vulnerabilities were found in 2002, it was discovered that many machines should not have been running SNMP at all. An important rule in hardening clients and servers against attack is to turn off all applications that are not absolutely needed to run the computer.

One important area for the future is security for Web services. Although some progress has been made, existing security is so embryonic that it would be premature to discuss.

MAKING THE INTERNET FORENSIC

Forensics is the application of science to criminal prosecution. One of the biggest problems in Internet security today is simply that there is almost no way to stop attackers. Victims must operate almost entirely on the defensive while under constant attack. As any military scientist will attest, this is a poor basis for security.

The Problem of Prosecution

Few attackers are ever prosecuted. In some cases, the laws of a country are insufficient. In the case of two of the worst viruses—CIH and Love Bug—governments gave up on prosecution because they could not prosecute under their countries' statutes. Although the legal situation has become clearer, prosecutors are reluctant to prosecute security incidents. In many cases, the perpetrator is a minor, and only slap-on-the wrist penalties are possible. In other cases, prosecutors feel that even for adult perpetrators, prosecution is too expensive for the minor penalties that the courts have tended to apply. Post-September 11 laws have toughened penalties somewhat, but prosecutors in general still feel that prosecution for hacking and other attack activities is a low priority. Unless losses are in the tens of thousands of dollars, law enforcement officials probably will not prosecute.

Administrative Sanctions

A more promising approach is to sanction perpetrators by cutting off their Internet access and making it difficult or impossible for them to restore Internet access. To take the onus off the ISP, a court could make the decision based on ISP data, and the legal system could even impose ticket-based fines for modest infractions (such as sending probing messages), as is done for automobile misbehavior.

Attacker Identification

In many cases, attackers do little or nothing to hide their identities. For instance, hackers often send out many probing messages against companies they plan to attack. These messages must have the attacker's IP address in the packet so that the attacker can see replies.

In other cases, attackers hide their identities by working through a string of computers, as shown in Figure 8. The last computer in the string can be identified fairly quickly, but tracing several attack computers in sequence tends to take so long that the attacker can hide his or her identity before defenders can trace the attack.

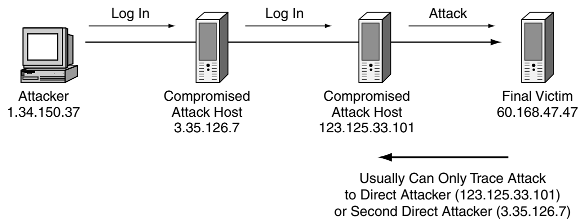


Figure 8: Using a sequence of computers to generate an attack.

To make attackers easier to identify, ISPs could keep connection information far longer than they do today. (Today IP source address information is only stored for a brief period of time for billing purposes if it is stored at all.) Because of the costs involved, ISPs have been unwilling to store connection information for long periods of time. However, several countries are now considering laws to require the long-term storage of connection information plus the ability of law enforcement personnel to get at such information quickly for traceback.

The IETF is now working on a new standard to make traceback easier. This is a new ICMP message type, traceback. Each router will randomly but rarely transmit ICMP traceback messages to the destination host. These traceback messages will list routers along the way.

Even if an IP source address is spoofed, occasional ICMP traceback messages that the victim received in a denial-of-service attack or any other type of sustained attack will still be able to determine the route of the attack. Routers may also send occasional traceback messages back to their sources to aid in the analysis of certain types of reflection attacks. It will take some time for the IETF to ratify the new message type and for many of the Internet core routers to implement the standard, if they implement it at all. (It would cost money for them to do so and would reveal the ISP's internal router architecture, leaving ISPs vulnerable to certain attacks.)

Proactive ISP Efforts to Stop Attacks

More proactively, ISPs could use intrusion detection systems to watch packets coming from their own customers. If many attack packets began to appear, the ISP could inform the attacker and threaten to cut off Internet access. If the attacker failed to desist, the ISP could cut off their service. Generally, ISPs do not wish to do this (although a few already do). First, they do not want to lose customers. Second, intrusion detection and analysis would be expensive. ISPs, in other words, would have little to gain and much to lose. Stopping hackers at their source connection is so attractive, however, that governments should force ISPs to do intrusion detection and to cut off attackers. There is now sufficient danger to the entire Internet system to make such a requirement reasonable. ISPs that fail to do filtering might even be "black holed" by other ISPs, much as some ISPs that fail to stop their users from spamming have been black-holed.

Of course, there would have to be procedural safeguards to protect users who are not really attackers but whose computers have been compromised by attackers.

(At the same time, it is reasonable to require users of compromised computers to clean out their computer if they wish to continue service.)

An Internet Network Operations Center?

One way to increase security on the Internet is to create a network operations center (NOC) for the Internet. Data on the status of the Internet would be delivered to the NOC from multiple sensors distributed throughout the Internet. This would include performance data to spot congestion on the Internet. It would also include attack data.

Centralizing attack data and adding trend detection software might allow the early detection of massive attacks so that actions to stop the attack could be undertaken before massive attacks spread too far. Early detection might also improve the probability of finding the sources of attacks.

Unfortunately, this type of NOC could also allow snooping on legitimate traffic for political purposes. Given diverse opinions on hacktivism (hacking for political purposes), separating legitimate from illegitimate traffic might be controversial.

Creating Multiple Internets?

Another possibility is to create multiple Internets with different levels of security. During the 1970s, the U.S. military attempted to do this by creating milnet (which became the Defense Communication Network). What we may see in the future is the emergence of a business-class Internet and a consumer Internet (probably today's Internet). The business-class Internet would have high security throughout, and the consumer Internet would either have today's casual security or the type of expanded security we have seen in this chapter.

Interconnecting Internets with different security levels would be difficult. Although firewalls between high-security and low-security Internets would lessen the danger, air gaps between them would be needed for high security. Even then, users might surreptitiously interconnect multiple Internets by connecting servers or clients to two or more Internets simultaneously.

THE STATE OF INTERNET SECURITY STANDARDS

Today if two communication partners wish to communicate securely, they can do so by adding dialogue security on top of nonsecure Internet transmission. For lightweight needs, they can turn to PPTP or TLS. For industrial-strength security, they can use IPsec.

General Insecurity

More generally, however, the standards that the Internet needs for message delivery (IP, TCP, UDP, and ICMP), Internet supervisory standards (DNS, SNMP, LDAP, etc.), and Internet application standards (the Web, e-mail, etc.) vary widely in security from none to semiadequate. Consequently, the Internet today is rather fragile and open to seriously damaging attacks. Given the pace of security implementation in individual standards, the fragility of the Internet is not likely to change radically in the immediate future. Even after more secure standards are developed, it will take several years for them to be widely adopted.

The Broad IETF Program

At the same time, there is extremely broad concern with security across the IETF. Although some standards have better security than others, standards working groups appear to be playing leapfrog in their efforts to improve security across a broad spectrum of Internet standards.

The Authentication Problem

The most difficult problem in Internet security standards is authentication. Quite simply, unless authentication uses public key authentication (such as digital signatures) coupled with digital certificates managed by a reliable and well-regulated network of certificate authorities, authentication strength will only be moderate. Creating large public key infrastructures will take years, however, and although European countries are moving to manage and regulate certificate authorities, the United States is has adopted the let-the-market-do-it philosophy that has worked so well recently in the energy industry and in corporate financial reporting.

The Primary Authentication Problem

One broad and deep problem in authentication is the primary authentication problem—proving the identity of a person or organization in the first place, say, to give them a digital certificate or accept them as an employee. Identity theft has long allowed impostors to obtain fraudulent drivers' licenses and other authentication instruments. Even within closed systems, such as corporate networks, the prime authentication problem can be daunting. For consumer authentication and other large communities, the technical and operational problems for credible initial authentication are extremely daunting.

Protection From Denial-of-Service Attacks

One security threat that has barely been faced so far is the prospect of denial-of-service attacks. Although a few standards have taken some steps to reduce these attack risks, this is an area of little general development within the IETF.

The Need for a More Forensics-Friendly Internet

Although technical security is important, tremendous strides in security could be made if ISPs were required to cut off service to attackers and if ISPs and corporations were required to keep audit data for at least a month instead of for a few hours or days. In general, we need laws to make the Internet a more forensics-friendly environment.

Corporate Responsibility and Operational Security

We also need to make the Internet more of a responsible community. In many attacks, adversaries use "innocent" computers as attack tools. This is possible because so many irresponsible users and companies fail to protect and harden their computers. Many attacks, especially denial-of-service attacks, can only be stopped at the source, and many attacks, including virus attacks, can only be reduced substantially by reducing the number of gullible e-mail users. At the corporate level, firms should harden their computers against being taken over and also should do egress filtering on their firewalls to prevent their computers from attacking others if they are taken over. Standards may help, but "operational security" is still the key to security on the Internet.

GLOSSARY

Community key A secret key shared by multiple devices within a network.

Cryptographic system A standards-based system for automatically providing multiple protections in a dialogue between two parties.

Denial-of-service attacks Attacks that attempt to render a computer or network useless; this type of attack is not addressed in most Internet security efforts.

Dialog security Security applied to a dialogue between two parties; often added on top of nonsecure Internet transmission.

Digital signature A message-by-message authentication system based on public key encryption. To work well, the parties should have digital certificates from trusted certificate authorities.

Forensics The application of science to prosecution.

Internet Engineering Task Force (IETF) The body that creates TCP/IP standards for the Internet. The IETF is now engaged in a broad effort to add security to Internet standards.

IPsec The family of standards for adding security to both version 4 and version 6 of Internet Protocol transmission.

Kerberos server An Authentication system in which a central server gives out authentication information.

Key-hashed message authentication code (HMAC) A message-by-message authentication system based on a secret key shared by the two parties.

Multilayer security Providing security at multiple layers to provide defense in depth.

Password authentication Authentication in which the two sides know a password. Sometimes the password is sent in the clear (without encryption).

Policy server In a security system, a server that tells devices how they must implement security.

PPP tunneling (PPTP) Delivering a point-to-point (PPP) frame across the Internet by encapsulating it in an IP packet.

Remote access Accessing a computer remotely for management; includes Telnet (nonsecure) and secure shell protocol (secure in version 2).

Transport Layer Security (TLS) A general standard for adding security at the transport layer; formerly called Secure Sockets Layer (SSL) and also called HTTPS because URLs beginning with "https" are requesting SSL/TLS security.

Virtual private network (VPN) Family of standards for adding security to a dialogue. Includes PPTP, IPsec, and TLS.

CROSS REFERENCES

See *Authentication; Digital Signatures and Electronic Signatures; Encryption; Passwords; Virtual Private Networks: Internet Protocol (IP) Based.*

REFERENCES

Panko, R. (2004). *Business computer and network security.* Upper Saddle River, NJ: Prentice-Hall.

FURTHER READING

Blumenthal, U., & Wijen, B. (1998, January). User-based security model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3). RFC 3413.

Dierks, T., & Allen, C. (1999, January). The TLS protocol version 1.0. RFC 2246.

Harrison, R. (2002, November). LDAP: Authentication methods and connection-level security methods, Internet draft.

Rose, S. (2002, September). DNS security document roadmap. Retrieved from http://draft-ietf-dnsext-dnssec- roadmap-06.IETF.org.

Thomas, S. A. (2002). *IP switching and routing essentials.* New York: Wiley.

Tonsley, W., Valencia, A., Rubens, A., Pall, G., Zorn, G., & Palter, B. (1999, August). Layer two tunneling protocol "L2TP." RFC 2661.

Ylonen, T., Kivinen, T., Saarinen, M., Rinne, T., & Lehtinen, S. (2002, September 20). SSH authentication protocol, Internet Draft.

Internet2

Linda S. Bruenjes, *Lasell College*
Carolyn J. Siccama, *University of Massachusetts Lowell*
John LeBaron, *University of Massachusetts Lowell*

INTRODUCTION TO INTERNET2

Context

The Internet2 project was established to support a variety of academic innovations requiring advanced networking capabilities. The mission of Internet2 is to improve scholarship and to serve the needs of academics in universities, colleges, and schools who require cutting-edge data processing solutions. This “invitation only” initiative is hosted by a consortium of academic, corporate, and government entities known as the University Corporation for Advanced Internet Development (UCAID). Over 200 U.S. educational institutions, recognizing the potential of technologically enhanced teaching, learning, and research tools, have partnered with corporate and government agencies to establish this “next-generation Internet.” UCAID’s vision is to facilitate and coordinate the development, deployment, operation, and technology transfer of advanced, network-based applications and network services to further U.S. leadership in research and higher education and accelerate the availability of new services and applications on the Internet (Houweling, 1999). This next-generation Internet promises to do what academics and researchers hoped the first-generation Internet would do: allow scholars to collaborate with colleagues and students around the world.

Need

The present-day Internet, originally fostered by a partnership of academia, industry, and government, was not designed to handle the volume of personal and commercial use that it is currently expected to support. Originally developed to exchange data through file transfer protocol (FTP) and share resources by remotely logging in through Telnet connections, the Internet’s popularity grew with the advent of the graphical user interface (GUI), which stimulated the use of electronic mail, the Internet’s most-used feature. According to Computer Industry Almanac, Inc. (2002), Internet usage has exploded, growing from 4 million users in 1991 to over 530 million users in 2001. By the end of 2005, it is projected to host over one billion users. At the same time the general population’s use of this medium has increased, the academic institutions that pioneered the Internet have been developing online educational programs that require far more speed and bandwidth than the current networks provide. Some of these programs involve revolutionary collaborative models, to be used by teachers, scientists, and researchers, that are expected to run on networks 1,000 times faster than today’s fastest networks.

The need to expand educational opportunities through the use of virtual learning laboratories and other distance learning opportunities prompted a number of educators to develop interactive multimedia applications, precipitating the need for a powerful and speedy network with capabilities far outreaching the original Internet. Out of this grew a collaborative effort by a number of universities and research centers to research and develop networking technologies that would support their needs. They expressed the need for a reliable and high-capacity technological infrastructure that would allow a collaborative exchange of research materials and an open atmosphere for the exchange of software development tools. Thus, in 1996, 30 universities collaborated on plans to build a communications network that was to be more powerful than the current network system. This new network system, Internet2, would assist these academics in returning to their original vision of a vehicle primarily used to promote education and research.

SHORT HISTORY OF THE INTERNET

Origins

Rooted in the military culture of the cold war, the Internet's original ancestor was established at the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense. It was called ARPANET and was created out of the need for a robust, redundant system capable of carrying critical military data throughout a relatively small network of then-powerful computers. Serving as a critical element of military preparedness and civil defense, ARPANET was designed to function regardless of any kind of network disruption, no matter how cataclysmic. The original ARPANET has been replaced by a succession of bigger, faster, more powerful networked systems, leading to today's Internet. Sensing the potential of distributed computing for science discovery and application, the National Science Foundation created a network dedicated to academic research and development based on the ARPA's Internet protocol. Operating from a system of regional computing systems interwoven by an Internet "backbone," the Internet offered researchers immediate access to colleagues and files throughout the world.

The Internet has grown into a loose amalgam of individual networks and computer systems, all operating according to agreed-upon network protocols, allowing the worldwide transfer of information between systems built on a variety of operating platforms. Unlike many of the well-known commercial online network service providers (e.g., America Online, AT&T Broadband), no individual, company, agency, or institution owns the Internet. It is a "distributed" system of interconnected computers and networks. This means that no single "hypercomputer" or "supernetwork" drives it. Rather, it is an enormous, electronic global "neighborhood." Some "houses" are bigger and more luxurious than others, but each one arranges its own furniture and landscaping.

The Internet's origin may be traced back to an American sense of shock and fear caused by the 1957 launch of Russia's first-ever space orbiting satellite, Sputnik. The earliest Internet-type operation was launched in 1969. At the time, this network comprised four hosts (the University of California, Los Angeles; Stanford University; the University of California, San Bernardino; and the University of Utah). At its midlife (1982), the Internet had grown to more than 200 hosts. By 1995, there were almost 6 million hosts, and at the start of 2002, the number of hosts had grown to approximately 147 million.

Nearly 200 countries are now connected to the Internet (post-Taliban Afghanistan has recently become connected). When we add other networks capable of interfacing with the Internet in some manner or other, the number of nations involved increases still further. By 1995, the number of Web home pages had surpassed 20,000. As of March 2002, Zakon (2002) placed the number of Web servers at more than 38 million. Because most servers host multiple Web sites, the number of sites should be in the hundreds of millions.

Some noteworthy Internet benchmarks are illustrated in Table 1.

With such growth has come a change in purpose and tone. In its earliest incarnation, the primary Internet purpose was military. As the Internet moved beyond military use, its function shifted to research and scholarship. At the dawn of the 1990s, the Web had neither entered the national lay consciousness nor become viable for commercial transaction. The 1990s changed all that. In the industrially developed world, the Internet became an indispensable tool for all kinds of business, serving vendors and consumers alike. As a result, the charitable camaraderie that characterized a relatively small number of Web users in the early 1990s has now become primarily a commercial marketplace, where transactions for information, misinformation, financial services, transportation, products, education, romance, entertainment, culture, hate, and pornography are routinely carried out. From an earlier culture of open information sharing, a more defensive posture, to protect commercial and personal interests, has emerged. By the year 2000, the old neighborhood had changed for good.

With this change has appeared an ever-growing danger of mischief, intellectual property theft, and malicious destruction. Ever more devastating computer viruses and worms spring up faster than updated virus protection software can thwart them. Major institutional networks have been disabled for days at a time and sometimes longer by such viruses as *Melissa* (1999), *Love Letter* (2000), *Nimda* (2001), and *Code Red* (2001). Network hackers have broken into the presumably secure private networks of major corporations (e.g., Microsoft), transportation services (e.g., British Rail), and the governments of several nations (e.g., the U.S. Senate). Predators have harmed children. Meanwhile, a passionate debate rages between those who advocate legal restraints on putatively harmful Internet exploitation and those who value intellectual freedom above everything.

Globalization

Although a worldwide affiliation of groups vocally opposes economic and cultural globalization, Internet-based e-commerce has not only made common business transactions more accessible to global markets, it has also created exponentially richer international choices in education, entertainment, and culture. For example, classical music buffs may choose from nearly 100 radio stations

Table 1 Internet Time Line

Year	Event
1957	First orbiting space satellite, Sputnik, launched by the former USSR
1963	First dial-up modem developed
1969	Four-node ARPANET established with network speed of 50 kpbs
1972	First e-mail message sent using the @ sign; first computer chat undertaken
1976	First Internet routers developed; Queen Elizabeth II sends e-mail
1977	First transport control protocol (TCP) developed
1979	Emoticons ";-)" first appear in e-mail messages as antidote to dry text
1980	Accidental virus brings ARPANET temporarily to a halt
1982	TCP/IP protocol established as technical standard to drive Internet traffic
Mid-80s	Loose coalition of networks in ARPANET becomes known as Internet
1988	"Worm" infects 10% of all Internet hosts
1989	Creation of interconnected regional academic networks
1990	ARPANET decommissioned, leaving a loose network of networks, the Internet
1991	World Wide Web and Gopher are launched
1992	First Internet audio and video transmissions
1993	MOSAIC, the first GUI Web browser released; White House comes online
1994	First banner ads appear on the Web; first pizza ordered online
1995	Real-time audio streaming technology launched; Radio Hong Kong hits the Internet
1996	Internet becomes truly global; U.S. Communications Decency Act (CDA) passed
1997	CDA ruled unconstitutional; Microsoft challenges Netscape with Internet Explorer browser
1998	U.S. Postal Service sells stamps for downloading and printing from the Web
1999	Securities fraud on Web raises a small company's stock price by 31% in one day
2000	Catastrophic "denial of service" attacks launched against Amazon, eBay, and Yahoo
2001	Napster forced to suspend free service; resumes shakily as a subscription service
2002	Internet boasts more than 147,000,000 hosts (727,000 in 1992; 235 in 1982)

Sources: Public Broadcasting Service (1997), "Life on the Internet: Timeline," available at http://www.pbs.org/internet/timeline/timeline-txt.html; Zakon, R. H. (2002), "Hobbes' Internet Timeline," available at http://www.zakon.org/robert/internet/timeline; Verizon Corporation (2001), "Who We Are: BBN Timeline," available at http://www.bbn.com/timeline.

worldwide, many of them uninterrupted by annoying commercials. Such a shift in the network zeitgeist has created many employment opportunities and threatened others (e.g., travel agents). The Internet domain-naming system, only recently limited to six categories (.com, .edu, .gov, .mil, .net, and .org), has added several new categories to accommodate growth (e.g., .biz, .info, .aero, .name, .pro, .coop, .us, and .museum). Partly because of the inchoate spawning of commercial traffic, Internet2 represents an effort to regain the prominent place research and education formerly held on the original Internet.

INTERNET DEEMED INADEQUATE

Traffic

Although the present-day Internet grew out of the national defense need for uninterrupted communications, it soon became a popular medium for researchers and academics who wished to share research and data. Once considered the domain of researchers, with the introduction of the graphical user interface browser, the Internet quickly became usable for a wider audience. The introduction of Mosaic, which was quickly followed by Netscape and then Microsoft Internet Explorer, forever changed the face of the Internet. Rather than a medium used only by those who understood how to navigate a text-only protocol, the Internet became accessible to anyone who could operate a mouse and a keyboard. This increased usage quickly became a problem for a network system that was not originally developed to handle this increased amount of traffic.

One solution to the Internet traffic problem has been to increase bandwidth as well as the speed with which the data is sent. A number of universities have added T1 and T3 lines, and a new Internet protocol is being tested. The current Internet protocol, TCP/IP, uses a packet-switching protocol that Vinton Cerf (1995) likens to "sending a novel on many postcards." Although this may be appropriate for e-mail transmission, it is inadequate for the smooth transmission of audio and video files (Lan & Gemmill, 2000). The need to speed up the delivery of information has led to the creation of IPv6, which is currently being promoted and coordinated by the Internet2 IPv6 Working Group. This new protocol, Internet protocol version 6, allows files to be transported as one packet to many locations. More important, this same protocol promotes the collaborative participation envisioned by the original designers of the Internet. For instance, the technology behind multicasting enables a number of medical experts from remote locations to collaborate during time-sensitive diagnostic medical procedures. The Internet2 backbone networks, Abilene and VBNS+, use IPv6 as their Internet protocols.

The Needs of the Educational Community

The Internet has impacted the teaching and learning community in a very important way. Interaction between students and faculty through such online communication tools as e-mail, instant messaging, synchronous chat, and asynchronous forums has allowed the learning experience to move outside the walls of the classroom. Internet tools have given faculty members the opportunity to post online quizzes, handouts, lectures, and digital materials for easy access by students. The Internet has also provided the availability of online library materials that can be accessed by students and faculty any time of the night or day from their homes or offices. Even though these advances are significant, Internet2 visionaries suggest that a new generation of network-based applications necessitates the capacity to build a unique environment that "will engender new possibilities for the synthesis of ideas that can ultimately lead to the creation of important new knowledge" (Wasley, 1996). The current generation Internet cannot accommodate these needs.

Quality of Service

The Internet does not distinguish between low-priority e-mail messages and medical collaborative broadcasts where a delay in transmission could be critical. Additional congestion is caused by sending live video streams because the broadcast must be sent from point of origination to each separate requestor. Distance learning opportunities are hindered by both lack of speed and bandwidth. Internet2 researchers are working on these quality-of-service issues by considering issues of priority and multicasting, and by increasing the number of possible Internet addresses.

WHAT INTERNET2 IS

Vision

The essence of this new generation of network-based applications revolves around collaboration among individuals with access to data and resource materials. Internet2 member institutions have developed a number of new initiatives, including digital video for video conferencing and remote control instrumentation, multicast streaming designed to deliver video efficiently, media-rich collaborative and interactive learning environments in such areas as medical education, interactive and simulation-based teaching initiatives, tele-immersion programs designed for educational virtual reality systems, remote auditioning opportunities, and transmission of live, high-quality, uncompressed audio.

Infrastructure

Currently, two major networks support the research and advanced networking technologies envisioned by this consortium, by offering high-capacity connections. Abilene, known as the network backbone of Internet2, and vBNS+ networks connect universities to gigaPoPs (gigabit points of presence), which serve as network hubs independently sponsored by regional networking communities. Capable of reaching speeds approximately 200 times faster than DSL (digital subscriber line) and cable connections of today, the next generation network is connected through a series of robust regional aggregation points using the most advanced protocol for transmission of data.

Commercialization

The University Corporation for Advanced Internet Development (UCAID) is a nonprofit organization that supports Internet2 specifically for educational and research institutions. According to one of UCAID's six founding principles, member organizations wishing to provide Internet2 services to commercial organizations must offer these services separately. That is, Internet 2 network hubs, referred to as gigaPoPs, and other commercial network hubs must "remain appropriately distinct." The commercial organizations currently involved in the Internet2 project are those that have offered their services in the development of the project. Cisco Systems, 3Com, Advanced Network & Services, AT&T, IBM Corporation, Intel Corporation, ITC DeltaCom, Juniper Networks, Lucent Technologies, Microsoft Research, Nortel Networks, Qwest Communication, SBC Technology Resources, Spirent Communications, Sun Microsystems, WorldCom, and Yotta Yotta are currently listed as corporate partners of the Internet2 consortium. Such partnerships have contributed to the development of security measures against potentially harmful network infiltration, have connected with educational institutions to support technologically advanced communications, have granted access to information, and have provided strong funding opportunities.

Security

Security on the traditional Internet has continued to defy surefire solutions. To address this issue, Internet2 member organizations are involved in a number of working groups. Some of these groups are working on middleware, which is described on the Internet2 Web site as the "glue (or the) layer of software between the network and the applications" (UCAID, 2002i). Participants in such working groups as VidMid (Internet2 Video Middleware Group) are focusing on common middleware solutions to security challenges. VidMid's February 2002 draft, *A Framework of Requirements, Threat Models, and Security Services for Videoconferencing over Internet2*, outlines its plans for security initiatives related to video conferencing tools and applications. Among the considerations outlined in this draft are issues related to theft of service, impersonation of servers, authentication of users, firewall issues, and encryption.

THE INSTITUTIONAL STAKES OF MEMBERSHIP

Once an institution becomes a member of Internet2, the stakes of membership are high. Representative groups within the research, academic, and corporate communities may need to take on various scholarly, ethical, financial, and instructional responsibilities within the Internet2 consortium.

The Stakeholders

Internet2 is an organization driven by its membership, which includes for-profit and not-for-profit organizations (UCAID, 2002h), including those in the research and corporate communities.

The Research Community

Individual stakeholders within the research community may include pure researchers, professors, information technology (IT) professionals, and research and education administrators. It is these groups who will facilitate and lead the Internet2 research and educational initiatives emanating primarily from universities. Available new technologies in research, teaching, and learning are forcing the creation of cutting-edge network capabilities (UCAID, 2002e). As a result of these advanced networking capabilities, researchers are developing new sets of teaching and learning applications (Lan & Gemmill, 2000).

To assist institutions in meeting these research and educational goals, the National Science Foundation, through its High Performance Network Connections (HPNC) program, has committed millions of dollars in grant funding for American higher education and other institutions with significant research and education missions "to establish high performance Internet connections (at or above 45mbits per second) to support cutting edge science and engineering research" (NSF, 2002).

Members of the research community who wish to keep abreast of Internet2 priorities and initiatives can look to the UCAID Applications Strategy Council (ASC) and the Network Research Liaison Council (NRLC) for current information. The ASC helps identify priorities for research and education. NRLC helps identify priorities related to adapting computer systems to fit the Internet2 infrastructure, including research, access, and prototype deployment (UCAID, 2002a).

The Corporate World

Corporate stakes in Internet2 can be more or less substantial depending on the type of membership. Corporations may agree to contribute goods, equipment, services, cash, grants, human resources, and capital ranging from $100,000 to $1,000,000 for their participation in the Internet2 project (UCAID, 2002c; O'Beay, 2002). Annual fees for corporations range from $10,000 to $25,000, depending on their total annual revenue.

Fortune 500 companies, such as IBM, Quest Communications, Lucent Technologies, and AT&T, have formed the foundation of Internet2 corporate memberships. However, the three levels of corporate membership (partner, sponsor, and member) have allowed content providers and research organizations, such as CREN and EDUCAUSE, to become Internet2 members. Members of the corporate community who wish to keep abreast of Internet2 priorities and initiatives can look to the UCAID Industry Strategy Council (ISC). The ISC provides a strategic vision related to advanced networking and applications development and helps focus technology transfer aspects of UCAID initiatives (UCAID, 2002a).

Higher Education

When a college or university commits to Internet2 membership, it pledges to provide a significant financial investment to the project. Institutions pay an annual $25,000 fee to be a regular member of Internet2. Everything considered, existing members typically estimate they will spend a minimum of $500,000 per year for their participation in Internet2 (UCAID, 2002k). This annual expense varies, depending on the current network infrastructure of the institution (UCAID, 2002k). At this level of financial support, applicants and members must be aware that in order to conduct cutting-edge research, they must make a substantial and continuing commitment to experimentation with, and to the evolution and use of, advanced networking technologies and applications (UCAID, 2002k). To help institutions with planning and decision making, the UCAID Network Planning and Policy Advisory Council (NPPAC) advises on matters related to the planning, development, financing, and management of advanced networks for research and education (UCAID, 2002a).

Although there are many grants available through federal sources, the Experimental Program to Stimulate Competitive Research (EPSCoR) is one program that some states may find particularly helpful (EPSCoR, 2002a). EPSCoR supports partnerships between the National Science Foundation and many states. Its goal is to leverage the potential of the states' science and technology funds to develop research infrastructure and to advance economic growth (EPSCoR, 2002b).

NUTSHELL SKETCHES ON WHO IS USING INTERNET2

The Academic Community and High-Definition Television

Since early 1999, the University of Washington, one of the original founders of the Internet2 initiative, has been working on the transmission of high-quality and high-speed video transmission using high-definition television (HDTV). At its fall 1999 Internet2 membership meeting, the University successfully demonstrated the transmission of a high-definition television stream from Stanford University to the University of Washington at 270 Mbits per second; approximately 5000 times faster than the speed of a 56K modem. Since that time, the researchers involved in this project have successfully sent multiple HDTV streams simultaneously. The University of Washington researchers credit the success of these experiments to the unique collaborative efforts of individual experts in the networking, computing, and video fields. The researchers believe that their reliance on industry standards has positive implications for future use by a number of interested parties, including those in the broadcast television industry.

The Academic Community and Video Conferencing

The University of Oklahoma has launched a program for teaching music through advanced network video conferencing. The increased bandwidth offered by the Internet2 backbone allows for a quality of sound and video transfer that was not available on the first-generation Internet. Rather than offering an alternative method to face-to-face teaching, music educators envision bringing world-class musicians to their courses, expanding their own teaching to new constituencies, increasing collaborative and research opportunities for faculty, and increasing outreach programs throughout the state of Oklahoma.

The Scientific Community and Distributed Computing

The Gemini Observatory's Internet2 connection offers astronomers the opportunity to use twin 8.1-m telescopes located in Hawaii and Chile to observe both the northern and the southern hemisphere skies from remote networked locations. An international partnership managed by the Association of Universities for Research in Astronomy, Gemini has a cooperative agreement with the National Science Foundation.

The Scientific Community and Tele-immersion

The Electronic Visualization Laboratory (EVL) (2002) at the University of Illinois at Chicago defines tele-immersion as

> collaborative virtual reality over networks, an extension of the "human/computer interaction" paradigm to "human/computer/human collaboration," with the computer providing real-time data in shared, collaborative environments, to enable computational science and engineering researchers to interact with each other (the "teleconferencing" paradigm) as well as their computational models, over distance.

Tele-immersions allow virtual excursions through the human body, machines, and buildings, as well as virtual visits to historical sites and to "environments that explore truths about irony, humor, music, and levels of consciousness" (Lemley, 2002), in which one can participate in such virtual physical experiences as jumping off a set of stairs. EVL's goal is to move their laboratory from the desktop to the next-generation Internet so that they can share their expertise with remote users and form a collaborative to encourage cooperative problem solving and development.

ADVANTAGES AND DISADVANTAGES OF INTERNET2

Access to Information

For Internet2 members, the potential for accessing large amounts of data and information is enormous. Through partnerships and collaboration, Internet2 applications, middleware, and engineering initiatives are all focused on increasing access to information for teaching, learning, research, and clinical missions of member institutions. Although this is certainly an opportunity for Internet2 members, those who are not Internet2 members do not have the same access to such information. This, in a sense, is creating a gulf between those who can only access the original Internet and those who can access Internet2—an academic digital divide of sorts. There are opportunities to bridge this gap, which include becoming a sponsored participant or a sponsored educational group participant (SEGP). A member university may sponsor organizations that are not eligible for Internet2 membership. Such organizations may include "educational institutions (including both not-for-profit and for-profit K-20, technical, and trade schools), museums, art galleries, libraries, hospitals, as well as other non-educational, not-for-profit or for-profit organizations" (UCAID, 2002f). The opportunity to become a sponsored participant or a SEGP allows the sponsored organizations to use Internet2 to collaborate with primary members or other sponsored members.

Although the potential for accessing information via Internet2 is great, in some disciplines there remain challenges to accessing the needed information efficiently. The health sciences and particle physics disciplines have unique challenges in that the computational research and subsequent data output create huge data sets. For faculty, such as those at Tufts University School of Medicine, Internet2 is helping to address access to information not previously available via traditional networks in health sciences, such as high-resolution magnetic resonance imaging and confocal microscopy.

On the other hand, in areas such as particle physics, some researchers believe that Internet2 will not be quite fast enough to meet their needs, and they await the creation of Internet3, which may be able to handle the amount of data generated by the computational outputs of particle physics research. At this time, these scholars continue to focus on creating simulations because there is still not a fast enough network to handle the petabytes (2^{50} bytes) of particle physics data.

Research Communication

The Internet2 philosophy is based on Metcalf's law, which states that the "value" or "power" of a network increases in proportion to the square of the number of nodes on the network (Robertson, 2002). Thus, the added value of Internet2 rests in the degree of collaboration among its members. Although Internet2 represents many disciplines, such as math, humanities, science, and engineering, some disciplines, such as social sciences, are not as well represented.

In the health sciences, Internet2 has allowed advanced research in instructional content and visualization (e.g., visible human project), simulations (e.g., virtual labs projects), and distance learning (e.g., virtual grand rounds and video conferencing); however, one overarching challenge in the health science area is the issue of network security and patient confidentiality. Policies and guidelines to these challenges are being identified by various health sciences Internet2 working groups.

Infrastructure Speed and Robustness

A particularly promising opportunity for Internet2 members is the speed at which they can now collaborate with colleagues and students around the world. In a recent Internet2 intercontinental land-speed record, "6.7 gigabytes of data was transferred across 6,800 miles (10,978 km) of network in less than one minute." The speed of this transfer was "923 megabits per second, or more than 3,500 times faster than a typical home broadband connection" (Wood, 2003).

A challenge that arises when we look at infrastructure speed and robustness is the issue of storage capacity.

Faster networks generate massive increases in data transmission, requiring comparable hardware improvements to store these data so they can be accessed and analyzed in an efficient manner.

Potential for Distance and Distributed Learning

One significant advantage of Internet2 is that it is changing the way professors teach and students learn. Faculty are enhancing their traditional face-to-face classrooms with real-time collaboration tools, simulations, and H.323 video conferencing. H.323 is an International Telecommunication Union (ITU) standard that allows transmission of Internet video conferencing over IP-based networks. Human sciences and arts have found tremendous opportunities in Internet2 that they do not have with the conventional Internet. For example, Internet2 facilitates access to full-motion video and high-fidelity audio. Without traveling, theatre directors can conduct remote auditioning using real-time video conferencing. Master musicians around the world can be present via video conferencing to conduct live master classes and teaching sessions with young musicians. These video conferencing opportunities have been made possible by Internet2 because the challenges of sample rates, frequency, and latency have been overcome to allow effective music education over the Internet.

As faculty members begin to expand their teaching techniques and adopt instructional materials that require advanced networks, comprehensive faculty development programs and corresponding support resources need to be in place. At individual institutions, adequate staff and resources are necessary to support such robust undertakings in teaching and learning. Faculty development and outreach initiatives can include training grants, online discussion forums, focus groups, desktop computer conferencing, structured workspace sharing, and consulting (online and face to face), typically in combinations rather than in single strategies.

Costs of Access and Infrastructure

Because the role of Internet2 is to promote networked teaching and bring scholarly research and development to fruition, the question of costly access to resources begs for attention. The capacity of some institutions to join the Internet2 community provides an opportunity that will allow these institutions to be at the forefront of advanced networking initiatives. To address this challenge of cost, institutions are now exploring various funding and cost-sharing models that may assist in defraying some of the initial and ongoing costs associated with being an Internet2 member. If the cost of membership is still prohibitive for smaller organizations, there are grant monies available to help offset the costs of membership and subsequent infrastructure upgrades. There are also other opportunities for Internet2 sponsorship agreements.

One advantage of being an Internet2 member is that members can obtain loaner equipment from the Internet2 applications working group. Members can request loaner equipment, such as portable access grid nodes, various types of video conferencing equipment, and preinstalled hardware with visual, immersive, and digital library applications.

HOW EXPERIENCE IS SHAPING INTERNET2

As it pioneers the meta-network of the future, as new research unfolds, Internet2 is taking precautions to develop appropriate policies and guidelines. Based on the experience of member institutions, as new information and knowledge is disseminated, current user policies governing Internet2 are reviewed and updated. To generate cutting-edge experience, Internet2 has announced a call for participation to its members. The call for participation asks members to propose "compelling, exemplary advanced content applications that will be valuable to the mission of member universities" (UCAID, 2002b). As a result of participation in such projects, Internet2 hopes to identify best practices for developers and applications management and monitoring and ways in which these experiences can support the teaching and research missions of member institutions.

Essential Partnerships

Thanks to close collaborations between academia and industry, Internet2 research and development initiatives have been made possible. At the core of these Internet2 collaborations are the partnerships formed to create Abilene, the network backbone of Internet2. In February 1999, Cisco Systems, Nortel Networks, Quest Communications, and Indiana University launched the Abilene project in the United States. In April 2002, Juniper Networks partnered with Abilene. A fall 2002 upgrade equips the U.S. Abilene network to use a routing product produced by Juniper Networks that can quadruple the speed of Internet2.

Other essential connections are those between member institutions and the Abilene backbone or gigaPoPs. Through their connections to regional and national networks, member research and development efforts are being joined. For example, the mid-Atlantic crossroads (MAX) is a gigaPoP in the Washington, D.C., area that connects to the Abilene network and to other national and international networks. Many universities, federal research laboratories, and commercial and nonprofit institutions in Washington, D.C, Maryland, and Virginia connect to this regional gigaPoP.

One example of a project utilizing MAX is that of the University of Maryland Space and Plasma Physics Group. This group uses MAX to compute large simulations at the Pittsburgh (PA) Supercomputing Center and the National Center for Supercomputing Applications at the University of Illinois at Urbana—Champaign. The computational results and graphical outputs of such massive data processing are shared collaboratively through MAX and the Abilene network with researchers at Dartmouth College and NASA. For institutions to join the MAX network and to benefit from its supercomputing capacities, an annual subscription fee of $12,000 is levied, with additional annual costs ranging from $45,000 to $180,000, depending on the institution's choice of connectivity and bandwidth.

Essential partnerships and research and development initiatives revolve around five technical areas of Internet2: applications, middleware, engineering, end-to-end performance, and backbone network infrastructure. Table 2 depicts research interests, working groups, and advisory councils within each of the five technical areas. These five technical areas rest on the foundation of a high-performance network where ideas and products can be evaluated and tested. Development initiatives and activities within each of the five technical areas are conducted in committees known as working groups (UCAID, 2002g). Each technical area has a director who oversees the activities of all related working groups and reports working group developments and highlights to the president of UCAID. In addition to these working groups, members can also participate in Internet2 activities through Sponsored Interest Groups (SIGs) and Birds of a Feather (BoF). UCAID also receives strategic and advisory guidance from the UCAID advisory councils on various matters related to advanced networking in higher education (UCAID, 2002b). The UCAID board of trustees elects council members, and each council is limited to 14 individuals; however, member institutions and other interested parties can keep abreast of council activities and council meeting minutes on the UCAID Web site.

Life-Long Learning

The Internet of today has made education available to students in new ways. Distance education, as it is called, allows students of all ages the convenience and flexibility of learning from their homes or offices. With such advanced applications as tele-immersion, digital libraries, and virtual laboratories, the possibilities for life-long learning are limitless. AccessGrid, for example, is a project of the Argonne National Laboratory Futures Lab and the Office of Computational and Technology Research of the U.S. Department of Energy. Through the use of new collaborative technologies, such as virtual reality environments, tele-immersion, and advanced two-way audio and video technology, AccessGrid supports collaborative teamwork sessions, seminars, lectures, tutorials, and training.

Sharing Intellectual Resources

Sharing intellectual resources is a cornerstone of Internet2 activity. Just as Internet2 is creating new applications, researchers are also inventing new middleware that will facilitate an unprecedented sharing of intellectual resources. At the core of the middleware initiative is the creation of directories that are known as the "operational linchpin" of Internet2 information resource sharing (UCAID, 2002d). Creation of these directories will enable a network security layer, allowing access to resources, people, and data within the Internet2 collaborative.

The goal in creating such directories is to "deploy a common middleware infrastructure to support the academic and administrative needs of the research and education community" (UCAID, 2002j). Directory projects that are currently under development include EduPerson and a Directory of Directories for Higher Education (DoDHE). These directories promote standardization in how applications and data are shared between institutions, with the aim of developing core middleware services allowing Internet2 members to communicate easily and quickly. These directories are the "first step in building an environment that opens new dimensions for sharing information across organizations" (EDUCAUSE, 2001).

With significant collaboration and information sharing across networks and between various academic, government, and corporate organizations, Internet2 and UCAID have recently approved an intellectual property framework. This framework addresses intellectual property, including hardware and software design, text, applications, documents, and innovations developed in connection with Internet2 activity.

CASE STUDIES

Case One: Internet2 and Sign Language Transmission

Gallaudet University, well known for its mission to educate a community of hard of hearing, deaf, and hearing students, has been a member of Internet2 since 1998. Cynthia King, Executive Director of Academic Technology, explains that Gallaudet has "long recognized the importance of a high-performance network—especially for sign language applications" (personal communication, May 25, 2002). Grants from IBM in 1998 and the National Science Foundation in 2001 have been used to explore and test video conferencing capabilities related to sign language transmission.

The traditional Internet has proved inadequate for the smooth flow of sign language streaming. Since April 2002, Gallaudet and the University of Tennessee—Knoxville have been conducting video conferencing experiments over Internet2. As King explains, "we are using the Delco Box, which is a Motion-JPEG system capable of 16-meg per second. This is about 20 times the typical best performance of H.323 video conferencing" (personal communication, May 25, 2002). Patrick Harris, Media Technical Designer at Galluadet, suggests that this advanced networking capability will "result in clearer high frame rate sign communication" (White, 2002).

As part of a Preparing Tomorrow's Teachers to Use Technology grant, education faculty have used the Internet2 connection to collaborate with other educators around the country. Thus far, the involvement of students has been limited; however, Dr. King explains that there are plans to involve students and faculty in the 2002–2003 academic year with the opening of a "new technology-rich" student academic center, which will commence the "official kick-off of the project where the Internet2 connection will be used more exclusively" (personal communication, May 25, 2002). A number of projects have been planned with a focus on science and education. Among these projects are the following: (a) some delivery of courses to schools in Pennsylvania that teach children in prekindergarten through 12th grade; (b) remote interpreting work with Gallaudet students who attend area universities for selected courses; (c) work with Gallaudet's regional centers to deliver instruction over the network; and (d) the continued Web casting of various Gallaudet events but with greater quality provided by the Internet2

Table 2 Research Interests, Working Groups, and Advisory Councils for Internet2 Technical Areas

Technical Area	Areas of Research	Working Group (WG), Sponsored Interest Groups (SIGs), Birds of a Feather (BoF) and Initiatives	Advisory Groups (AG) & Councils
Applications	Tele-immersion	Arts and Humanities WG	Application strategy council
	Virtual labs	Internet2 Commons WG	Health Sciences
	Digital libraries	Digital Video WG	Initiative AG
	Distributed instruction	Multicast WG	Performance Events
		Voiceover IP WG	Advisory Committee
		Geospatial WG	
		Medical Middleware WG	
		Orthopedic Surgery WG	
		Veterinary Medical SIG	
		High Energy and Nuclear Physics SIG	
		Network Storage SIG	
		ResearchChannel SIG	
		Remote Instrumentation BOF	
		Very Long Baseline Interferometry (VLBI) BOF	
		Health Sciences Initiatives	
Middleware	Middleware integration	Middleware Architecture Committee for Education (MACE)-Shibboleth WG	Network Research Liaison Council
	Core middleware	MACE-WebISO WG	Directories Technical
	Identifiers	MACE-Dir (Directories) WG	Advisory Board
	Authentication	VidMid-Video Conferencing (VC) WG	Federating
	Directories	VidMid-Video on Demand (VoD) WG	Organizations
	Authorization	HEPKI (Higher Education PKI)-Policy Activities Group (PAG) WG	Organization AG
	Certificates and Public Key Infrastructure (PKI)	HEPKI-Technical Activities Group (TAG) WG	HEPKI AG
	Upper middleware	OpenSAML (Open Source Security Assertion Markup Language)	MACE Advisory Board
		Medical Middleware (MedMid)	PKI Labs AG
		Multicampus Middleware	
		Early Harvest	
		Early Adopters	
		Identifier Mapping	
		DoDHE	
		EduPerson	
		LDAP Recipe	
		Secure Multipurpose Internet Mail Extensions	
		(S/MIME)	
		PKI Labs	
		Related Projects & Organizations	
		Federal PKI Technical WG	
		Net@EDU PKI WG	
		Coalition for Networked Information (CNI)	
		The Corporation for Research and Educational Networking (CREN)	
		The Common Solutions Group (CSG)	
		Global Grid Forum	
		Globus I2-DSI (Distributed Storage Infrastructure)	
		Legion	
		Open Knowledge Initiative	
			(*cont.*)

Table 2 (Continued)

Technical Area	Areas of Research	Working Group (WG), Sponsored Interest Groups (SIGs), Birds of a Feather (BoF) and Initiatives	Advisory Groups (AG) & Councils
Engineering	IPv6	Campus Bandwidth Management WG	Network Planning and Policy Advisory Council
	Measurement	IPv6 WG	Industry Strategy Council
	Multicast	Multicast WG	
	Quality of Service	Security WG	
	Routing	Topology WG	
	Security	Measurement SIG	
	Topology	Quality of Service (QoS) SIG	
		Routing SIG	
Backbone Network Infrastructure		IP Optical WG	
End-to-End Performance (E2Epi)		Peer-to-Peer WG	E2Epi Technical AG

Sources: Internet2 (2003a). Applications. Retrieved February 18, 2003 from http://apps.internet2.edu/; Internet2 (2003b). Internet2 Working Groups, Sponsored Interest Groups and Advisory Groups. Retrieved February 18, 2003 from http://www.internet2.edu/working-groups.html; Internet2 (2003c). Middleware. Retrieved February 18, 2003 from http://middleware.internet2.edu/.

connection. Dr. King further explains that there will be a continuation of faculty participation in curriculum and video-enhanced learning projects, with assistance from the academic technology staff, who will be instrumental in producing the digital media once the faculty have identified their curricular and video needs. "These related efforts will continue and be expanded in light of the new technologies we have available with Internet2" (personal communication, May 25, 2002).

A number of challenges remain. The biggest challenge, however, relates to the quality of video for sign language. Even with the advanced capabilities available through Internet2 networks, quality of service (QoS) issues need to be addressed. "We are interested in QoS issues, but understand that solutions won't be occurring in that space for some time" (personal communication, May 25, 2002).

Case Two: Crime & Punishment

In 1996, Kent Portney, Professor of Political Science at Tufts University, wished to provide his students with a better understanding of variability in criminal sentencing. He wanted his students to complete his undergraduate course on judicial politics and know how to isolate factors to explain variability in judicial sentencing. With the help of a FIPSE grant, he created a computer-based simulation entitled *Crime & Punishment,* in which the students act as judges. After being presented with real-life information about cases, the students sentence the defendants before them.

One of the challenges that presents itself when teaching variability of criminal sentencing in the classroom is that students have preconceived ideas about what causes variations in judicial sentencing. The goal of the *Crime & Punishment* simulation is to change those preconceived notions and help students to understand why judges sentence the way they do and to help students examine their own prejudices (Rosenberg, 1997).

Now in its third generation, the project has come a long way. Having Internet2 on the Tufts University campus allows quick access to the video in the simulation without loss of performance in displaying the video. Making this project available to students over Internet2 also accommodates several users simultaneously, without loss of speed or performance of the video. The current generation of this project allows all user inputs and case information to be stored in a database, which provides Portney access to his students' selections and sentencing decisions. This is critical when reporting information back to the students. Portney can manipulate the data received in the database and create graphs that represent students' choices and variability decisions based on the characteristics of the defendant.

Future generations and enhancements of this simulation are planned that will make full use of the collaborative nature of Internet2. Enhancements will make it accessible to other faculty members and allow all faculty to access the database of student decisions and compare their courses to other courses in this and other institutions. A longer term enhancement will allow other faculty to modify the simulation by adding their own video or cases, which would then become part of a large digital library for all users of the *Crime & Punishment* simulation.

Another significant achievement of this project is shown by the results of the learning assessments. Substantial learning effects have been shown as a result of students' interaction with this computer-based simulation. Students' misconceptions have been corrected because of their experience. The result is that students begin to look for more substantial and theoretical reasons for variation, focusing on theories of punishment and sentencing. As a result of using this simulation, Portney reports that student response to this simulation is "very well received and it is the highlight of his class" (personal communication, May 20, 2002).

As with any new-technology-based project, technical challenges are part of the territory. This project is no exception. The developers of *Crime & Punishment* struggled in the early stages to develop a model for capturing the real-life subjects in a video format. They have since developed a model that is efficient and that meets their needs. To optimize achievement of their enhancement goals, working with current Internet browsers and media players (such as RealPlayer), presents the challenge of maintaining control over what the end user experiences.

Although many years of research and development have gone into this successful project, Portney cautions that supplemental funding (such as this project's FIPSE grant) needs to be lined up before undertaking such an initiative and that it would be unacceptably frustrating to develop such highly interactive procedures if the appropriate resources were not in place.

CONCLUSION: PLAUSIBLE FUTURES FOR INTERNET2

Ted Hanss, Director of Applications Development for Internet2, suggests that "Internet2 networks are like a time machine, anticipating what may be possible in the future when everyone has access to high-speed broadband connections at home, school, and work" (Hanss, 2001). Such cutting-edge networks are also being planned by academic, government, business, and research communities worldwide. Canada, Europe, and Australia are developing advanced networks, such as CA*net3, Geant, and AARNet, making it possible to connect researchers from remote universities to resources ranging from real-time medical consultation to robust data sets. Such networks also have the potential to expand the capabilities of distributed computing by sharing work, space, and processing power to complete computations that are too sophisticated for even the current complement of supercomputers to handle.

In parallel with the Next Generation Internet (NGI) project sponsored by the U.S. government, Internet2 is focused on developing networking capabilities that will expand well beyond the capacity of the current Internet. NGI has significantly affected the Internet2 projects by directly funding connections between a number of academic institutions and the advanced networks of Abilene and vBNS+. These powerful connections have the potential to bring education to higher levels as institutions work together to develop new learning tools, collectively seek solutions to security issues, and strategically plan for future initiatives.

GLOSSARY

AARNet The Australian academic and research network; operates AARNet2 network, which provides Internet service to all 37 of Australia's universities.

Abilene An advanced backbone network that connects regional aggregation points (gigaPoPs) and supports the work of Internet2 universities.

CA*net3 The advanced Canadian research and education network developed by CANARIE.

Geant The pan-European academic and research network, reaching over 3,000 research and education communities in over 30 countries.

Gigabit Point of Presence (gigaPoP) A regional internetworking aggregation point used for accessing the Internet2 network.

H.323 An International Telecommunication Union (ITU) standard that allows transmission of Internet video conferencing over IP-based networks.

Internet Protocol version 6 (IPv6) Called the next version of the Internet protocol replacing Internet Protocol version 4, which is currently used by the current Internet.

Middleware The layer of software between the application and the Internet2 network, which provides services such as identification, authentication, and security.

University Corporation for Advanced Internet Development (UCAID) A nonprofit consortium of university members working in conjunction with corporations and others to provide leadership for advanced networking development.

vBNS+ (Backbone Network Service Plus) A very-high-performance, Internet2 backbone network.

CROSS REFERENCES

See *History of The Internet; Internet Literacy; Internet Navigation (Basics, Services, and Portals); Standards and Protocols in Data Communications.*

REFERENCES

Cerf, V. G. (1995). *How the internet works, part I. The Cerf Report archives.* Retrieved May 22, 2002, from http://www1.worldcom.com/global/resources/cerfs_up/prose/hownetworks.xml

Computer Industry Almanac (2002). *Internet users will top 1 billion in 2005. Wireless internet users will reach 48% in 2005.* Retrieved May 22, 2002, from http://www.c-i-a.com/pr032102.htm

EDUCAUSE (2001). *Key component of higher education directory service released: New standard facilitates access to applications and resources across higher education.* Retrieved May 22, 2002, from http://www.educause.edu/news/2001/02/eduperson.html

Electronic Visualization Laboratory (2002). *Research: Tele-immersion.* Retrieved May 22, 2002, from http://www.evl.uic.edu/research/telei.html

EPSCoR (2002a). *Home page.* Retrieved May 22, 2002, from http://www.ehr.nsf.gov/epscor/

EPSCoR (2002b). *State programs.* Retrieved May 22, 2002, from http://www.ehr.nsf.gov/epscor/state_program/about/start.cfm

Hanss, T. (2001). Digital video: Internet2 killer app or Dilbert's nightmare? *Educause Review, 36* (3), 17–25.

Houweling, D. V. (1999). *The launching of UCAID and Internet2. CREN TechTalks*. Retrieved May 22, 2002, from http://www.cren.net/know/seminars/trans/i2/launch.html

Internet2 (2003a). Applications. Retrieved February 18, 2003, from http://apps.internet2.edu/

Internet2 (2003b). Internet2 Working Groups, Sponsored Interest Groups and Advisory Groups. Retrieved February 18, 2003, from http://www.internet2.edu/working-groups.html

Internet2 (2003c). Middleware. Retrieved February 18, 2003, from http://middleware.internet2.edu

Lan, J., & Gemmill, J. (2000). The networking revolution for the new millennium: Internet2 and its educational implications. *International Journal of Educational Telecommunications, 6,* 179–198.

Lemley, B. (2002). Internet2: *A supercharged new network with true tele-presence puts the needs of science first*. Retrieved May 22, 2002, from http://www.discover.com/may_02/featinternet2.html

National Science Foundation (2002). *High performance network connections*. Retrieved May 22, 2002, from http://www.nsf.gov/pubs/2002/nsf02073/nsf02073.html

O'Beay, A. (n.d.) *UCAID/Internet 2 corporate relations*. Retrieved May 22, 2002, from http://www.internet2.edu/presentations/Corp-Ann/Corp-Ann.ppt

Public Broadcasting Service (1997). *Life on the Internet: Timeline*. Retrieved May 22, 2002, from http://www.pbs.org/internet/timeline/timeline-txt.html

Rosenberg, S. (1997). Crime & Punishment. Retrieved February 17, 2003, from http://www.tufts.edu/tccs/services/css/crimeandpunishment.html

Robertson, J. (2002). *Metcalf's law*. Retrieved August 9, 2002, from http://www-ec.njit.edu/~robertso/infosci/metcalf.html

UCAID (2002a). *Advisory councils*. Retrieved May 22, 2002, from http://www.ucaid.edu/ucaid/html/councils.html

UCAID (2002b). *Collaboration in innovation approaches to content delivery over Internet2 networks*. Retrieved May 22, 2002, from http://www.internet2.edu/html/cfp-contentcollab.html

UCAID (2002c). *Corporate membership*. Retrieved May 22, 2002, from http://www.ucaid.edu/ucaid/html/corporate.html

UCAID (2002d). *Directories*. Retrieved May 22, 2002, from http://middleware.internet2.edu/core/directories.shtml

UCAID (2002e). *Internet 2 corporate opportunities and benefits*. Retrieved May 22, 2002, from http://www.internet2.edu/members/html/corpbenefits.html

UCAID (2002f). *Internet2 and terms of affiliation*. Retrieved May 22, 2002, from http://www.internet2.edu/members/html/terms_of_affiliation.html#SponsoredPart

UCAID (2002g). *Internet2 working groups*. Retrieved May 24, 2002, from http://www.internet2.edu/html/working-groups.html

UCAID (2002h). *Membership information kit*. Retrieved May 22, 2002, from http://www.internet2.edu/resources/Internet2_Infokit.PDF

UCAID. (2002i). *Middleware*. Retrieved May 22, 2002, from http://middleware.internet2.edu/

UCAID (2002j). *Middleware architecture committee for Education*. Retrieved May 22, 2002, from http://middleware.internet2.edu/MACE/

UCAID (2002k). *Regular membership*. Retrieved May 22, 2002, from http://www.ucaid.edu/ucaid/html/regular.html

Verizon Corporation. 2001. *Who we are: BBN timeline*. Retrieved May 22, 2002, from http://www.bbn.com/ timeline

Wasley, D. L. (1996). *The focus of the internet ii vision*. Retrieved May 22, 2002, from http://www.ucop.edu/irc/wp/wp_Docs/wpd004.html

White, S. (2002). *Academic technology news and events: Internet2 sign communication*. Retrieved May 25, 2002, from http://academic.gallaudet.edu/events.nsf/ID/C4A2E3ABADABDC1985256B6D005EE0EC?OpenDocument

Wood, G. (2003). *New intercontinental Internet performance records set in Internet2 land speed record competition. Message posted to the I2-News ListProc Listserv*. Retrieved February 16, 2003, from http://archives.internet2.edu/guest/archives/I2-NEWS/log200301/msg00005.html

Zakon, R.H. (2002). *Hobbes' internet timeline v5.6*. Retrieved May 22, 2002, from http://www.zakon.org/robert/internet/timeline

Intranets

William T. Schiano, *Bentley College*

INTRODUCTION

An intranet is defined by the use of Internet technologies (HTTP, TCP/IP, FTP, SMTP) within an organization. By contrast, the Internet is a global network of networks connecting myriad organizations. The line becomes blurred when the internal system is opened to remote access and parts of the system are made available to customers and suppliers. This extension of an intranet to selected outsiders is often called an extranet. As networks continue to develop, distinctions among types of systems will become increasingly artificial and contrived, just as the once clear distinctions among hardware such as personal computers, servers, and minicomputers have lost meaning. When reading this chapter, concentrate on how the issues described may affect Internet technology-based systems within your organizations, rather than on the specific definition of any one system.

Because they are based on open Internet standards, intranets are easy to implement technologically. This can be a blessing and a curse. Although they can offer robust functionality with little investment, they are often rolled out with little forethought and therefore fail to make a significant contribution to the organization. This chapter outlines what intranets are; how they are used; the technologies for constructing and running them; how to implement, secure, and maintain them; and how to make them more efficient.

Much of the research and background information applicable to intranets is now being published in other areas. The technological issues overlap with the broader Internet, and many of the internal applications offered on intranets are now encompassed by enterprise applications and knowledge management. Some authors even refer to business-to-employee (B2E) applications (Hansen & Deimler, 2001), emphasizing the importance of focusing on the access to systems by those within the organization. This chapter draws on these and other literatures to offer a broad perspective on intranets and their application.

FEATURES OF AN INTRANET

Few organizations question the value of having an intranet in some form. Indeed, over 90% of major U.S. corporations have intranets (Baker, 2000). The benefits can range from a reduction of paper and headcount, always having the current version of any data or document available, to a central interface, or portal, and repository for all corporate systems and data. Intranets may encompass many different types of content and features. Other benefits include the use of Internet-standard network protocols that facilitate connections to the broader Internet.

Portal

> A door or gate; hence, a way of entrance or exit, especially one that is grand and imposing. (*Webster's Revised Unabridged Dictionary*, 1913)

Intranets are often designed to be "portals," serving as the central point of access to all information resources within an organization. The implication of grandiosity is deliberate and helpful, as good design can be a crucial determining factor in the success of an intranet. A portal also needs to be available regularly to a wide community, with access to multiple sources of data, both internal and external, with a useful search mechanism. The term portal often evokes third-party commercial portals such as Excite.com (http://www.excite.com) and vertical portals, also called vortals, such as aluminium.com (http://www.aluminium.com). These portals have struggled to establish successful business models, especially given the collapse of Internet advertising rates. Portals inside organizations, in contrast, have thrived, often with the support of the information systems (IS) department and the business users.

One of the greatest frustrations for information technology users inside organizations is the necessity of multiple systems with varied interfaces, including command-line based interfaces. With a portal, users have the ability to go to a single location, use a common, intuitive, and well-established graphical interface, and find their information. Types of the information and services found on intranets are outlined below.

Human Resource Materials

Human resource materials are an obvious and common type of content for intranets. Benefits management is a significant cost for organizations; tax withholding, health care election choices in the United States, and retirement and pension accounting all involve substantial processing costs. These processing costs include printing, distributing, completing, collecting, and entering data from paper forms, providing support to fill out the forms, and managing subsequent changes. With a well-designed intranet, employees can process their own benefits, saving substantial overhead and time. The forms tend to be highly structured, with well-defined fields, and the application can be scaled to all employees, reducing the cost of development per employee. These forms can also change regularly. Changing the online versions is much less expensive, and eliminates the possibility of employees filling out the wrong version of a form.

Human resources is also an excellent choice for intranet applications, particularly early in an intranet rollout, because every employee uses its services. Beginning with a common set of applications with which all or most employees interact may lead users to become familiar with other services located on the intranet as well. Self-service benefits also allow the human resource department to train less computer literate employees on an application that human resources knows intimately and that all employees must use (Meuse, 1999). This makes it an excellent introduction to the intranet. Unfortunately, most users only need to visit the employee benefits site once or twice a year, so the human resource function cannot be counted on to drive traffic throughout the year. However, newsletters can be published on the intranet, with notifications sent out via e-mail that may bring some traffic regularly.

Purchasing

For many organizations, especially larger ones, purchasing is a major cost that can be difficult to control. Managing the purchasing function and its information requirements is an inconvenience, and much of the cost is incurred in the overhead of processing purchases. The story of Ford's and Mazda's accounts payable departments has been retold many times and is famously cited by Hammer as an example of reengineering. Ford, with 500 people working in accounts payable, decided to compare itself to Mazda as a benchmark for its efficiency. Ford found that Mazda had only five people working in a similar department (Hammer & Champy, 1993). Although Mazda's efforts predated the Web, information processing was a crucial part of their efforts. This underscores the complexity and overhead often involved in purchasing. Intranets can greatly reduce purchasing overhead by simplifying the process and centralizing the collection of order data. Approval processes can be automated or at least facilitated, reducing paperwork and related clerical work.

Many office and maintenance, repair, and operations suppliers offer customizable versions of their product catalogs to be put onto client company intranets. This customization, which automates the process of billing and shipping, allows employees to order products directly without intervention from corporate purchasing departments. The order data are captured once, at the point of origin, stored in the appropriate database, and reused as needed without reentry, reducing mistakes and clerical staffing needs.

Operations

The intranet can be at the center of all operations for the company, serving in essence as an interface to all enterprise systems. This can include logistics, inventory, project management, and operational systems. With the increasing focus on customer responsiveness, time to market, and value chain integration, such operational transparency is becoming increasingly valued.

Directories

One simple, useful application is an electronic directory of employees. The directory can easily be kept current, and is regularly used by most employees, particularly if paper directories are no longer published.

Menus

Something as simple as cafeteria menus and catering information can be a useful tool to drive traffic to the site.

Calendar Systems

Centralized calendar systems are an excellent intranet application for many reasons. First, it is an extremely valuable service to all users who schedule meetings or other appointments. It is also something most people are likely to use on a daily basis. Benefits include the reduction of time spent in arranging meetings or other events, the elimination of "double booking," and the increased efficiency of all involved staff.

The system may be used to schedule meetings and conference rooms and to coordinate individual schedules. Such technologies can also synchronize personal digital assistants (PDAs), allowing employees to keep their calendars with them and also have them stored centrally.

Organizations may also enable the PDAs to access the intranet remotely via wireless technologies including WiFi and Bluetooth. Convergence of standards, advances in security, and decreasing hardware cost make such wireless access relatively inexpensive and viable.

Intranets may also include time clocks. As organizations move closer to activity-based costing, they require additional data on employee productivity. In particular, as more employees become knowledge workers, time tracking becomes an increasingly valuable tool and intranet applications make the input and analysis of these data easier, with widely accessible, familiar, intuitive graphical interfaces.

Group Collaboration

An increasing amount of work in organizations is being done by multiple people working collaboratively in disparate geographic locations, often at different times. Many intranet- and Internet-based applications are available to support such work. These systems include those for group authoring, document management, change control, moderated and threaded message boards, shared whiteboards, and workflow management software. Benefits of these applications can be realized by employees working on-site and remotely, making the intranet a necessary component of any organization's attempts to expand remote work arrangements such as telecommuting, hoteling, and accommodating those employees who are frequently on the road.

Hewlett–Packard found that such traditional group collaboration tools were not sufficient because they lacked the "causal proximity" necessary for productive group work. They implemented passive cameras to indicate whether someone was at or near their desk, and included an intercom-enabled telephone system and instant messaging, to foster regular, brief conversations (Sieloff, 1999). Such richer media helped emulate the benefits of working in the same physical location.

Syndicated Data

Many organizations make regular use of purchased external data feeds for strategic and operational work. Stock prices, weather, and news streams may all be purchased and made available through the intranet, customized for users based on their needs. Coordinating these streams requires careful management to balance easy access to current data with security.

Knowledge Management

Knowledge management has received a great deal of attention in the business and popular press. As organizations move away from manufacturing toward services, more knowledge is in the minds of workers than embedded in physical systems, or even documentation. This increases the cost of employee turnover and can impede growth.

Part of the challenge of knowledge management is uncovering, storing, and retrieving tacit knowledge. While many organizations struggle with simply managing explicit knowledge that can be readily articulated and documented, the greatest benefits often come from capturing tacit knowledge. The regular use of an intranet may encourage employees to routinely store information and knowledge and make explicit previously tacit knowledge.

Given the volume of information created in organizations, not only is indexing and retrieval important, but forgetting is also crucial. Archiving functions can move information out of the main databases to improve the relevancy of searches.

Implementing an intranet can fundamentally change the nature of the organization by reorganizing its business processes. By doing so, intranets can be a powerful lever of control (Simons, 1995), serving as a catalyst for change. General Motors credits its intranet, which links GM's 14 engineering centers with computer-aided design software and 3-D simulators, with an increase in creativity and even now recruits from the motion picture industry rather than just fine arts and automotive design programs (Rifkin, 2002) to take advantage of the Web-based skills. Such intranet-driven changes are so common that Pitt et al. have developed I-CAT, an instrument to measure the effectiveness of an intranet as a catalyst for change (Pitt, Murgolog-Poore, & Dix, 2001).

Technology

The technology for intranets is straightforward and similar to any other Web-based system. If the corporate network is TCP/IP, as most are, then starting an intranet site is as simple as placing a machine on the network, or finding one already connected, and installing or enabling a Web server. Current versions of Microsoft Windows desktop and server software come with Web servers installed and can be easily configured by anyone comfortable with personal computers. Once the server is up and running, the server may operate in background on an employee's ordinary PC workstation. Other employees may access the server by entering the IP address of the machine. Obviously, most users would prefer to type in a text domain name rather than a series of numbers. Such a text name would require either registration on the company's domain name server, which maps IP addresses to text names, or the use of Windows Internet Name Service (WINS), which serves the same purpose as DNS, but on Windows networks.

Content may be added to the Intranet in the form of HTML pages, which may be coded by hand, saved as HTML from word processors such as Microsoft Word, or written in HTML editors such as FrontPage and DreamWeaver. Simple programs may also be written to generate content in response to requests on the intranet through scripting in such languages as Cold Fusion, UltraDev, JSP, ASP, and Perl. HTML editors can generate simple scripts automatically for novice users. Such dynamic Web pages allow greater flexibility in the design of intranet Web pages.

Establishing the corporate portal/intranet as the access point for all information services is an appealing architecture for the IS department. Assuming the workstations are Windows-based, then all that needs to be installed on each machine is Microsoft Office with the Internet Explorer Web browser. This simplicity greatly reduces the support necessary for personal computers on the network because every employee in the company receives an identically configured computer with only standard Microsoft applications. This simplicity is one overarching benefit of an intranet over legacy network systems, since the need for individual computing support is significantly reduced by the common hardware and software, and user access is enabled through graphical, largely intuitive, interfaces.

All major software vendors have addressed the intranet market. Microsoft is working aggressively to integrate its desktop productivity software in Office with Web-based intranet tools, including video, Exchange server, and Microsoft Project (Microsoft, 2002). Microsoft even ships wizards for the creation of intranets, such as the 60 Minute Intranet Kit. Server software vendors such as Sun, IBM, and open-source Apache all offer intranet configuration suggestions and options for network and systems administrators familiar with those companies' products.

Enterprise system vendors have also focused heavily on Web-based access in the past several years. In that time, SAP, PeopleSoft, Baan, Oracle, and JD Edwards have all launched enterprise portal products to integrate their systems with a Web interface and other corporate intranet applications. Customer relationship management products such as Applix and Siebel have also launched portal interfaces in response to demands from the corporate market. The intranet/Web browser has become a common and accepted interface for corporate applications.

Extensible markup language (XML) is rapidly gaining support as a standard for Web documents. This is good news for intranets for several reasons. First, XML, with its focus on metadata, supports the complex, structured documents that populate most intranets. Second, as XML becomes a standard format for word processors, saving files to the intranet will become even easier. This approach has gained attention since Microsoft announced it intends to base its Office applications on XML. Finally, XML documents can be readily stored in databases and searched and retrieved quickly and easily, improving performance for large intranets.

Dot-coms were a major market for many of the software vendors who also compete in the intranet software market. With the collapse of the dot-coms, prices for Internet, and therefore intranet, software have dropped considerably. In addition, open-source software is becoming more robust, offering much less expensive alternatives to the major software vendors. Many companies use the open-source Apache suite to develop and manage their corporate intranets, and most major software vendors have started to support interface with Apache and other common open-source products.

As intranets grow, and increasing amounts of data are available, there must be a way to find information effectively. Information anxiety (Wurtman, 1989) is an inevitable challenge for intranets, given the volume of data likely to proliferate. Hewlett–Packard, for instance, found that within two years of launching its intranet, there were over two million documents stored on thousands of servers throughout the organization, and fewer than 5% of them were traditional official communications (Sieloff, 1999). Such volumes make information architecture (Rosenfeld & Morville, 2002) essential. Information architecture is a crucial component of successful intranets, and the hierarchical structure and navigation themes must be well thought out and implemented. An efficient and effective search engine available for the entire intranet is also required. Many organizations license third-party search engines to reduce the problems inherent in searching the wide variety and amount of information stored in an intranet.

Unless entries are key worded with metadata, users' searches are reduced to full-text searches, which are likely to return massive numbers of hits even with the best search engines. Proper metadata coding requires an appropriate taxonomy. Such a taxonomy may be available within the industry, or extant in the organization. If not, a thorough metadata taxonomy must be developed and implemented. Without a pre-hoc taxonomy, it is unlikely the metadata will be very useful. One survey found that only 31% of companies with enterprise portals had implemented a taxonomy (Answerthink, 2002), emphasizing the need for corporate information metadata management.

Establishing a taxonomy is only the first step; it must be used correctly and consistently to be effective. This requires all of those entrusted with entering data to use the taxonomy properly. This can be accomplished by simplifying the taxonomy so that there is no ambiguity in the interpretation, and there must be documentation and/or training to ensure consistency of the taxonomy's application at all data entry points. It is also likely that the taxonomy will need to evolve over time, requiring a plan for maintenance and continued training.

It may be possible to post much of the content on the Intranet with limited metadata coding because most users would find the content via browsing, or the search engine would likely find it and rate it as highly relevant.

Personalization

For users, having the most applicable material readily accessible can be extremely powerful and empowering. Customization strategies for intranet access can be individualized or group based. For example, a different interface can be developed for selective categories of employees, offering some customization at low cost, or personalization software can be installed to generate pages based on an individual's expressed or inferred preferences or behaviors. Most Web servers, including Microsoft IIS, iPlanet, and Apache, now come with basic personalization functionality.

However, Web developers have found that personalization is not a silver bullet. Indeed, Yahoo!, one of the largest personalizable sites on the Web, learned that most users, if given the choice, do not customize their interfaces (Manber, 2000). This makes designing the default pages and customizations more important than extensive customization options. Yahoo! also discovered that any customizations applied in one section of the site should apply to all areas, or users will become frustrated because they expect to need to set up customization only once. Yahoo! did find that a cadre of power users customized their interfaces beyond all expectations of the developers. Lessons learned from these power users can be applied to other users through redesign of the default or group-customized pages, underscoring the power of an information architecture steering group with input from all user groups.

Content Management

The Yankee Group outlines a three-stage content delivery cycle: creation, management, and presentation (Perry, 2001). Content management systems such as Vignette and Interwoven's TeamSite manage all three of these stages by storing all Web site data in a database. They facilitate content management by accepting external and internal automated data streams through syndicated, operational, and financial data, and receive input from publishers throughout the organization in word processing, spreadsheet, presentation, XML or HTML formats. These systems manage the data with change control, security, and workflow tools. Content management systems handle presentation by working with application servers such as BEA's WebLogic and IBM's WebSphere, to create pages based on templates created by designers.

For large intranets, content management can be a worthwhile solution. Much of the data on the intranet changes regularly, and updating a database is more efficient than making changes to all Web pages that contain it. Content management systems facilitate the distribution of intranet content over multiple channels, including multiple Web sites, possibly including externally accessible sites, high- and low-bandwidth versions, wireless/mobile, syndication, and even hard copy. Document management systems such as Documentum and Hummingbird and even document imaging software serve a similar purpose. In general, all of these products are evolving toward enterprise content management.

Ease of Use

The importance of usability is well established for all Web sites. In addition, usability has been linked to effective knowledge management (Begbie & Chudry, 2002). For intranets to be effective, they need a wide array of users across the organization. Providing extensive training and technical support to all users is not likely to be possible for reasons of cost and logistics. As employee turnover rises, reducing the learning curve for using the intranet becomes more important. Therefore the intranet must be well designed and simple to use.

IS DEPARTMENT INTRANET CHALLENGES

While intranets may be developed easily within departments with little support from IS, if the intranet is to scale and to function well, there are several areas the IS department must address. Indeed, intranets are likely to fundamentally change the nature of systems analysis and design, as users become more intimately involved in the process (Perry, 1998).

Standards vs Flexibility

The question of how much to constrain users in the interest of promoting standards for efficiency has plagued information systems professionals for decades. With intranets, the implications are substantially greater, as many people in the organization become, essentially, systems developers.

Controlling the use of spreadsheets and other personal productivity software was hard enough, but at least those systems were only accessed by one user, or perhaps disseminated over a sneaker net. With intranet data, potentially the entire organization has access to what is developed, and the developers have no formal training in systems analysis or design. Substantial duplication may arise, adding to costs and confusion. Poorly designed applications or content may lead to user rejection of the intranet, despite realizing other benefits.

The power of such distributed systems lies in the empowerment of employees with access to the data, and the ability to enter and manipulate it online. This empowerment is consistent with the principles of reengineering, "just in time," customer relationship management, supply chain management, and other management trends of the past several decades: enter data only once, at the source.

Intranet standards can be set for data storage, color schemes, templates, navigation, file naming, graphics formats, file formats, file sizes, and narrative voice. Content management systems can automate much of this standardization, and also establish approval processes for implementation.

Intranets have the benefit of limited or no access by those outside the organization, and this insulation offers many a false sense of security about the lack of need to protect the brand image. It is likely that much of the data on the intranet will make its way to suppliers and customers, whether indirectly or directly, because a successful intranet will be the primary source of information throughout the organization.

Availability

Once an intranet site is established, IS will be held responsible for ensuring that the site is available whenever needed. This may mean 24/7 in some instances. For a traditional system, with a limited number of professional developers operating in a staged development environment with source control on systems running in a data center, this is a well-solved problem, depending on resource constraints. However, for systems often hosted within departments outside IS control, and developed by untrained personnel, keeping the systems running can be a challenge. One poorly constructed script can bring a system to its knees and make diagnosis difficult. These systems are often not backed up, and nearly all lack any form of source control. However, such carelessness is not inevitable, and traditional IS methodologies may be applied to ensure the reliability of intranets.

Training and Support

As applications and services are moved onto the intranet, and traditional physical channels are closed down, there will be a significant need for training and support. Often the systems are brought live by departments without thought of the need for training personnel, but the support demands are made to the IS department. A training program can obviate many of the support calls, but requires a coordinated effort and a willingness on the part of departments to support in principle, if not financially, such training programs. Given the expense of lost productivity, such programs are often politically unpopular, but necessary for a successful implementation of any internal system.

Support for the systems is also a challenge because these applications may change frequently and the IS department may not have been involved in their construction. It is unrealistic, however, to believe that the call center, or its equivalent in smaller companies, will not be the main source of intranet support, whether de facto or de jure. Successful, sustainable intranets involve the IS department, especially for support.

Access

For the intranet to achieve its potential, it must displace other channels such as paper, phone, in person, and even e-mail. As these other channels are shut down, many employees who did not previously need access to a computer now need it to manage their benefits, enter a request for a repair with physical plant, look up a phone number, etc. This means providing access to all employees, many of whom have never needed computers before. Custodial staff, warehouse employees, and others may not need individual workstations, but do need convenient access to a computer connected to the intranet. This may mean extending the network to areas of buildings previously left unwired. While modern construction and wireless technologies make this less of an issue than previously, there may still be significant expense and effort to ensure intranet access to all. Access must also be established to legacy software systems that may not be Web-enabled.

Networks

Intranets may be more secure and reliable than the external Internet because companies have control over the network and servers on which they run. Intranet access may substantially increase the need for bandwidth on the network. Bandwidth demands can easily grow exponentially when an intranet is introduced, especially as the cost of multimedia hardware continues to drop. Although the cost of bandwidth is also dropping, the complexity of network management increases substantially as more bandwidth is used. Also, as demand for the services increases, the need for network reliability also rises. Good networks are more than bandwidth. As more work is intranet-based and dependent on the corporate network, availability, reliability, service, support, and scalability of the network all become crucial.

Security

> If a Sun employee attaches a modem inside our network without going through the firewall, he's fired. —Sun Microsystems chief network officer Geoffrey Baebr (*Information Week*, 1997)

Intranets may contain a great deal of sensitive personal and organizational data. These clearly need to be protected. The fact that intranets are for internal use only offers some a false sense of security. This laxity can lead to greater risks if there is any exposure to external access.

Intranet security is similar to security for any Web-based system, but potentially more complex because of the larger number of people with authorized access. When the Navy began implementing its Navy/Marine Corps Intranet (NMCI), it quickly found that the U.S. Department of Defense Information Technology Security Certification and Accreditation Process (DITSCAP) was difficult to apply. The Department of Defense needed to adapt the process to fit the needs of intranet development (Gerstmar, 2002).

Given that most security lapses are from authorized users within the organization, intranets create numerous potential breaches. Access controls for data and process may be set at the operating system, directory, application server, and Web server levels. Authorization can be done through passwords, or even biometrics, depending on the value of the information. Firewalls can be established not only at the border with the Internet to keep those outside the organization out, but also inside the organization to limit access to authorized employees.

Implementation

> Get the chairman in early, but get him out quick.
> —Alan Boehme, DHL Director of Business Planning (*InformationWeek*, 1997b)

It is well established that top management support is a crucial factor in the success of all information systems implementations. However, it is not always sufficient. Many intranets are small, bootstrapped applications implemented with little fanfare. For an intranet to be successful, the goals for the initiative must be articulated. This involves determining the audiences for the system, the services to be offered, and the extent to which they are to be used. Once the goals are known, then metrics can be established for measuring success. With metrics in place, a plan can be created for implementing the system. This is another reason to have a corporate information systems architecture that covers the development, implementation and maintenance of all intranets.

Run a Pilot Study

Because intranet systems are Web-based, prototypes are easy and generally inexpensive to build. This makes pilot testing an effective tool for determining the relevance of this technology in relation to a company's needs. However, it is important to keep in mind the goal of the pilot study. Many people incorrectly view pilot studies as always determining definitively whether a given system will work within the target community. Rather, most initial pilot studies should be tests for proof of concept, and should be repeated several times in different areas of the organization to ensure the generalizability of the concept test.

If the pilot is testing for proof of concept, then the population chosen for the test should be the one likely to adopt it. The initial pilot will be closely watched by many in the organization, and a failure could set the project back significantly. Participation in the pilot study is also an opportunity for potential users of the production system to invest psychologically in it. Potential users may be coopted into adopting the technology by being brought into the development process as well as testing. IS research has shown that involvement in the design process makes people more likely to support and use the intranet system when implemented, and with intranet technologies, these people can be involved not only in the abstract design but in the actual construction, increasing their actual and perceived investment in the system, and therefore the likelihood of their support.

Successful intranets require momentum to achieve critical mass. Gladwell (2000) refers to a tipping point, where phenomena, including technology adoptions, either succeed or fail. Gladwell notes three factors that determine the tipping point of a technology: the law of the few, stickiness, and context.

The law of the few reflects the fact that not all users or decision makers are equally influential. Gladwell identifies three types of people that are particularly influential: connectors, who know and interact with a large number of people; mavens, who are considered experts in a given area; and salespeople, who will advocate for a position in which they believe. When implementing an intranet, identifying and engaging people from each of these groups is crucial to the project's success.

Stickiness reflects the extent to which something is memorable. The term stickiness is often used in reference to Web pages to describe how long visitors stay. Time spent by visitors is an easily gathered measure of user interest in a page or site.

Finally, context refers to the way in which information about the innovation or technology is disseminated. For instance, something advocated by members of the business units is likely to be more successful than something promoted by an e-mail from the IS department.

When the concept has been firmly established as viable, and there is support within the organization, then a representative group that will be more likely to predict success in the organization as a whole may be chosen.

Once the system is implemented, it is essential to eliminate other sources of the same information or service. This can be done drastically or gradually. For instance, calls to the help desk or human resources could ask whether the employee has gone through the Web site. If she/he has not accessed the intranet before calling for support, IS or HR staff could offer to guide the caller through the site for resolution before providing an answer directly.

Maintenance and Management

A major component of a successful intranet is a plan for its control. The first obvious question is who should control it. Four common configurations have emerged: IS, corporate, human resources, and distributed. Any one of the models may work, depending on the goals of the intranet and the structure of the organization.

With the IS department in control of the intranet, the technology is likely to be well aligned with the existing systems and architecture, making it more reliable. However, in terms of prioritization, the IS department is likely to be the worst choice for deciding what needs to be updated first. At a minimum, putting IS in charge adds an additional layer of control beyond the creators of the content. Technologies such as content management can make it more viable for business units to control the system.

In a corporate structure, there is a decision maker within the organization who oversees the business units involved in the intranet. This person makes strategic decisions about the intranet. Tactical decisions may be delegated farther down the organization. The systems themselves may be built by the business units or the IS department.

Although human resources applications are among the most common and heavily used functions of intranets, as more operational information is added, most human resource departments are ill equipped to manage the systems.

The distributed model is appealing because the subject matter experts can be put in control of the data. With content management systems supported by the IS department, this model can be viable as the business users are relieved of responsibility for design, but retain control of content.

A plan for maintenance of the system is crucial. At a minimum, such a plan should address how often the information should be updated, who is responsible for removing out-of-date information, how information will be archived, and how often the design should be revisited. Redesigns should have a formal structure and budget process (Fichter, 2001). There should also be a disaster recovery plan commensurate with the importance of the intranet. Depending on the organizational value of the intranet, the plan for business continuity in the event of a disaster may range from periodic backups with a restoration plan, to off-site mirrored servers.

An acceptable use policy is an integral part of any intranet plan. Such a plan makes clear what is appropriate for the intranet, and may also include guidelines for Internet access as well. Although filters can be helpful in reducing inappropriate content, they are not a substitute for a well-articulated policy. For guidelines on creating an acceptable use policy, see Lichtenstein and Swatman (1997).

Cost/Benefit Calculations

Calculating the value of information technology has been the subject of extensive research over the past decade. The term productivity paradox is often used to describe the argument that despite rapid increases in spending on information technology, there is not a concurrent increase in productivity.

Measuring the return on intranet investments poses the same complexities as other information systems: calculating all of the costs involved is difficult and attributing benefits to them is even more challenging. However, such calculations are feasible, and there is reason to believe that a well-planned intranet can be profitable (Healey, 2001).

Among the obvious and easily calculated costs of an intranet are the hardware and software required for the servers, including any support contracts, and development software purchased. The costs of development may be relatively easy to measure for IS personnel involved in the system construction if they are contracted, or if the internal IS shop tracks costs on a project basis. However, much of the construction will involve the time of those outside IS, and there may be no extant tools for tracking their time, nor easy access to data on the cost of their time. Network costs for running the intranet may not be easy to track, because both utilization and marginal cost would need to be calculated. A methodological change in how such costs are calculated could make the difference in positive or negative ROI. Training costs must be included, both for the trainers' time in development and delivery and also lost productivity for trainees.

Application service providers can provide all of the infrastructure and functionality of an intranet for an organization with only an Internet connection. Infostreet.com (http://www.infostreet.com), Intranets.com (http://www.intranets.com), and others offer a broad array of cost-effective intranet services to companies of all sizes.

The decision to outsource intranet development is much like the "make or buy" decision for types of information technology. Organizations must decide what, if any, aspects of the intranet should be produced or maintained outside the organization.

While cost calculation is difficult, benefit calculation is even more difficult. Benefit calculation should include obvious direct cost reductions such as headcount, paper, and software licenses. One commonly used metric for efficiency is to credit the system for the pro-rated salary saved by using the intranet rather than the previous methods. However, this makes the assumption that all of the difference is saved. In many cases, the previous methods were not a complete loss; for instance, workers were doing other work while on the phone, or filled out forms during their commute. Not all of the recovered time will be used productively. Potentially more important than such bottom line cost reductions are top line enhancements in revenue. However, such improvements are harder to measure. Specifically, has the intranet contributed to better service, higher quality products, or improved time to market? If so, how much is that worth to the organization?

One excellent source of metrics for evaluating an intranet is the server access logs. All page accesses are automatically logged, along with the time and date of access, IP address of the user, which can be mapped to the specific intranet user, and the referring page, if any. These data may be analyzed to determine how employees are using the intranet and to identify potential enhancements. Estimates of the most and least popular pages, typical paths through the site, most used search terms, performance data, and other metrics are easily generated with simple analysis tools. This can be supplemented with direct surveys of users, online and through other channels.

WHY DO INTRANETS FAIL?

Many surveys show intranet initiative failure rates over 50%. Of course, it is important to keep this number in perspective, as surveys show similar percentages of all IS projects fail. These rates of failure may reflect unrealistic goals in some organizations, but often there are fundamental causes for failure. This section outlines common reasons for failures of intranets.

Other Channels Are Not Shut Down

Electronic systems only save paper if the paper versions are eliminated. If an employee handbook is still printed, then the intranet version is simply an added cost, and confusion arises about which version is correct.

Intranet Is Not Used

Many intranet systems languish after expensive and extensive construction. Often organizations do not bother to track intranet access, failing to identify the lack of interest earlier.

No Maintenance

Many intranets fall victim to good intentions. Intranets are often implemented, but are not made anyone's primary responsibility. There is often an initial burst of energy, but then no information is updated.

Poor Planning

Many intranets are implemented without adequate consideration of the issues described above.

Inadequate Content

In some cases, there is simply not enough reason to visit the intranet to sustain it.

Poor Performance

Intranets are likely to see lumpy access patterns, and the machines should be specified to handle these peak demands. Intranets are often implemented on disused hardware because intranets are rarely a budget priority. Poorly optimized databases can quickly bog down under volume.

Poor Interface Design

Many intranets are designed by people inside the organization with little or no design training, leading to obtuse and cumbersome systems and user frustration.

CONCLUSION

Intranets can be useful for organizations of any size to improve communication and information dissemination, and to lower costs. Powerful tools can be installed and configured quickly with relatively low investment. The technologies have standardized and are reliable and reasonably priced. Qualified, experienced personnel are readily available to build and maintain intranets. Success in intranet development and management simply requires the application of traditional business and information systems skills for establishing goals, planning, employing the proper architecture and technologies in response to business needs, and measuring outcomes.

GLOSSARY

Acceptable use policies How users may employ a system.

Content management systems Systems that use databases to store text, graphics, and other material to be displayed on a Web site, and then dynamically build pages based on previously designed templates.

Extensible markup language (XML) A language for describing the content and format of documents containing structured information.

Extranets Corporate systems containing sensitive data made accessible over the Internet to selected customers, suppliers, and business partners.

Information architecture The modeling of the structure of a Web site, including functionality and data, based on articulated goals.

Intranets Systems that use Web technologies such as HTTP, TCP/IP, FTP, and SMTP within organizations to facilitate communication and the distribution of information.

Knowledge management The process of making explicit and codifying the knowledge in an organization.

Hoteling An office configuration that allocates space temporarily on a first-come, first-served basis, similar to a hotel stay rather than assigning permanent space to workers who may not spend all of their time in the office.

Metadata The structure and content of data to facilitate future use of the data, particularly for searching.

Portals Web pages or sites that aggregate access to information and to other systems, serving as a starting point for users.

Reengineering The process of redesigning business processes to achieve radical improvements.

Scripting languages Programming tools for handling simple processing needs, two popular Web ones being Perl and JavaScript.

Self-service benefits The online management by employees of their human resource benefits administration without intervention by support staff.

Syndicated data Feeds that may be purchased to provide consistent or streaming updates such as weather, news, and stock prices.

CROSS REFERENCES

See *Extranets; GroupWare; Internet Literacy.*

REFERENCES

Answerthink (2002). Despite widespread portal technology adoption, business benefits remain elusive. *PR Newswire*. Retrieved May 17, 2003, from http://www.answerthink.com/news_and_events/press_release_2002_detail.asp?ident=252

Baker, S. (2000). Getting the most from your intranet and extranet strategies. *Journal of Business Strategy, 21*(4), 40–43.

Begbie, R., & Chudry, F. (2002). The intranet chaos matrix: A conceptual framework for designing an effective knowledge management intranet. *Journal of Database Marketing, 9*(4), 325–338.

Fichter, D. (2001). The intranet of your dreams and nightmares: Redesign issues. *Online, 25*(5), 74–76.

Gerstmar, T. (2002). Legacy systems, applications, challenge intranet rollout. *Signal, 57*(4), 29–31.

Gladwell, M. (2000). *The tipping point: How little things can make a big difference*. Boston: Little, Brown.

Hammer, M., & Champy, J. (1993). *Reengineering the corporation*. New York: Harper Business.

Hansen, M. T., & Deimler, M. S. (2001). Cutting costs while improving morale with B2E management. *Sloan Management Review, 43*(1), 96–100.

Healey, A. (2001). Using ROI to champion workplace portals. *Workspan, 44*(3), 22–30.

InformationWeek (1997a, June 2). *InformationWeek*, 10.

Information Week (1997b, October 6). *InformationWeek*, 13.

Lichtenstein, S., & Swatman, P. M. C. (1997). Internet acceptable usage policy for organizations. *Information Management and Computer Security, 5*(5), 182–190.

Manber, U. (2000, August). Experience with personalization on Yahoo! *Communications of the ACM*, 35–39.

Meuse, D. (1999). Making employee self-service work for employees and your company. *Benefits Quarterly, 15*(3), 18–23.

Microsoft. (2002). *Microsoft solution for intranets*. Retrieved December 20, 2002, from http://www.microsoft.com/solutions/msi/

Perry, R. (2001). *Managing the content explosion into content-rich applications (Internet computing strategies)*. Boston: The Yankee Group.

Perry, W. G. (1998). What is an intranet, and how is it going to change systems analysis and design? *Journal of Computer Information Systems, 39*(1), 55–59.

Pitt, L., Murgolog-Poore, M., & Dix, S. (2001). Changing change management: The intranet as catalyst. *Journal of Change Management, 2*(2), 106–114.

Rifkin, G. (2002). GM's Internet overhaul. *Technology Review, 105*(8), 62–67.

Rosenfeld, L., & Morville, P. (2002). *Information architecture for the World Wide Web* (2nd ed.). Cambridge, MA: O'Reilly.

Sieloff, C. G. (1999). If only HP knew what HP knows: The roots of knowledge management at Hewlett-Packard. *Journal of Knowledge Management, 3*(1), 47–53.

Simons, R. (1995). *Levers of control: How managers use innovative control systems to drive strategic renewal*. Boston: Harvard Business School Press.

Webster's Revised Unabridged Dictionary (1913). Springfield, MA: C. & G. Merriam Co.

Wurtman, R. (1989). *Information anxiety*. New York: Doubleday.

Intrusion Detection Techniques

Peng Ning, *North Carolina State University*
Sushil Jajodia, *George Mason University*

INTRODUCTION

Intuitively, *intrusions* in an information system are the activities that violate the security policy of the system, and *intrusion detection* is the process used to identify intrusions. Intrusion detection has been studied for approximately 20 years. It is based on the beliefs that an intruder's behavior will be noticeably different from that of a legitimate user and that many unauthorized actions will be detectable.

Intrusion detection systems (IDSs) are usually deployed along with other preventive security mechanisms, such as access control and authentication, as a second line of defense that protects information systems. There are several reasons that make intrusion detection a necessary part of the entire defense system. First, many traditional systems and applications were developed without security in mind. In other cases, systems and applications were developed to work in a different environment and may become vulnerable when deployed in the current environment. (For example, a system may be perfectly secure when it is isolated but become vulnerable when it is connected to the Internet.) Intrusion detection provides a way to identify and thus allow responses to attacks against these systems. Second, due to the limitations of information security and software engineering practice, computer systems and applications may have design flaws or bugs that could be used by an intruder to attack the systems or applications. As a result, certain preventive mechanisms (e.g., firewalls) may not be as effective as expected.

Intrusion detection complements these protective mechanisms to improve the system security. Moreover, even if the preventive security mechanisms can protect information systems successfully, it is still desirable to know what intrusions have happened or are happening, so that we can understand the security threats and risks and thus be better prepared for future attacks.

In spite of their importance, IDSs are not replacements for preventive security mechanisms, such as access control and authentication. Indeed, IDSs themselves cannot provide sufficient protection for information systems. As an extreme example, if an attacker erases all the data in an information system, detecting the attacks cannot reduce the damage at all. Thus, IDSs should be deployed along with other preventive security mechanisms as a part of a comprehensive defense system.

Intrusion detection techniques are traditionally categorized into two methodologies: *anomaly detection* and *misuse detection*. Anomaly detection is based on the normal behavior of a subject (e.g., a user or a system); any action that significantly deviates from the normal behavior is considered intrusive. Misuse detection catches intrusions in terms of the characteristics of known attacks or system vulnerabilities; any action that conforms to the pattern of a known attack or vulnerability is considered intrusive.

Alternatively, IDSs may be classified into host-based IDSs, distributed IDSs, and network-based IDSs according to the sources of the audit information used by each IDS. Host-based IDSs get audit data from host audit trails and usually aim at detecting attacks against a single host; distributed IDSs gather audit data from multiple hosts and possibly the network that connects the hosts, aiming at detecting attacks involving multiple hosts. Network-based IDSs use network traffic as the audit data source, relieving the burden on the hosts that usually provide normal computing services.

This chapter starts with an overview of current intrusion detection techniques. Next, it reviews the various types of anomaly detection methods, such as statistical models and machine learning methods, followed by an overview of various types of misuse detection methods, including rule-based languages, the colored Petri-net-based method, and the abstraction-based method. The section following that discusses additional techniques for intrusion detection in distributed systems, including distributed IDSs, network-based IDSs, and interoperation between (heterogeneous) IDSs.

ANOMALY DETECTION

Statistical Models

Statistical modeling is among the earliest methods used for detecting intrusions in electronic information systems. It is assumed that an intruder's behavior is noticeably different from that of a normal user, and statistical models are used to aggregate the user's behavior and distinguish an attacker from a normal user. The techniques are applicable to other subjects, such as user groups and programs. Here, we discuss two statistical models that have been proposed for anomaly detection: NIDES/STAT and Haystack.

NIDES/STAT

The Stanford Research Institute's next-generation real-time intrusion detection expert system statistical component (NIDES/STAT) observes behaviors of subjects on a monitored computer system and adaptively learns what is normal for individual subjects, such as users and groups (Axelsson, 1999). The observed behavior of a subject is flagged as a potential intrusion if it deviates significantly from the subject's expected behavior.

The expected behavior of a subject is stored in the profile of the subject. Different measures are used to measure different aspects of a subject's behavior. When audit records are processed, the system periodically generates an overall statistic, T^2, that reflects the abnormality of the subject. This value is a function of the abnormality values of all the measures comprising the profile. Suppose that n measures $M_1, M_2, \ldots, M_n$ are used to model a subject's behavior. If $S_1, S_2, \ldots, S_n$ represent the abnormality values of M_1 through M_n, then the overall statistic T^2 is evaluated as follows, assuming that the n measures are independent of each other:

$$T^2 = {S_1}^2 + {S_2}^2 + \cdots + {S_n}^2.$$

The profile of a subject is updated to reflect the changes of the subject's behavior. To have the most recently observed behaviors influence the profile more strongly, NIDES/STAT multiplies the frequency table in each profile by an exponential decay factor before incorporating the new audit data. Thus, NIDES/STAT adaptively learns a subject's behavior patterns. This keeps human users from having to manually adjust the profiles; however, it also introduces the possibility of an attacker gradually "training" the profile to mask his/her intrusive activities as normal behavior.

Haystack

Haystack used a different statistical anomaly detection algorithm, which was adopted as the core of the host monitor in the distributed intrusion detection system (DIDS) (Axelsson, 1999). This algorithm analyzes a user's activities according to a four-step process.

First, the algorithm generates a session vector to represent the activities of the user for a particular session. The session vector $\mathbf{X} = <x_1, x_2, \ldots, x_n>$ represents the counts for various attributes used to represent a user's activities for a single session. Examples of the attributes include session duration and number of files opened for reading.

Second, the algorithm generates a Bernoulli vector to represent the attributes that are out of range for a particular session. A threshold vector $\mathbf{T} = <t_1, t_2, \ldots, t_n>$, where t_i is a tuple of the form $<t_{i,min}, t_{i,max}>$, is used to assist this step. The threshold vector is stored in a user's profile. The Bernoulli vector $\mathbf{B} = <b_1, b_2, \ldots, b_n>$ is generated so that b_i is set to 1 if x_i falls outside the range t_i, and b_i is set to 0 otherwise.

Third, the algorithm generates a weighted intrusion score, for a particular intrusion type, from the Bernoulli vector and a weighted intrusion vector. Each group and intrusion type pair has a weighted intrusion vector $\mathbf{W} = <w_1, w_2, \ldots, w_n>$ in which each w_i relates the importance of the ith attribute in the Bernoulli vector to detecting the particular intrusion type. The weight intrusion score is simply the sum of all weights, w_i, where the ith attribute falls outside the range t_i. That is,

$$\text{the weighted intrusion score} = \sum_{i=1}^{n} b_i \cdot w_i.$$

Finally, the algorithm generates a suspicion quotient to represent how suspicious this session is compared with all other sessions for a particular intrusion type. Specifically, the suspicion quotient is the probability that a random session's weighted intrusion score is less than or equal to the weighted intrusion score computed in the previous step.

Unlike NIDES/STAT, the Haystack algorithm has a step that determines resemblance to known attacks. The advantages are that more knowledge about the possible attacks can be derived from this step and better responses can follow the alarms. However, extra knowledge about possible intrusion types is required: We need to understand the impact of the intrusion types on the attributes of the session vectors and assign appropriate weights to these attributes to reflect the impact. In reality, the process of generating the weighted intrusion vectors is time consuming and error prone.

Machine Learning and Data Mining Techniques

Time-Based Inductive Machine

Teng, Chen, and Lu (1990) proposed the use of a time-based inductive machine (TIM) to capture a user's behavior pattern. As a general-purpose tool, TIM discovers temporal sequential patterns in a sequence of events. The sequential patterns represent highly repetitive activities and are expected to provide prediction. The temporal patterns, which are represented in the form of rules, are generated and modified from the input data using a logical inference called *inductive generalization*. When applied to intrusion detection, the rules describe the behavior patterns of either a user or a group of users based on past audit history. Each rule describes a sequential event pattern that predicts the next event from a given sequence of events. An example of a simplified rule produced in TIM is

$$E1 - E2 - E3 \rightarrow (E4 = 95\%; E5 = 5\%),$$

where $E1$, $E2$, $E3$, $E4$, and $E5$ are security events.

This rule says that if $E1$ is followed by $E2$, and $E2$ is followed by $E3$, then there is a 95% chance (based on the previous observation) that $E4$ will follow, and a 5% chance that $E5$ will follow. TIM can produce more generalized rules than the above. For example, it may produce a rule in the form

$$E1 - {}^{*} \rightarrow (E2 = 100\%),$$

where an asterisk matches any single event. Any number of asterisks is allowed in a rule.

The limitation of TIM is that it only considers the *immediately following* relationship between the observed events. That is, the rules only represent the event patterns in which events are adjacent to each other. However, a user may perform multiple tasks at the same time. For example, a user may check his/her e-mail during the editing of a document. The events involved in one application, which tend to have strong patterns embedded in the sequence of events, may be interleaved with events from other applications. As a result, it is very possible that the rules generated by TIM cannot precisely capture the user's behavior pattern. Nevertheless, TIM may be suitable for capturing the behavior patterns of such entities as programs that usually focus on single tasks.

Instance Based Learning

Lane and Brodley (1998) applied *instance based learning* (IBL) to learn entities' (e.g., users) normal behavior from temporal sequence data. IBL represents a concept of interest with a set of instances that exemplify the concept. The set of instances is called the instance dictionary. A new instance is classified according to its relation to stored instances. IBL requires a notion of "distance" between the instances so that the similarity of different instances can be measured and used to classify the instances.

Lane and Brodley did several things to adapt IBL to anomaly detection. First, they transformed the observed sequential data into fixed-length vectors (called *feature vectors*). Specifically, they segmented a sequence of events (e.g., a sequence of user commands) into all possible overlapping sequences of length l, where l is an empirical parameter. (Thus, each event is considered the starting point of a feature vector, and each event is replicated l times.) Second, they defined a similarity measure between the feature vectors. For a length l, the similarity between feature vectors $\mathbf{X} = (x_0, x_1, \ldots, x_{l-1})$ and $\mathbf{Y} = (y_0, y_1, \ldots, y_{l-1})$ is defined by the functions

$$w(\mathbf{X}, \mathbf{Y}, i) = \begin{cases} 0, & \text{if } i < 0 \text{ or } x_i \neq y_i \\ 1 + w(\mathbf{X}, \mathbf{Y}, i-1), & \text{if } x_i = y_i \end{cases}$$

and

$$Sim(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{l-1} w(\mathbf{X}, \mathbf{Y}, i).$$

The converse measure, distance, is defined as $\text{Dist}(\mathbf{X}, \mathbf{Y}) = \text{Sim}_{\max} - \text{Sim}(\mathbf{X}, \mathbf{Y})$, where $\text{Sim}_{\max} = \text{Sim}(\mathbf{X}, \mathbf{X})$. Intuitively, the function $w(\mathbf{X}, \mathbf{Y}, i)$ accumulates weights from the most recently consecutively matched subsequences between $\mathbf{X}$ and $\mathbf{Y}$ at position i, whereas $\text{Sim}(\mathbf{X}, \mathbf{Y})$ is the integral of total weights.

A user profile is built to contain a collection of sequences, $\mathbf{D}$, selected from a user's observed actions (e.g., commands). The similarity between the profile and a newly observed sequence, $\mathbf{X}$, is defined as $\text{Sim}_{\text{D}}(\mathbf{X}) = \max_{\text{Y} \in \text{D}} \{\text{Sim}(\mathbf{Y}, \mathbf{X})\}$. That is, the similarity between $\mathbf{X}$ and $\mathbf{D}$ is defined as the similarity between $\mathbf{X}$ and a vector in $\mathbf{D}$ that is most similar to $\mathbf{X}$. Then a threshold r is chosen. If the similarity between an observed sequence $\mathbf{X}$ and the profile $\mathbf{D}$ is greater than r, $\mathbf{X}$ is considered normal; otherwise, $\mathbf{X}$ is abnormal.

To reduce the storage required by the profile, Lane and Brodley used the least-recently-used pruning strategy to keep the profile at a manageable size. As new instances are acquired and classification is performed, the profile instance selected as most similar is time stamped. Least-recently-used instances are removed when the profile is constrained to the desired size. In addition, they applied a clustering technique to group the instances in the profile and used a representative instance for each cluster.

This attempt shares a problem similar to that of TIM, that is, it tries to find patterns from sequences of consecutive events. As they have noted, a user may interrupt his/her normal work (e.g., programming) and do something different (e.g., answer an urgent e-mail) and thus yield a different sequence of actions from his/her profile. Their solution (Lane & Brodley, 1998) is to use a time average of the similarity signals; however, such a solution may make real anomalies unnoticeable. In addition, the least-recently-used pruning strategy gives an attacker a chance to train the profile slowly, so that intrusive activities are considered normal ones.

Neural Network

Fox, Henning, Reed, and Simmonian (1990) were the first to attempt modeling system and user behaviors using neural networks. Their choice of neural network is Kohonen's *self-organizing map* (SOM), which is a type of unsupervised learning technique that can discover underlying structures of the data without prior examples of intrusive and nonintrusive activities.

They used SOM as a real-time background monitor that alerts a more complex expert system. In their prototype system, 11 system parameters accessible from the system's statistical performance data are identified as the input to the SOM model. These parameters include (a) central processing unit (CPU) usage, (b) paging activity, (c) mailer activity, (d) disk accesses, (e) memory usage, (f) average session time, (g) number of users, (h) absentee jobs, (i) reads of "help" files, (j) failed log-ins, and (k) multiple log-ins. However, their study showed the results of only one simulated virus attack, which are not sufficient to draw serious conclusions.

In another attempt to apply neural network to anomaly detection, Ghosh, Wanken, and Charron (1998) proposed using a *back-propagation* network to monitor running programs. A back-propagation network is developed for supervised learning. That is, it needs examples of intrusive and nonintrusive activities (called *training data*) to build the intrusion detection model. Such a network consists of an input layer, at least one hidden layer (neurons that are not directly connected to the input or output nodes), and an output layer. Typically, there are no connections between neurons in the same layer or between those in one layer and those in a previous layer.

The training cycle of a back-propagation network works in two phases. In the first phase, the input is submitted to the network and propagated to the output through the network. In the second phase, the desired output is compared with the network's output. If the vectors do not agree, the network updates the weights starting at the output neurons. Then the changes in weights are calculated for the previous layer and cascade through the layers of neurons toward the input neurons.

Ghosh et al. proposed using program input and the program internal state as the input to the back-propagation network. One interesting result is that they improved the performance of detection by using randomly generated data as anomalous input. By considering randomly generated data as anomalous, the network gets more training data that is complementary to the actual training data.

Similar to statistical anomaly detection models, deciding the input parameters for neural network anomaly detectors is a difficult problem. In addition, assigning the initial weights to the neural networks is also an unresolved question. The experiments by Ghosh et al. (1998) showed that different initial weights could lead to anomaly detectors with different performance. Nevertheless, research on applying neural networks to anomaly detection is still preliminary; more work is needed to explore the capability of neural networks.

Audit Data Analysis and Mining

Audit data analysis and mining (ADAM) proposes applying data mining techniques to discover abnormal patterns in large amounts of audit data (e.g., network traffic collected by TCPdump, which is a program used to sniff and store packets transmitted in the network) (Barbara, Couto, Jajodia, & Wu, 2001; Barbara, Couto, Jajodia, & Wu, 2002). In particular, the existing research focuses on the analysis of network audit data, such as the transmission control protocol (TCP) connections. Using data mining techniques, ADAM has the potential to provide a flexible representation of the network traffic pattern, uncover some unknown patterns of attacks that cannot be detected by other techniques, and accommodate the large amount of network audit data that keeps growing in size.

ADAM uses several data-mining-related techniques to help detect abnormal network activities. The first technique ADAM uses is inspired by association rules. Given a set I of items, an association rule is a rule of the form $X \rightarrow Y$, where X and Y are subsets (called item sets) of I and $\mathrm{X} \cap \mathrm{Y} = \phi$. Association rules are usually discovered from a set T of transactions, where each transaction is a subset of I. The rule $X \rightarrow Y$ has a *support s* in the transaction set T if $s\%$ of the transactions in T contain $X \cup Y$, and it has a *confidence c* if $c\%$ of the transactions in T that contain X also contain Y.

However, ADAM doesn't use association rules directly; instead, it adopts the item sets that have large enough support (called *large item sets*) to represent the pattern of network traffic. Specifically, it assumes that each network event (e.g., a TCP connection) is described by a set of attribute values and considers each event a transaction. The large item sets discovered from the network traffic then represent the frequent events in the network traffic. The power of such a mechanism lies in the flexible representation of events.

ADAM builds a profile of normal network activities in which the frequent events (represented by large item sets) are stored. During the detection time, it adopts a sliding-window method to incrementally examine the network events. Within each window, ADAM looks for the large item sets that do not appear in the profile and considers them suspicious.

The second technique ADAM uses is called *domain-level mining*. Intuitively, it tries to generalize the event attribute values used to describe a network event. For example, an IP address that belongs to the subnet *ise.gmu.edu* can be generalized to *ise.gmu.edu*, *gmu.edu*. and *edu*. Then it discovers large item sets using the generalized attribute values. An advantage of this approach is that it provides a way to aggregate the events that share some commonality and may discover more attacks. However, the scheme used to generalize the attribute values is ad hoc; only generalization from IP addresses to subnets and from smaller subnets to larger subnets are studied.

The third technique ADAM uses is classification. ADAM is innovative in that classification is used to classify the output of the mining of large item sets. Four classification algorithms have been studied to date: C4.5 decision tree, naive Bayes, cascading classifier (which uses decision tree followed by naive Bayes and vice versa), and inductive rule learner. The results show that classification is quite effective in reducing false alarms.

Finally, ADAM uses the pseudo-Bayes estimator to accommodate unknown attacks. It is assumed that unknown attacks are the attacks that have not been observed. The training data is represented as a set of vectors, each of which corresponds to an event and is labeled as normal or as a known class of attacks. An additional class is then considered to represent the unknown attacks. Because the unknown attacks haven't been observed in the training data, the probability $P(\mathbf{x} \mid \text{class} = \text{unknown})$, where $\mathbf{x}$ is a training vector, is zero. The pseudo-Bayes estimator is used to smooth all the conditional probabilities $P(\mathbf{x} \mid \text{class})$ so that $P(\mathbf{x} \mid \text{class} = \text{unknown})$ is assigned a (small) probability. These conditional probabilities are then used to build a naive Bayes classifier.

The limitation of ADAM is that it cannot detect stealthy attacks. In other words, it can detect an attack only when it involves a relatively large number of events during a short period of time. This limitation occurs because ADAM raises an alarm only when the support of an unexpected rule (i.e., association of event attributes) exceeds a threshold. Indeed, this limitation is not unique to ADAM; most of the anomaly detection models suffer from the same problem.

Computer Immunological Approach

The computer immunological approach is based on an analogy of the immune system's capability of distinguishing self from non-self (Hofmeyr, Forrest, & Somayaji, 1998). This approach represents self as a collection of strings of length l, where l is a system-wide parameter. A string of length l is considered non-self if it does not match any string belonging to self. To generate detectors that can distinguish non-self from self, a naive approach is to randomly generate a string of length l and check whether it matches any self-string. If yes, the generated string is discarded; otherwise, it is used as a detector. However, the naive approach takes time exponential to the number of self-strings. To address this problem, Forrest et al. proposed a "r-contiguous-bits" matching rule to distinguish self from non-self: two l-bit strings match each other if they are identical in at least r contiguous positions. As a result, detectors can be generated more efficiently for this particular matching rule.

Hofmeyr et al. proposed using short sequences of system calls to distinguish self from non-self for anomaly detection. Given a program in a particular installation, the immunological approach collects a database of all unique system call sequences of a certain length made by the program over a period of normal operation. During the detection time, it monitors the system call sequences and compares them with the sequences in the aforementioned database. For an observed sequence of system calls, this approach extracts all subsequences of length l and computes the distance $d_{\min}(i)$ between each subsequence i and the normal database as $d_{\min}(i) = \min\{d(i,j)\}$ for all sequences j in the normal database}, where $d(i,j)$ is the Hamming distance between sequences i and j (i.e., the number of different bits in sequences i and j). The anomaly score of the observed sequence of system calls is then the maximum $d_{\min}(i)$ normalized by dividing the length of the sequence. This approach raises an alarm if the anomaly score is above a certain threshold.

The advantage of this approach is that it has a high probability of detecting anomalies using a small set of self-strings based on short sequences of system calls. In addition, it does not require any prior knowledge about attacks. The disadvantage is that it requires self to be well understood. That is, it requires a complete set of self-strings in order not to mistake self for non-self. This requirement may be trivial for such applications as virus detection, but it is very difficult for intrusion detection, where some of normal behaviors cannot be foreseen when the detectors are being generated.

Sekar, Bendre, Dhurjati, and Bollineni (2001) further improve the computer immunological method by using an automaton to represent a program's normal behavior. The program counter is used as the state of the automaton, and the system calls made by the program are used as the events that cause the state transitions. As a result, the automaton representation can accommodate more information about the programs' normal behavior and, thus, reduce false alert rate and improve detection rate. In addition, the automaton representation is more compact than the previous alternative. Consequently, such automata are easier to build and more efficient to use for intrusion detection.

Specification-Based Methods

Ko, Ruschitzka, & Levitt (1997) proposed a specification-based approach for intrusion detection. The idea is to use *traces*, ordered sequences of execution events, to specify the intended behaviors of concurrent programs in distributed system. A specification describes valid operation sequences of the execution of one or more programs, collectively called a (*monitored*) *subject*. A sequence of operations performed by the subject that does not conform to the specification is considered a security violation. Each specification is called a *trace policy*. A grammar called *parallel environment grammars* (PE-grammars) was developed for specifying trace policies.

The advantage of this approach is that in theory, it should be able to detect some new types of attacks that intruders will invent in the future. The drawback of this approach is that substantial work is required to specify accurately the behavior of the many privileged system programs, and these specifications will be operating-system specific. To address this issue, Ko (2000) proposed the use of *inductive logic programming* to synthesize specifications from valid traces. The automatically generated specifications may be combined with manual rules to reduce the work involved in specification of valid program behaviors.

Wagner and Dean (2001) further advanced the specification-based approach. The basic idea is to automatically generate the specification of a program by deriving an abstract model of the programs from the source or binary code. Wagner and Dean studied several alternative models, including the call-graph model and the abstract stack model. Central to these models is the control flow graph of a program; these models adopt different ways to represent the possible system call traces according to the control flow graph. Attractive features of this approach are that it has the potential to detect unknown patterns of attacks and it has no false alerts, although it may miss some attacks.

Information-Theoretic Measures

Lee and Xiang (2001) proposed the use of information-theoretic measures to help understand the characteristics of audit data and build anomaly detection models. The well-known concept *entropy* is the first information-theoretic measure. Given a set of classes C_X and a data set X, where each data item belongs to a class $x \in C_X$, the entropy of X relative to C_X is defined as

$$H(X) = \sum_{x \in C_X} P(x) \log \frac{1}{P(x)},$$

where $P(x)$ is the probability of x in X. For anomaly detection, entropy reveals the regularity of audit data with respect to some given classes.

The second information-theoretic measure is *conditional entropy*. The conditional entropy of X given Y is the entropy of the probability distribution $P(x \mid y)$; that is,

$$H(X \mid Y) = \sum_{x,y \in C_X, C_Y} P(x,y) \log \frac{1}{P(x \mid y)},$$

where $P(x,y)$ is the joint probability of x and y and $P(x \mid y)$ is the conditional probability of x given y. The conditional entropy is proposed to measure the temporal or sequential characteristics of audit data. Let X be a collection of sequences where each is a sequence of n audit events and where Y is the collection of prefixes of the sequences in X that are of length k. Then $H(X \mid Y)$ indicates the uncertainty that remains for the rest of the audit events in a sequence x after we have seen the first k events of x. For anomaly detection, conditional entropy can be used as a measure of regularity of sequential dependencies.

The limitation of conditional entropy is that it only measures the sequential regularity of contiguous events. For example, a program may generate highly regular sequences of events; however, if these events are interleaved with other sequences of events, the conditional entropy will be very high, failing to reflect the regularity embedded in the interleaved sequences.

The third information-theoretic measure, *relative entropy*, measures the distance of the regularities between two data sets. The relative entropy between two probability distributions $p(x)$ and $q(x)$ that are defined over the same $x \in C_X$ is

$$\text{relEntropy}(p \mid q) = \sum_{x \in C_X} p(x) \log \frac{p(x)}{q(x)}.$$

The fourth information-theoretic measure, *relative conditional entropy*, measures the distance of the regularities with respect to the sequential dependency between two data sets. The relative entropy between two probability distributions $p(x \mid y)$ and $q(x \mid y)$ that are defined over the same $x \in C_X$ and $y \in C_Y$ is

$$\text{relCondEntropy}(p \mid q) = \sum_{x,y \in C_X,C_Y} p(x,y) \log \frac{p(x \mid y)}{q(x \mid y)}.$$

Viewing intrusion detection as a classification problem, Lee and Xiang proposed the fifth information-theoretic measure, *information gain*, to measure the performance of using some features for classification. The information gain of attribute (i.e., feature) A on data set X is

$$\text{Gain}(X, A) = H(X) - \sum_{v \in \text{Values}(A)} \frac{|X_v|}{|X|} H(X_v),$$

where Values(A) is the set of values of A and X_v is the subset of X where A has value v. This measure can help choose the right features (i.e., the features that have high information gain) to build intrusion detection models. The limitation of information gain is that it requires a relatively complete data set to help choose the right features for the classification model. Nevertheless, an intrusion detection model cannot be better than the data set from which it is built.

Limitation of Anomaly Detection

Although anomaly detection can accommodate unknown patterns of attacks, it also suffers from several drawbacks. A common problem of all anomaly detection approaches, with the exception of the specification-based approach, is that the subject's normal behavior is modeled on the basis of the (audit) data collected over a period of normal operation. If undiscovered intrusive activities occur during this period, they will be considered normal activities. In addition, because a subject's normal behavior usually changes over time (for example, a user's behavior may change when he moves from one project to another), the IDSs that use the above approaches usually allow the subject's profile to gradually change. This gives an intruder the chance to gradually train the IDS and trick it into accepting intrusive activities as normal. Also, because these approaches are all based on summarized information, they are insensitive to stealthy attacks. Finally, because of some technical reasons, the current anomaly detection approaches usually suffer from a high false-alarm rate.

Another difficult problem in building anomaly detection models is how to decide the features to be used as the input of the models (e.g., the statistical models). In the existing models, the input parameters are usually decided by domain experts (e.g., network security experts) in ad hoc ways. It is not guaranteed that all and only the features related to intrusion detection will be selected as input parameters. Although missing important intrusion-related features makes it difficult to distinguish attacks from normal activities, having non-intrusion-related features could introduce "noise" into the models and thus affect the detection performance.

MISUSE DETECTION

Misuse detection is considered complementary to anomaly detection. The rationale is that known attack patterns can be detected more effectively and efficiently by using explicit knowledge of them. Thus, misuse detection systems look for well-defined patterns of known attacks or vulnerabilities; they can catch an intrusive activity even if it is so negligible that the anomaly detection approaches tend to ignore it.

The major problem in misuse detection is how to represent known patterns of attacks. The detection algorithms usually follow directly from the representation mechanisms. In this section, we discuss the typical ways to represent attacks.

Rule-Based Languages

The rule-based expert system is the most widely used approach to misuse detection. The patterns of known attacks are specified as rule sets, and a forward-chaining expert system is usually used to look for signs of intrusions. Here we discuss two rule-based languages, rule-based sequence evaluation language (RUSSEL) (Mounji, Charlier, Zampunieris, & Habris, 1995) and production-based expert system tool set (P-BEST) (Lindqvist & Porras, 1999). Other rule-based languages exist, but they are all similar in the sense that they all specify known attack patterns as event patterns.

RUSSEL

RUSSEL is the language used in the advanced security audit trail analysis on UNIX (ASAX) project (Mounji, Charlier, Zampunieris, & Habris, 1995). It is a language specifically tailored to the problem of searching arbitrary patterns of records in sequential files. The language provides common control structures, such as conditional, repetitive, and compound actions. Primitive actions include assignment, external routine call, and rule triggering. A RUSSEL program simply consists of a set of rule declarations that are made of a rule name, a list of formal parameters and local variables, and an action part. RUSSEL also supports modules sharing global variables and exported rule declarations.

When intrusion detection is being enforced, the system analyzes the audit records one by one. For each audit record, the system executes all the active rules. The execution of an active rule may trigger (activate) new rules, raise alarms, write report messages, or alter global variables, for example. A rule can be triggered to be active for the current or the next record. In general, a rule is active for the current record because a prefix of a particular sequence of audit records has been detected. When all the rules active for the current record have been executed, the next record is read and the rules triggered for it in the previous step are executed in turn. User-defined and built-in C-routines can be called from a rule body.

RUSSEL is quite flexible in describing sequential event patterns and corresponding actions. The ability to work with user-defined C-routines gives the users the power to describe almost anything that can be specified in a programming language. The disadvantage is that it is a low-level language. Specifying an attack pattern is similar to writing a program, although it provides a general condition-trigger framework and is declarative in nature. The feature that rules can share global variables introduces the possibility of bugs along with the convenience of sharing information among different rules.

P-BEST

P-BEST was developed for the multiplexed information and computing service (Multics) intrusion detection and alerting system (MIDAS) and later employed by the intrusion detection expert system (IDES), NIDES, and the event monitoring enabling responses to anomalous live disturbances (EMERALD) (Lindqvist Y. Porras, 1999). The P-BEST toolset consists of a rule translator, a library of runtime routines, and a set of garbage collection routines. Rules and facts in P-BEST are written in production rule specification language. The rule translator is then used to translate the specification into an expert system program in C language, which can then be compiled into either a stand-alone, self-contained executable program or a set of library routines that can be linked to a larger software framework.

The P-BEST language is quite small and intuitive. In P-BEST, the user specifies the structure of a fact (e.g., an audit record) through a template definition referred to as a *pattern type*. For example, an event consisting of four fields—*event_type* (an integer), *return_code* (an integer), *username* (a string), and *hostname* (a string)—can be defined as ptype[event *event_type*: int, *return_code*: int, *username*: string, *hostname*: string].

Thus, P-BEST does not depend on the structure of the input data. One important advantage of P-BEST is that it is a language preprocessor (i.e., it generates a precompiled expert system) and can extend its ability by invoking external C functions. However, it shares a similar problem with RUSSEL: It is a low-level language. Specification of attack patterns in P-BEST is time consuming. When many related rules are included in a system, correctness of the rules is difficult to check due to the interaction of these rules.

State Transition Analysis Tool Kit

Though rule-based languages are flexible and expressive in describing attack patterns for misuse detection, in practice, they are usually difficult to use. As observed in Ilgun, Kemmerer, and Porras (1995) "in general, expert rule-bases tend to be non-intuitive, requiring the skills of experienced rule-base programmers to update them." STAT was developed to address this problem.

In STAT, state transition analysis technique was adopted to facilitate the specification of the patterns of known attacks (Ilgun, Kemmerer, & Porras, 1995). It is based on the assumption that all penetrations share two common features. First, penetrations require the attacker to possess some minimum prerequisite access to the target system. Second, all penetrations lead to the acquisition of some ability that the attacker does not have prior to the attacks. Thus, STAT views an attack as a sequence of actions performed by an attacker that leads from some initial state on a system to a target-compromised state, where a *state* is a snapshot of the system representing the values of all memory locations on the system. Accordingly, STAT models attacks as a series of state changes that lead from an initial secure state to a target-compromised state.

To represent attacks, STAT requires some critical actions, called *signature actions*, to be identified. Signature actions refer to the actions that, if omitted from the execution of an attack scenario, would prevent the attack from successful completion. With the series of state changes and the signature actions that cause the state changes, an attack scenario is then represented as a state transition diagram, where the states in the diagram are specified by assertions of certain conditions and the signature actions are events observable from, for example, audit data.

STAT has been applied for misuse detection in UNIX systems, distributed systems, and networks. USTAT is the first prototype of STAT, which is aimed at misuse detection in UNIX systems (Ilgun, Kemmerer, & Porras, 1995). It relies on Sun Microsystems' C2-Basic Security Module (BSM) to collect audit records. In addition to detecting attacks, USTAT is designed to be a real-time system that can preempt an attack before any damage can be done. USTAT was later extended to process audit data collected on multiple UNIX hosts. The resulting system is called *NSTAT*. NSTAT runs multiple daemon programs on the hosts being protected to read and forward audit data to a centralized server, which performs STAT analysis on all data.

A later application of STAT to network-based misuse detection resulted in another system, named *NetSTAT* (Vigna & Kemmerer, 1999). In this work, the network topology is further modeled as a hypergraph. Network interfaces, hosts, and links are considered the constituent elements of hypergraphs, with interfaces as the nodes, and hosts and links as hyperedges. Using the network topology model and the state transition description of network-based attacks, NetSTAT can map intrusion scenarios to specific network configurations and generate and distribute the activities to be monitored at certain places in a network.

STAT was intended to be a high-level tool to help specify attack patterns. Using STAT, the task of describing an attack scenario is much easier than using rule-based languages, although the analysis required to understand the nature of attacks remains the same. In the implementations of STAT techniques (i.e., USTAT, NSTAT, and NetSTAT), the attack scenarios are transformed into rule bases, which are enforced by a forward-chaining inference engine.

Colored Petri Automata

Kumar and Spafford (1994) and Kumar (1995) viewed misuse detection as a pattern-matching process. They proposed an abstract hierarchy for classifying intrusion signatures (i.e., attack patterns) based on the structural interrelationships among the events that compose the signature. Events in such a hierarchy are high-level events that can be defined in terms of low-level audit trail events and used to instantiate the abstract hierarchy into a concrete one. A benefit of this classification scheme is that it clarifies the complexity of detecting the signatures in each level of the hierarchy. In addition, it also identifies the requirements that patterns in all categories of the classification must meet to represent the full range of commonly occurring intrusions (i.e., the specification of context, actions, and invariants in intrusion patterns).

Kumar and Spafford adopted colored Petri nets to represent attack signatures, with guards to represent signature contexts and vertices to represent system states. User-specified actions (e.g., assignments to variables) may be associated with such patterns and then executed when patterns are matched. The adapted colored Petri nets are called *colored Petri automata* (CPA). A CPA represents the transition of system states along paths that lead to intruded states. A CPA is also associated with pre- and postconditions that must be satisfied before and after the match, as well as invariants (i.e., conditions) that must be satisfied while the pattern is being matched. CPA has been implemented in a prototype misuse detection system called Intrusion Detection In Our Time (IDIOT).

CPA is quite expressive; it provides the ability to specify partial orders, which in turn subsume sequences and regular expressions. However, if improperly used, the expressiveness may lead to potential problems: If the intrusions are described in every detail, the attacker may be able to change his/her attacking strategy and bypass the IDSs. Nevertheless, CPA is not the root the problem.

Automatically Build Misuse Detection Models

Lee and Stolfo (2000) looked at intrusion detection as a data analysis process and applied several data mining techniques to build misuse detection models. The research efforts were conducted under a project entitled JAM, the Java Agent for Meta-learning (meta-learning is a general strategy that provides the means of learning how to combine and integrate a number of separately learned classifiers or models). In particular, association rules and frequent episodes are used to automatically discover features that should be used to model a subject's behavior, and the meta classification is used to combine the results of different classifiers to get better classification results.

Lee and Stolfo extended the original association rules to take into account the "order of importance" relations among the system features (the notion of association rule was discussed above regarding ADAM). *Axis* refers to the features that are important to intrusion detection; only the association rules involving axis features are considered. For example, in the shell command records, the command is likely to reflect the intrusive activities and thus be identified as an axis feature. In some sense, axis features incorporate expert knowledge into the system and thus improve the effectiveness of association rules.

To represent frequent sequential patterns of network events, Lee et al. extended frequent episodes, which were originally proposed in Mannila, Toivonen, and Verkamo (1995), to represent the sequential interaudit record patterns. Their algorithm finds frequent sequential patterns in two phases. First, it finds the frequent associations among event attributes using the axis attributes, and then it generates the frequent sequential patterns from these associations. The algorithm also takes advantage of the "reference" relations among the system features. That is, when forming an episode, the event records covered by the constituent item sets of the episode share the same value for a given feature attribute. The mined frequent episodes are used to construct temporal statistical features, which are used to build classification models. Thus, new features can be derived from the training data set and then used for generating better intrusion detection models.

Another innovation of the JAM project is meta classification, which combines the output of several base classifiers and generates the best results out of them. Specifically, from the predictions of base classifiers and the correct classes, a meta classifier learns which base classifier can make the best prediction for each type of input. It then chooses the prediction of the best base classifier for different input and combines the powers of the base classifiers.

Lee and Stolfo (2000) advanced state-of-the-art knowledge of intrusion detection by introducing a framework that helps generate intrusion detection models automatically. The limitation is that the framework depends on the volume of the evidences. That is, intrusive activities must generate a relatively noticeable set of events so the association of event attributes or frequent episodes can reflect them. Thus, the generated models must work with some other complementary systems, such as NetSTAT.

Abstraction-Based Intrusion Detection

The implementation of many misuse detection approaches shares a common problem: Each system is written for a single environment and has proved difficult to use in other environments that may have similar policies and concerns. The primary goal of abstraction-based intrusion detection is to address this problem.

The initial attempt of the abstraction-based approach is a misuse detection system named the adaptable real-time misuse detection system (ARMD) (Lin, Wang, & Jajodia, 1998). ARMD provides a high-level language for abstract misuse signatures, called MuSigs, and a mechanism to translate MuSigs into a monitoring program. With the notion of abstract events, the high-level language specifies a MuSig as a pattern over a sequence of abstract events, which is described as conditions that the abstract event attributes must satisfy. The gap between abstract events and audit records is bridged by an audit subsystem, which transforms the actual audit records into abstract events. In addition, on the basis of MuSigs, the available audit trail, and the strategy costs, ARMD uses a strategy generator to automatically generate monitoring strategies to govern the misuse detection process.

ARMD is a host-based misuse detection system. In addition to the features mentioned above, it also employs database query optimization techniques to speed up the processing of audit events. The experiences with ARMD show that knowing the characteristics of the audit trail helps estimate the cost of performing misuse detection and gives the security officers the opportunity to tune the misuse detection system.

A limitation of ARMD is that it requires users to have a precise understanding of the attacks and to make careful plans for the abstraction of events. This planning is not an easy job, especially when a user does not know how his/her MuSigs may be used. In particular, unforeseen attacks may invalidate previously defined abstract events and MuSigs, thus forcing the redevelopment of some/all of the MuSigs.

The work by Ning, Jajodia, and Wang (2001) further extends the result in ARMD to address the aforementioned limitation. It provides a framework for distributed attack specification and event abstraction. In this framework, abstraction is considered an ongoing process. The structures of abstract events are represented as system views, and attack signatures are represented as generic patterns on the basis of system views. This new approach allows the semantics of a system view to be modified by defining new signatures and view definitions without changing the specifications of the views or the signatures specified on the basis of the system views.

As a result, signatures in this model can potentially accommodate unknown variants of known attack patterns. Although the specification of attack signatures and the choice of right abstraction still partially depend on the users' skill, this framework provides guidance and alleviates the burden of writing and maintaining signatures.

Limitation of Misuse Detection

Current misuse detection systems usually work better than anomaly detection systems for known attacks. That is, misuse detection systems detect patterns of known attacks more accurately and generate much fewer false alarms. This better performance occurs because misuse detection systems take advantage of explicit knowledge of the attacks.

The limitation of misuse detection is that it cannot detect novel or unknown attacks. As a result, the computer systems protected solely by misuse detection systems face the risk of being comprised without detecting the attacks. In addition, due to the requirement of explicit representation of attacks, misuse detection requires the nature of the attacks to be well understood. This implies that human experts must work on the analysis and representation of attacks, which is usually time consuming and error prone. Lee and Stolfo (2000) have improved this process by automatically building the intrusion detection model; however, identification of attacks in past events is still required.

INTRUSION DETECTION IN DISTRIBUTED SYSTEMS

The rapid growth of the Internet not only provides the means for resource and information sharing, but it also brings new challenges to the intrusion detection community. Due to the complexity and the amount of audit data generated by large-scale systems, traditional IDSs, which were designed for individual hosts and small-scale networked systems, cannot be directly applied to large-scale systems.

Research on intrusion detection in distributed systems is currently focusing on two essential issues: *scalability* and *heterogeneity*. The IDSs in large distributed systems need to be scalable to accommodate the large amount of audit data in such systems. In addition, such IDSs must be able to deal with heterogeneous information from component systems of different types and that constitute large distributed systems and can cooperate with other types of IDSs.

Research on distributed intrusion detection is being conducted in three main areas. First, people are building scalable, distributed IDSs or are extending existing IDSs to make them capable of being scaled up to large systems. Second, network-based IDSs are being developed to take advantage of the standard network protocols to avoid heterogeneous audit data from different platforms. Third, standards and techniques are being developed to facilitate information sharing among different, possibly heterogeneous IDSs.

Distributed Intrusion Detection Systems

Early distributed IDSs collect audit data in a distributed manner but analyze the data in a centralized place, for example, DIDS (Snapp et al., 1991) and ASAX (Mounji, Charlier, Zampunieris, & Habra, 1995). Although audit data is usually reduced before being sent to the central analysis unit, the scalability of such systems is still limited. When the size of the distributed system grows large, not only might audit data have to travel long distances before arriving at the central place, but the central analysis component of the IDS may be overwhelmed by large amount of audit data being generated.

Recent systems, such as EMERALD (Lindqvist & Porras, 1999), GrIDS (graph-based intrusion detection systems; Axelsson, 1999), and AAFID (the autonomous agents for intrusion detection system; Spafford &

Zamboni, 2000), pay more attention to the scalability issue. To scale up to large distributed systems, these systems place IDS components in various places in a distributed system. Each of these components receives audit data or alerts from a limited number of sources (e.g., hosts or other IDS components), so the system is not overwhelmed by large amounts of audit data. Different components are often organized hierarchically in a tree structure; lower level IDS components disseminate their detection results to higher level components, so the intrusion related information from different locations can be correlated together.

Although most of the recent distributed IDSs are designed to be scalable, they only provide a partial solution to the scalability problem. The IDS components are either coordinated in an ad hoc way or are organized hierarchically. Although coordinating the IDS components in an ad hoc way is certainly not a general solution, organizing the components hierarchically does not always provide an efficient solution, especially when the suspicious activities are spread in different, unpredictable locations in a large system. In a hierarchical system, when the activities involved in a distributed attack fall beyond the scope of one IDS component, the audit data possibly related to the attack will have to be forwarded to a higher level IDS component to be correlated with data from other places. In a worst-case scenario, the audit data may have to be forwarded several times before arriving at a place where the data can finally be correlated with the related information. This process not only wastes the network's bandwidth, it also limits the scalability of the detection of distributed attacks.

The abstraction-based intrusion detection (Ning, Jajodia, & Wang, 2001) addresses this problem by generating a hierarchy of IDS components dynamically rather than statically. Intuitively, this method defines attack signatures as generic patterns (called *generic signatures*) of abstract events that may be observed in different places in a distributed system. When a particular type of attack is to be detected in a distributed system, the corresponding generic signature is mapped to the specific systems. The resulting signature is called a *specific signature*. This method then decomposes the specific signature into components called *detection tasks*, each of which corresponds to the intrusion detection activities required to process a type of event involved in the attack. A coordination mechanism is developed to arrange the messages passing between the detection tasks, so the distributed detection of attacks is equivalent to having all events processed in a central place. The abstraction-based method is more flexible and more efficient than the previous methods; however, it is limited in that it is applicable only to misuse detection.

Network-Based Intrusion Detection Systems

Network-based IDSs collect audit data from the network traffic, as opposed to host-based IDSs, which usually collect audit data from host audit trails. Examples of network-based IDSs include NSM (Network Security Monitor; Axelsson, 1999), NetSTAT (Vigna & Kemmerer, 1999), and Bro (Axelsson, 1999).

One challenge that the network-based intrusion detection is facing is the speed of high-performance networks. The great speed of the high-performance networks makes it very difficult to capture the network traffic, let alone perform intrusion detection in real time.

Several efforts have addressed enabling intrusion detection in high-speed networks. Snort, an open source IDS that specializes in network intrusion detection (Snort, 2002), was developed by Roesch (1999). It employs a fast pattern-matching algorithm to detect network misuse. However, early versions of Snort detected attacks individually, and its performance degrades when the number of attack signatures (rules) increases. Since version 1.9, Snort has incorporated new pattern-matching algorithms to address this problem.

Sekar, Guang, Verma, and Shanbhag (1999) developed a high-performance network IDS based on efficient pattern-matching algorithms. A distinguishing feature is that the performance of the system is independent of the number of misuse rules (signatures).

Kruegel, Caleur, Vigna, and Kemmerer (2002) proposed a partition approach to intrusion detection that supports misuse detection on high-speed network links. This approach is based on a slicing mechanism that divides the overall network traffic into subsets of manageable size. Thus, each subset can be processed by one or several misuse detection systems. The traffic partitioning is done such that each subset of the network traffic contains all of the evidence necessary to detect a specific attack.

Network-based IDSs offer several advantages. First, network-based IDSs can take advantage of the standard structure of network protocols, such as TCP/IP. This is a good way to avoid the confusion resulting from heterogeneity in a distributed system. Second, network-based IDSs usually run on a separate (dedicated) computer; thus, they do not consume the resources of the computers that are being protected.

Conversely, network-based IDS are not silver bullets. First, because these IDSs do not use host-based information, they may miss the opportunity to detect some attacks. For example, network-based IDSs cannot detect an attack launched from a console. Second, the standard network protocols do not solve the entire problem related to the heterogeneity of distributed systems because of the variety of application protocols and systems that use these protocols. For example, network-based IDSs must understand the UNIX shell commands if their goal is to monitor intrusive remote log-ins. As another example, network-based IDSs usually describe the suspicious network activities using the structure of the packet that standard network protocols such as TCP/IP support, which makes the specification of the suspicious activities to be detected very difficult.

Network-based IDSs have the same scalability problem as do general distributed IDSs. For example, existing network-based IDSs analyze network traffic data in a centralized place, as was done by the early distributed IDSs, although they may collect data from various places in the network. This structure limits the scale of the distributed systems that such IDSs can protect.

Sharing Information Among Intrusion Detection Systems

With the deployment of so many commercial IDSs, these IDSs must be able to share information so that they can interact and thus achieve better performance than if operating in isolation. Research and development activities are currently under way to enable different, possibly heterogeneous, IDSs to share information.

The Common Intrusion Detection Framework (CIDF) was developed to enable different intrusion detection and response (IDR) components to interoperate and share information and resources (Porras, Schnackenberg, Staniford-Chen, Stillman, & Wu, 1999). It began as a part of the Defense Advanced Research Project Agency (DARPA) Information Survivability program, with a focus on allowing DARPA projects to work together. CIDF considers IDR systems as composed of four types of components that communicate via message passing: event generators (E-boxes), event analyzers (A-boxes), event databases (D-boxes), and response units (R-boxes). A communication framework and a common intrusion specification language are provided to assist the interoperation among CIDF components.

Researchers involved in CIDF started an Intrusion Detection Working Group (IDWG) in the Internet Engineering Task Force (IETF), trying to bring the impact of CIDF to a broader community. The IDWG has been working to develop data formats and exchange procedures for sharing information among IDSs, response systems, and management systems. The extensible markup language (XML) has been chosen to provide the common format, and an intrusion detection message exchange format (IDMEF) has been defined in an Internet draft. The IDWG uses the blocks extensible exchange protocol (BEEP) as the application protocol framework for exchanging intrusion detection messages between different systems; an intrusion detection exchange protocol (IDXP) is specified as a BEEP profile, and a tunnel profile is provided to enable different systems to exchange messages through firewalls. At the time of this writing, these Internet drafts are under consideration for being adopted as IETF requests for comments (RFCs).

A negotiation protocol for CIDF components has been developed as part of the intrusion detection intercomponent adaptive negotiation (IDIAN) project (Feiertag, et al., 2000). It allows distributed IDS components to discover the intrusion detection services of other components, and to negotiate, manage, and adjust the use of those services. In particular, the notion of a filter was introduced to assist the negotiation process. A filter is essentially a pattern of CIDF messages, and the negotiation process helps determine one or several filters between two CIDF components. A CIDF component sends a CIDF message to another component only when the message matches the filters of the receiver. Similarly, a receiving CIDF component accepts an incoming CIDF message only if the message matches one of its filters. IDIAN partially addresses the information-sharing problem by enabling different IDS components to discover, negotiate, and adjust the use of services from each other.

Another effort in sharing information among different IDSs is the Hummer project (Frincke, Tobin, McConnell, Marconi, & Polla, 1998). In particular, the relationships among different IDSs (e.g., peer, friend, manager/subordinate) and policy issues (e.g., access control policy, cooperation policy) were studied, and a prototype system, HummingBird, was developed to address these issues. A limitation of the Hummer project is that it only addresses the general data-sharing issue; what information needs to be shared and how the information would be used are out of its scope. Thus, it should be used along with mechanisms such as IDIAN (Feiertag, et al., 2000) and the decentralized coordination mechanism in the abstraction-based approach (Ning, Jajodia, & Wang, 2001).

CONCLUSION

Intrusion detection continues to be an active research field. Even after 20 years of research, the intrusion detection community still faces several difficult problems. How to detect unknown patterns of attacks without generating too many false alerts remains an unresolved problem, although recently, several results have shown there is a potential resolution to this problem. The evaluating and benchmarking of IDSs is also an important problem, which, once solved, may provide useful guidance for organizational decision makers and end users. Moreover, reconstructing attack scenarios from intrusion alerts and integration of IDSs will improve both the usability and the performance of IDSs. Many researchers and practitioners are actively addressing these problems. We expect intrusion detection to become a practical and effective solution for protecting information systems.

GLOSSARY

Anomaly detection One of the two methodologies of intrusion detection. Anomaly detection is based on the normal behavior of a subject (e.g., a user or a system); any action that significantly deviates from the normal behavior is considered intrusive. The other methodology is misuse detection.

Audit trail Records a chronology of system resource usage. This includes user log-in, file access, other various activities, and whether any actual or attempted security violations occurred.

False negative An actual intrusive action that the system allows to pass as nonintrusive behavior.

False positive Classification of an action as anomalous (a possible intrusion) when it is legitimate.

IDS Intrusion detection system.

Intrusion Any activity that violates the security policy of an information system.

Intrusion detection The process of identifying intrusions by observing security logs, audit data, or other information available in computer systems and/or networks.

Misuse detection One of the two methodologies of intrusion detection. Catches intrusions in terms of the characteristics of known patterns of attacks or system vulnerabilities; any action that conforms to the pattern of a known attack or vulnerability is considered intrusive. The other methodology is anomaly detection.

Misuse signature A known pattern or attack or vulnerability, usually specified in a certain attack specification language.

Profile A set of parameters used to capture the pattern of a subject's (e.g., a user or a program) normal behavior. It is normally used to conduct anomaly detection.

Security policy A set of rules and procedures regulating the use of information, including its processing, storage, distribution, and presentation.

CROSS REFERENCES

See *Data Mining in E-Commerce; Guidelines for a Comprehensive Security System; Machine Learning and Data Mining on the Web.*

REFERENCES

Axelsson, S. (1999). *Research in intrusion-detection systems: A survey. Technical report TR 98-17.* Göteborg, Sweden: Department of Computer Engineering, Chalmers University of Technology.

Barbara, D., Couto, J., Jajodia, S., & Wu, N. (2001). ADAM: A testbed for exploring the use of data mining in intrusion detection. *ACM SIGMOD Record, 30* (4), 15–24.

Barbara, D., Couto, J., Jajodia, S., & Wu, N. (2002). An architecture for anomaly detection. In D. Barbara & S. Jajodia (Eds.), *Applications of Data Mining in Computer Security* (pp. 63–76). Boston: Kluwer Academic.

Feiertag, R., Rho, S., Benzinger, L., Wu, S., Redmond, T., Zhang, C., Levitt, K., Peticolas, D., Heckman, M., Staniford, S., & McAlerney, J. (2000). Intrusion detection inter-component adaptive negotiation. *Computer Networks, 34,* 605–621.

Fox, K. L., Henning, R. R., Reed, J. H., & Simonian, R. P. (1990). A neural network approach towards intrusion detection. In *Proceedings of 13th National Computer Security Conference* (pp. 125–134). Baltimore, MD: National Institute of Standards and Technology.

Frincke, D., Tobin, D., McConnell, J., Marconi, J., & Polla, D. (1998). A framework for cooperative intrusion detection. In NIST (Ed.), *Proceedings of the 21st National Information Systems Security Conference* (pp. 361–373). Baltimore, MD: National Institute of Standards and Technology.

Ghosh, A. K., Wanken, J., & Charron, F. (1998). Detecting anomalous and unknown intrusions against programs. In K. Keus (Ed.), *Proceedings of the 14th Annual Computer Security Applications Conference* (pp. 259–267). Los Alamitos, CA: IEEE Computer Society.

Hofmeyr, S., Forrest, S., & Somayaji, A. (1998). Intrusion detection using sequences of system calls. *Journal of Computer Security, 6,* 151–180.

Ilgun, K., Kemmerer, R. A., & Porras, P. A. (1995). State transition analysis: A rule-based intrusion detection approach. *IEEE Transactions on Software Engineering, 21* (3), 181–199.

Ko, C., Ruschitzka, M., & Levitt, K. (1997). Execution monitoring of security-critical programs in distributed systems: A apecification-based approach. In G. Dinolt & P. Karger (Eds.), *Proceedings of 1997 IEEE Symposium of Security and Privacy* (pp. 175–187). Los Alamitos, CA: IEEE Computer Society.

Ko, C. (2000). Logic induction of valid behavior specifications for intrusion detection. In M. Reiter & R. Needham (Eds.), *Proceedings of 2000 IEEE Symposium of Security and Privacy* (pp. 142–153). Los Alamitos, CA: IEEE Computer Society.

Kruegel, C., Valeur, F., Vigna, G., & Kemmerer, R. (2002). Stateful intrusion detection for high-speed networks. In M. Abadi & S. Bellovin (Eds.), *Proceedings of 2002 IEEE Symposium on Security and Privacy* (pp. 285–293). Los Alamitos, CA: IEEE Computer Society.

Kumar, S. (1995). *Classification and detection of computer intrusions.* Unpublished doctoral dissertation, Purdue University, West Lafayette, IN.

Kumar, S., & Spafford, E. H. (1994). A pattern-matching model for misuse intrusion detection. In *Proceedings of the 17th National Computer Security Conference* (pp. 11–21). Baltimore, MD: National Institute of Standards and Technology.

Lane, T., & Brodley, C. E. (1998). Temporal sequence learning and data reduction for anomaly detection. In L. Gong & M. Reiter (Eds.), *Proceedings of 5th Conference on Computer and Communications Security* (pp. 150–158). New York: ACM Press.

Lee, W., & Stolfo, S. J. (2000). A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information and System Security, 3* (4), 227–261.

Lee, W., & Xiang, D. (2001). Information-theoretic measures for anomaly detection. In R. Needham & M. Abadi (Eds.), *Proceedings of 2001 IEEE Symposium on Security and Privacy* (pp. 130–143). Los Alamitos, CA: IEEE Computer Society.

Lin, J., Wang, X. S., & Jajodia, S. (1998). Abstraction-based misuse detection: High-level specifications and adaptable strategies. In S. Foley (Ed.), *Proceedings of the 11th Computer Security Foundations Workshop* (pp. 190–201). Los Alamitos, CA: IEEE Computer Society.

Lindqvist, U., & Porras, P. A. (1999). Detecting computer and network misuse through the production-based expert system toolset (P-BEST). In L. Gong & M. Reiter (Eds.), *Proceedings of the 1999 IEEE Symposium on Security and Privacy* (pp. 146–161). Los Alamitos, CA: IEEE Computer Society.

Mannila, H., Toivonen, H., & Verkamo, A. I. (1995). Discovering frequent episodes in sequences. In U. Fayyad & R. Uthurusamy (Eds.), *Proceedings of the 1st Conference on Knowledge Discovery and Data Mining* (pp. 210–215). Menlo Park, CA: AAAI Press.

Mounji, A., Charlier, B. L., Zampuniéris, D., & Habra, N. (1995). Distributed audit trail analysis. In D. Balenson & R. Shirey (Eds.), *Proceedings of the ISOC'95 Symposium on Network and Distributed System Security* (pp. 102–112). Los Alamitos, CA: IEEE Computer Society.

Ning, P., Jajodia, S., & Wang, X. S. (2001). Abstraction-based intrusion detection in distributed environments. *ACM Transactions on Information and System Security, 4* (4), 407–452.

Porras, P., Schnackenberg, D., Staniford-Chen, S., Stillman, M., & Wu, F. (1999). *Common intrusion detection framework architecture*. Retrieved on February 13, 2003 from http://www.isi.edu/gost/cidf/

Roesch, M. (1999). Snort–lightweight intrusion detection for networks. In D. Parter (Ed.), *Proceedings of the 13th Systems Administration Conference*. Retrieved on February 13, 2003 from http://www.usenix.org/publications/library/proceedings/lisa99/technical.html

Sekar, R., Bendre, M., Dhurjati, D., & Bollineni, P. (2001). A fast automaton-based method for detecting anomalous program behaviors. In R. Needham & M. Abadi (Eds.), *Proceedings of 2001 IEEE Symposium on Security and Privacy* (pp. 144–155). Los Alamitos, CA: IEEE Computer Society.

Sekar, R., Guang, Y., Verma, S., & Shanbhag, T. (1999). A high-performance network intrusion detection system. In J. Motiwalla & G. Tsudik (Eds.), *Proceedings of the 6th ACM Conference on Computer and Communications Security* (pp. 8–17). New York: ACM Press.

Snapp, S. R., et al. (1991). DIDS (distributed intrusion detection system)—Motivation, architecture, and an early prototype. In *Proceedings of 14th National Computer Security Conference* (pp. 167–176).

Snort—The open source intrusion detection system. (2002). Retrieved February 13, 2003, from http://www.snort.org

Spafford, E. H., & Zamboni, D. (2000). Intrusion detection using autonomous agents. *Computer Networks 34,* 547–570.

Teng, H. S., Chen, K., & Lu, S. C. (1990). Adaptive real-time anomaly detection using inductively generated sequential patterns. In *Proceedings of 1990 IEEE Symposium on Security and Privacy* (pp. 278–284), Los Alamitos, CA: IEEE Computer Society.

Vigna, G., & Kemmerer, R. A. (1999). NetSTAT: A network-based intrusion detection system. *Journal of Computer Security, 7* (1), 37–71.

Wagner, D., & Dean, D. (2001). Intrusion detection via static analysis. In R. Needham & M. Abadi (Eds), *Proceedings of 2001 IEEE Symposium on Security and Privacy* (pp. 156–168). Los Alamitos, CA: IEEE Computer Society.

Inventory Management

Janice E. Carrillo, *University of Florida*
Michael A. Carrillo, *Oracle Corporation*
Anand Paul, *University of Florida*

INTRODUCTION

Motivation: The Importance of Effective Inventory Management

According to a recent study, U.S. firms in retail managed $1.1 trillion in inventory to support $3.2 trillion in annual sales in the year 2002 (C. Lee, 2002). This inventory was spread across enterprise supply chains with $400 billion in retail, $290 billion in distribution centers, and $450 billion in manufacturing. Although the benefits associated with carrying such high levels of inventory typically include higher service levels, this is not always the case. During the same time period, approximately 8% of retail customers were unable to find their products, a supply chain stock-out glitch, which led to a 3% sales revenue opportunity loss for U.S. retailers. The crux of these problems lies in a firm's inability to manage its inventory effectively throughout the supply chain.

Furthermore, the long-term implications of poor inventory management are far reaching, ultimately having an impact on the market value of the firm. Supply chain shortfalls, stock-outs, and product delays significantly affect shareholder value by as much as 18% (Singhal & Hendricks, 2002). To illustrate, Boeing allocated $1 billion in 1997 for "production disruption" costs incurred for parts or supply shortages associated with production ramp-up for airplane production. Similarly, earnings for the candy maker Hershey's in 1999 were low because of problems created by the unsuccessful implementation of enterprise software, effectively incapacitating their order-taking and distribution systems during the crucial Halloween season. Therefore, it is clear that the market has a negative reaction of any firm that is unable to manage product supply and demand effectively within its value chain.

Inventory Measures in a Supply Chain Context

Although the importance of effective supply chain and distribution management is apparent, an appropriate set of metrics measuring a firm's performance relative to its supply chain is essential. Three important measures of supply chain success include those that assess the value of a firm's assets (including both inventory and cash), customer service and speed with which it can deliver its products (Johnson, 1998). Traditional inventory measures have included the monetary value of inventory, the time supply (e.g., 3 weeks of inventory), and inventory turnovers. The implications of effective inventory management now extend beyond the single enterprise to the entire supply chain. Moreover, the linkage between inventory management and other essential measures of supply chain success is critical for a firm's success in the marketplace.

Within the realm of supply chain metrics, customer service, a firm's ability to mediate supply with demand, is an important factor affecting business performance. A key measure of a firm's customer service level success requires the firm to ensure delivery of the right product at the right time at the right place in the right quantity while combating excessive inventory costs. The complexity of successfully delivering high customer service levels is further exacerbated by product uncertainties in demand (i.e., demand variability, short life cycles, proliferation of product variety, product volume, etc.) and supply (i.e., supplier sources, globalization, capacity constraints, process, quality, etc. [H. Lee, 2002]).

Consider the events of the eToys corporation during the 1999 holiday season (Hallowell, 2000). eToys, an online retailer of toys, experienced an unanticipated high volume of demand for their products. Although high demand in and of itself is usually a welcome phenomena amongst a firm's stockholders, eToys was also plagued with numerous supply problems that intensified its inability to meet demand. Items that were out of stock were actually listed as available on the company Web site. Orders were late or not delivered at all. Furthermore, service representatives were unable to utilize information systems effectively to answer customer questions concerning product tracking and delivery. In the short term, eToys's stock value plunged because of high costs of order fulfillment. Eventually, eToys was unable to recover from the supply debacle, and the firm went bankrupt.

One way that firms can manage such product uncertainties is through the adoption of appropriate supply chain strategies and subsequent inventory policies. In a highly competitive global environment, firms have to manage inventory within evolving distribution and supply channels at a rate commensurate with the firm's industry. To maintain competitive advantage in quick-paced industries, firms must continually evaluate how to manage inventory considering the following factors: new product introductions, sourcing, collaboration, differentiation, and lead-time compression initiatives. For example, firms manufacturing innovative products with short life cycles (typically 3 months to 1 year) require flexible and responsive supply chains that can facilitate introducing new products into the market quickly (Fisher, 1998). Utilizing inventory in the form of either parts or finished goods can be an effective buffer against uncertainty in demand and supply glitches. On the other hand, firms manufacturing functional products with longer life cycles (typically greater than 1 year) and predictable demand should focus on establishing an efficient supply chain to deliver the goods at the lowest cost. To facilitate such a strategy, firms in this environment require minimal levels of inventory and high rates of inventory turnovers.

Classical Inventory Management

Classical inventory techniques that address each of these unique environments are well established within the operations management area. For innovative products with high demand uncertainty, stochastic methods (such as reorder point variations) can be used to aid managers in inventory planning decisions. For functional products with longer life cycles and predictable demand, however, other deterministic models (such as EOQ) are appropriate. Factors typically used in these inventory decisions include the following relevant cost drivers: inventory holding, ordering and purchase, setup, stockout, and goodwill. Numerous variations of these inventory replenishment methodologies are embodied in enterprise software, including min–max, reorder point, and time-phased approaches for single and multi-echelon systems.

Technologies Enabling Effective Inventory Management

The introduction of Internet technologies has led to increases in a firm's operational efficiencies through real-time information sharing (e.g., key performance indicators, point-of-sale data) and decision sharing (e.g., vendor managed inventory) within its organization and across its dynamic supply chain. Firms today can select from among a number of Internet architected software products from supply chain management and enterprise resource planning vendors that no longer simply manage simple transactions (e.g., orders, invoices, etc.) but instead enable supply chain collaboration.

The purpose of this chapter is as follows. First, in the introduction, we provide an overview of the complexities inherent in inventory management and the implications of inventory management gone awry. Second, we present a summary of classic inventory management policies and offer guidance concerning the correct application of such policies under various circumstances. Third, we discuss more complex variations of inventory theory as it applies to planning for multiple items and multiple stocking points. Fourth, we compare and contrast traditional inventory fulfillment methods with more recently developed e-commerce models. Finally, we discuss how current Internet technologies facilitate effective inventory management. Conclusions summarizing key managerial implications for inventory management are included at the end of the chapter.

INVENTORY REPLENISHMENT METHODOLOGIES

There are two separate demand processes underlying every inventory system; it is essential to bear this in mind throughout the discussion that follows. The first demand process, that of consumer demand, may be regarded as an external one; the second demand process comprises internal requests for replenishment of stock. Clearly, the consumer demand process drives the stock replenishment process. In the real world, the consumer demand process is uncertain, to one degree or another, whereas the internal replenishment process unfolds by design. The essential task of the inventory planner is to use information about the consumer demand process to design rules governing the replenishment process.

It is relatively simple to devise optimal replenishment procedures for a single item with known demand, provided we have accurate cost estimates. At the other end of the spectrum, it is a complex task to craft an optimal inventory control system for a group of items with individual demands that are difficult to predict. The crux of the problem is to balance costs from excessive or obsolete inventories against the consequences of running out of inventory prematurely. This section offers a basic tutorial on inventory replenishment decisions for a single firm making a choice with regard to a single item with independent demand. In the following section, we generalize these policies for the multiple item, multi-echelon, and supply chain environments.

Table 1 Summary of Inventory Models

INVENTORY MODEL	DETERMINES	APPROPRIATE FOR	SIGNIFICANT FACTORS
Continuous review	Order frequency	Critical items with costly stock-outs	Information intensive
Periodic review	Order frequency	General purpose items	
Economic order quantity	Order quantity (inventory level)	Stable demand	Monthly demand Holding cost Fixed order cost Lead time
Reorder point	Order quantity (inventory level)	Uncertain demand	Average demand Standard deviation demand Customer service level
Min–max	Order quantity (inventory level)	Uncertain demand	Min stock level Max stock level
Order up-to	Order quantity (inventory level)	Uncertain demand	Max stock level

The optimal inventory replenishment decision breaks down into two interlinked choices: how often to order (i.e., frequency), and how much to order (i.e., quantity). In general, there are two approaches to determine how frequently to place orders, called the continuous and periodic review systems. Note that the decision of how much to order is directly interlinked to the frequency decision. In addition, another key factor driving the total order quantity decision concerns the stability of demand. In environments with fairly stable demand, consistent order quantities determined by the economic order quantity (EOQ) model are appropriate. In environments with unknown or unstable demand, reorder point, min-max, and order up-to policies are appropriate. Table 1 contains a summary of these inventory policies. A detailed discussion of each of these systems follows.

Periodic and Continuous Review Systems

Regardless of the underlying structural features, there are two broad classes of inventory systems determining the frequency of orders: *continuous review systems* and *periodic review systems*. A continuous review system is one in which inventory is monitored continuously. Conversely, a periodic review system is one in which inventory is checked at certain points in time separated by a fixed interval, the period. A continuous review system allows for a replenishment order to be triggered when the inventory level declines to any preset value, but this is not possible with a periodic review system. This implies that the type of inventory monitoring policy chosen determines the range of replenishment options (i.e., the order quantity decision) available.

Continuous review is clearly an information intensive policy. If a firm deals with a large number of items, it may not be feasible to review the inventory of all the items continuously. As a general rule, an item needs to be monitored continuously if a stock-out of that item is especially costly for the firm. The more critical the item, the more pressing the need to set in place a continuous review system. For example, in a wafer fabrication facility, running out of a critical part may lead to hours of costly downtime. Items of the nuts-and-bolts variety, on the other hand, may be monitored periodically; the length of the period will depend on the magnitude and variability of demand and the cost of running out of the item.

All inventory systems work by generating triggers for replenishment. The trigger is generally set off when the stock level falls below a certain level. Once the replenishment signal is triggered, the system must wait for a certain period of time before the refill arrives. This time window is the *lead time*. Lead time represents the time required to fulfill a replenishment order. It may be attributable to production time or transportation time or both.

Compared with a periodic review system, a continuous review system reduces the safety stock required to guard against unexpected surges in demand. *Safety stock* is inventory carried specifically to tide over variable demand during lead time. This is because the maximum period of time for which a continuous review system may be out of stock is simply the lead time. On the other hand, the maximum period of time for which a periodic review may be out of stock is the sum of the lead time and the period length. To illustrate this point, consider a continuous review system in which the stock level required to trigger a replenishment order is 5 or fewer units of an item. Suppose that the current stock level is 10. Now an order for 10 units of the item will trigger a replenishment signal and simultaneously plunge the stock level to zero. The time for which this system will remain out of stock is exactly the lead time. Next, consider a periodic review system for the same item in which a replenishment order is triggered at a review point whenever the stock level at the point is 5 or fewer units. Suppose that the stock level is equal to 10 units at the current review point; then no replenishment signal is triggered. But if a demand for 10 units of the item occurs immediately after the review point, it is apparent that the system will remain out of stock for a time interval equal to the sum of the period length and the lead time.

Despite the need for higher safety stock, a periodic review system may be more convenient to administer than a continuous review system. For instance, the stocks of several related items may be simultaneously checked at each review point and replenishment signals for multiple items sent out in a coordinated fashion. This is not possible with a continuous review system in which the replenishment order signals for different items are, in general, generated independently of each other.

Replenishment Rules for Known Demand: The EOQ Formula

The preceding discussion brings us to the single most famous formula in inventory management, the Economic Order Quantity (EOQ). A stable demand process permits a simple replenishment process involving a fixed, as opposed to a variable, order quantity. The EOQ formula bears out this rationale.

Consider a situation in which customer demand occurs at a steady, known rate of D units per month and a continuous review system is in place. Suppose the relevant inventory costs are an inventory holding cost of h per unit of stock per month, and a fixed cost K associated with every replenishment order. The holding cost represents the cost of physical storage, as well as the opportunity cost of capital tied up in inventory. The set up cost represents the transaction cost associated with a replenishment order. If replenishment is achieved by in-house production on a machine, K may represent the cost to set up the machine for a production run. The holding cost is in part a notional cost while the set up cost entirely reflects cash outflow and is therefore easier to quantify. Under these conditions, it may be shown using elementary calculus that the optimal policy is to order a quantity equal to

$$\sqrt{\frac{2KD}{h}}$$

units. The diagram in Figure 1 illustrates how the total costs vary with order quantity. The holding cost increases linearly while the order cost decreases nonlinearly as the order size increases. The total cost is smallest at precisely the point where order cost balances holding cost.

The trigger point for order replenishment depends on the lead time. If the lead time is practically zero, so is the trigger point; if the lead time is positive, the trigger point will also be some positive number that can be calculated by using information about the magnitude of consumer demand. In general, a replenishment is triggered when the stock on hand drops to a level equal to lead time demand.

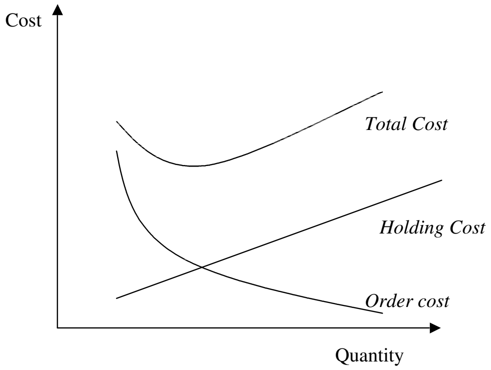


Figure 1: Total inventory costs determining economic order quantity.

Although the EOQ formula applies with complete accuracy only under idealized conditions, it is remarkably robust in the sense that the order quantity prescribed does not change appreciably even if the costs in the equation are off target. Furthermore, if the demand is variable rather than constant, a useful rule of thumb is to use the EOQ formula with average demand substituting for demand. Although the EOQ formula is not theoretically optimal when the demand process is variable, it may provide a useful benchmark. A more sophisticated procedure for handling variable demand is described in the next section. Another practical situation in which the EOQ formula does not apply directly is when the supplier offers quantity discounts. In this case, some care must be exercised in using the EOQ formula (Nahmias, 1990).

Replenishment Rules for Uncertain Demand: Reorder Point

The EOQ formula is designed to optimize an inventory system facing steady demand. If demand is expected to be variable, there are two useful statistical metrics that are used to capture the anticipated pattern of demand: *average demand* and *standard deviation of demand*. Both these metrics may be easily computed using spreadsheet functions from a database of historical demand.

Suppose we have demand for the past 12 months on file; let us denote the 12 monthly demand numbers by D_1, $D_2, \ldots, D_{12}$. Then we may use the following formulae to estimate the average and the standard deviation of demand:

$$Average = \frac{D_1 + D_2 + \cdots + D_{11} + D_{12}}{12}$$

$$Standard\ deviation = \sqrt{\frac{(D_1 - Average)^2 + (D_2 - Average)^2 + \cdots + (D_{12} - Average)^2}{12}}$$

As remarked earlier, one may substitute mean demand in the EOQ formula to obtain the order quantity for replenishment. It is advisable, on theoretical grounds, to complement information about average demand with information about demand variability, however; standard deviation is a metric that nicely captures demand variability. As we mentioned in the previous section, a simple rule of thumb for replenishment order quantity, given knowledge about both average demand and standard deviation of demand, is the EOQ with average demand substituting for demand in the formula. How does one incorporate standard deviation into the replenishment rule?

The answer lies in safety stock and the reorder point. *Safety stock* is inventory carried specifically to tide over variable demand during lead time, and the magnitude of safety stock is precisely the reorder point. The greater the anticipated variability in lead time demand, the greater the safety stock that must be carried and the greater the reorder point. A convenient rule of thumb is to set the reorder point equal to anticipated average demand during lead time *plus* a multiple (say Z) of the standard deviation of demand during lead time.

The appropriate determination of the multiplier parameter Z represents an explicit trade-off to managers between excessive inventory carrying costs and the chance of a *stock-out*. A stock-out occurs when demand is higher than current inventory levels. Another important measure of inventory performance is the *customer service level,* which is the probability that demand can be met from current inventory levels. Consider the following illustration relating the appropriate reorder point level (i.e., through choice of Z) to the overall customer service level. If demand during lead time follows a bell curve, one may use standard tables to decide on a suitable value of Z; the greater the value of Z, the greater the inventory holding cost but the smaller the risk of stock-outs. For instance, setting $Z = 0.67$ results in a 75% customer service level and a 25% chance of stock-outs, whereas setting $Z = 1.64$ results in a 95% customer service level and only a 5% chance of stock-outs. Note that the total inventory carried in the system is the sum of two distinct components: *cycle stock*, arising from the order replenishment process, and *safety stock*, carried to buffer the inventory system against demand shocks.

Replenishment Rules for Uncertain Demand: The Min–Max Replenishment Rule

One of the most popular inventory replenishment rules is the two bin or min–max policy. The rule works as follows: If the inventory level at any point is smaller than some predetermined number s, a quantity sufficient to bring the inventory level up to a level S is ordered; of course S is greater than s by design. The academic literature on inventory management generally refers to this as an (s, S) policy (Porteus, 2002). Under this rule, the specific order quantity requisitioned will vary, depending on the points in time at which successive reviews are made and on the demand conditions in the interim between two reviews.

Suppose we apply the min–max rule to an item with an inventory that is monitored continuously. If customer demand occurs in discrete units and if each order is for exactly one unit, then an order can be placed at precisely the point when the inventory level declines to s. Furthermore, the replenishment order quantity will be exactly S minus s every time a replenishment order is called for. On the other hand, under similar demand conditions, if the inventory system in place is a periodic review system, then the inventory level triggering a replenishment order will generally be smaller than s. This implies that the order quantity will be at least S minus s rather than exactly equal to it. The same observations apply to a continuous review system in which customer orders may be for multiple units of an item. It is instructive to note that the simplicity or complexity of the stock replenishment mechanism reflects the simplicity or complexity of the customer demand process. If every customer demands precisely one unit, a fixed quantity is requisitioned for replenishment under a continuous review system. If customer demand deviates from this fixed pattern, the order quantity must also vary.

Replenishment Rules for Uncertain Demand: The Order Up-To Rule

A simple rule associated with periodic review systems is to order at every review point in such a way as to bring the inventory level up to S. The order quantity resulting from this policy will, in general, vary from one review period to another; the greater the variation in demand, the greater will be the variation in order quantity. If *backorders* are allowed, the order quantity may even exceed S. Backorders occur when unsatisfied demand can be fulfilled during the following period. We introduce the notion of *inventory position* to state the "order up-to" rule in a way that covers backordered situations. The inventory position at any point in time is defined as the amount of physical stock on hand plus the amount of stock on order less the amount of backorders. So the order up-to rule is simply to restore the inventory position to S at every review point. The min–max rule, restated in a similar way, is to restore the inventory position to S provided the inventory position is smaller than s at a review point.

INVENTORY MANAGEMENT FOR MULTIPLE ITEMS AND MULTIPLE ECHELONS

We have, up to this point, discussed inventory management principles for a single item in isolation. In practice, all firms deal with several types of items, divided into groups. Inventory management rules are generally devised for groups of items stored at various places throughout the company and the supply chain. This section describes the theory and mechanics behind inventory decision making considering multiple items and multiple echelon policies. The next section offers managerial guidance concerning strategic decision making for inventory policies throughout the supply chain.

Inventory Management With Multiple Items

When managing the replenishments for several items, coordination becomes a factor. There are two conflicting dimensions to coordination in this context. When a group of items is produced on a single production line for in-house replenishment, it is desirable to trigger the replenishment orders of the items at different points in time to avoid overloading the machine. On the other hand, setup costs may be reduced, if a group of related items is produced during the same run. For instance, filling an entire truckload of items may lower transportation costs. If the items are ordered rather than produced, we may want to trigger replenishment orders for a group of items at the same time to exploit quantity discounts offered by the supplier.

When demands for different items are relatively stable over an extended period, cyclic schedules are often used to smooth the production load on a machine. This means that each item is ordered periodically, and the ordering periods of the various items are set so as to even out the production load.

Consider a situation in which a machine is used to produce four items with approximately the same demand. One feasible cyclic schedule is to produce item 1 in weeks

1, 5, 9, . . . , item 2 in weeks 2, 6, 10, . . . , item 3 in weeks 3, 7, 11, . . . , and item 4 in weeks 4, 8, 12, ...

The schedule can be implemented by using a periodic review order up-to *S* rule with a review period of 4 weeks for each of the four items. Every week, a review will occur, but in any 4 consecutive weeks, four different items will be reviewed. This inventory management policy will result in a very even production load on the machine. It may be a complicated mathematical problem to find feasible cyclic schedules for a number of items produced on a number of different production facilities; this example is a particularly simple one (Axsater, 1999).

Another important example of an inventory system in which multiple items must be managed concurrently occurs in discrete parts manufacturing, where producing one unit of a certain product may require a multiple units of different components that must be assembled to fabricate the end product. The components themselves may have to be assembled using multiple units of different subcomponents. The resulting product structure is called a *bill of material,* and the corresponding inventory planning methodology is called material requirements planning (MRP). There is one crucial distinction between the setup in MRP and that in the multiple item example discussed in the previous paragraph. An MRP system deals with a collection of items whose demands are dependent; specifically, the demands of the components are driven by the demand for the end product. On the other hand, the previous example assumed that the demands of the various items were independent of one another. In an MRP system, the stocks and replenishment policies for the components can be derived from the replenishment policy formulated for the end product, simply by staggering the replenishments of components in accordance with the lead times involved (Hopp & Spearman, 1996).

Multi-Echelon Inventory Systems

All the replenishment rules we have been discussing assume that items are produced, or procured, at a single place within a single firm. In practice, a large number of stocking points are coupled to one another and the inventory replenishment processes at the separate places must be managed in synchronism. For example, several companies use an inventory system with a central warehouse close to the production facility, and a number of dispersed stocking points serving local markets. On the production front, the site of the second of the two demand processes characterizing the inventory system, stocks of raw materials, components, and finished products are coupled to each other. To design efficient inventory control procedures for such multi-echelon systems, it is necessary to use special methods that take interdependencies between the inventories at different stocking points into account. Note that these multi-echelon inventory decisions can occur both within a single firm and throughout an entire supply chain.

One common way of controlling a multi-echelon inventory system is to apply a *reorder point, fixed quantity* replenishment rule to each facility. This rule is a variation of the min–max rule explained earlier. It operates as follows: If the inventory position at any review point is less than *R*, we place an order for *Q* units. Here *R* is the reorder point and *Q* is the order quantity. If the control system is of the continuous review type, an order for Q units is placed as soon as the inventory position declines to *R*; in a periodic review system, the inventory position will generally be smaller than *R* when a replenishment signal is sent out. Different stocking points will, in general, be governed by distinct reorder points and distinct order quantities. We call this system of multi-echelon inventory control an *installation stock reorder point policy* (Axsater, 2000).

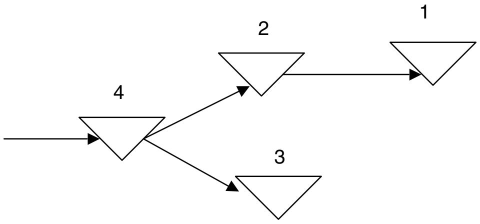


Figure 2: Echelon stock distribution example.

Another type of multi-echelon control policy is the *echelon stock reorder point policy.* Let us explain this policy with a specific example. Consider a distribution system with four installations as shown in Figure 2. The *installation stock* at each of the distribution points is simply the inventory position at that point. The *echelon stock* at each distribution point is the sum of the inventory positions of that distribution point and all the downstream points. So if the installation stocks at installations 1, 2, 3, and 4 number 5, 4, 1, and 10, respectively, then the corresponding echelon stocks work out to be 5, 9, 1, and 20. Observe that the echelon stock at each point is, by definition, greater than the installation stock at that point. An *echelon stock reorder point policy* is one that uses echelon stocks rather than installation stocks to compute reorder points.

Note that installation stocks use only local inventory information in designing the inventory control system, whereas echelon stocks process inventory information pertaining to multiple installations. An important fallout of using exclusively local information to design inventory control systems is that it tends to result in large variations in the demands on individual installations. Indeed, demand variations at the consumer level tend to get *magnified* as we traverse the supply chain from the point of external demand, downstream, all the way upstream. This effect, christened the *bullwhip effect* in the inventory literature, degrades supply chain performance (Lee, Padmanabhan, & Whang, 1997). The remedy is to formulate and employ echelon level inventory control policies that use global, rather than merely local, information.

E-COMMERCE INITIATIVES AFFECTING INVENTORY MANAGEMENT

Although the advent of the Internet has had an impact on inventory management in several ways, the traditional inventory techniques previously discussed still yield many important managerial insights. In particular, many of the parameters, variables, and costs previously discussed may be altered, but the same basic relationships still hold. In this section, we briefly discuss several of the newer e-commerce initiatives relating them directly to the traditional inventory management techniques previously discussed. In the following subsections, we describe many of these Internet-related initiatives in greater detail.

First, the availability of Internet procurement applications has altered several of the key variables traditionally used in inventory management. According to a recent report (Aberdeen Group, 1999), early adopters of Internet procurement initiatives experienced the following: (a) a 5–10% reduction in the price of materials and services, (b) a reduction in the average length of the purchase and fulfillment cycles (i.e., *lead time*) from 7.3 days to 2 days, and (c) a reduction in the administrative costs per order (i.e. *order cost K*) from $107 to $30. Consequently, these firms also decrease their overall levels of inventory, including *safety stock* and *order quantities* to suppliers.

Use of ERP and Web-based technologies has enabled more accurate prediction of demand, thereby facilitating the calculation of appropriate *reorder point* policies. Third, the advent of supply chain management techniques has emphasized the importance of analyzing the inventory position of each individual supplier in determining appropriate *echelon stock policies* throughout the supply chain. Many firms are conceding control of their inventory decisions to supply chain partners with the hopes of achieving better coordination and higher customer service levels throughout the supply chain. Moreover, functionalities offered in modern day ERP and Web services software facilitate sharing of inventory information throughout the supply chain, thereby minimizing *bullwhip effect*.

Collaborative Planning Forecasting and Replenishment

To use many of the inventory management techniques previously discussed, a firm must develop a reasonably accurate picture of demand. Companies usually rely on a variety of both subjective and objective techniques to calculate forecasts for potential product sales, depending on the types of products under consideration. For example, when planning the sales for an item with fairly stable demand, mathematical analysis of the historical sales for this product should be adequate to characterize previous demand patterns and predict future demand. For items that are completely new in the marketplace or trendy in nature (e.g., apparel and toys), subjective methods can be used such as polling experts in the industry to project the potential acceptance in the market. See Nahmias (2001) for a complete review of the topic of forecasting.

Recently, companies have recognized the value of developing a common forecast to share with supply chain partners. Collaborative planning forecasting and replenishment (CPFR) initiatives were developed to help facilitate such information sharing (see http://www.cpfr.org). Collaboration within the chain is executed through confidentiality agreements, dispute management protocols, shared performance metrics and common goal sharing. Additionally, the scope of the collaboration is qualified in terms of product (e.g., SKU, family, category, etc.), location (e.g., site, geography, etc.), and trading partner participation. The planning component implies collaborative product planning for the product level in scope as well as the target operations measures (e.g., sales volume, margins, fill rates, pricing, inventory levels, safety stocks, etc.). Collaborative planning also applies to business planning coordination of promotions, new product introductions, product phase-outs, inventory policy changes, and so forth.

The third component, forecasting or collaborative forecasting, seeks to eliminate the independent demand decisions that cause information distortion up the supply chain (i.e., the bullwhip effect). Trading partners agree to the final working forecast and jointly resolve forecast variance. Once a sales or a downstream customer forecast is reviewed, the forecast is then converted into an operations forecast using a demand planning tool commonly included with ERP systems. Finally, an appropriate inventory replenishment methodology can be adopted based on characteristics of the product, life cycle, and demand. Note that replenishment also extends into the efficient delivery of products through transportation. CPFR not only improves replenishment among trading partners but links demand and supply planning into a continuous process.

Historically, CPFR was initiated and used by retailers and consumer products companies such as Walmart, Safeway, Proctor & Gamble, Kimberly Clark, and Nabisco. More recently, however, CPFR case studies have extended beyond retail to other industries that require efficient chains such as Hewlett Packard. All of these companies have sought benefits such as greater product or category sales, greater customer service through reduced stockouts, reduced inventory held within the supply chain, compression of lead times among trading partners and increased forecast accuracy. Additionally, the adoption rate of CPFR has increased through the use of technologies such as enterprise resource planning software, EDI, Internet and Web applications, and radio frequency identification (RFID) scanning tools. However, drawbacks to CPFR exist, and include trust issues (i.e. many companies are reluctant to share the level of data necessary due to a potential loss of competitive information), technology issues (i.e., many companies don't have the technological capabilities for this type of sharing) and high implementation costs.

Virtual Inventory Management

In an effort to increase supply chain efficiencies, responsiveness and customer service, many firms engage in contracts to manage a downstream partners inventory. We term this policy virtual inventory management, and such contracts may occur between any echelon of a supply chain network. For example a Hewlett Packard (HP) network may include a contract manufacturer, an assembly manufacturer, an HP hub distributor, HP (the original equipment manufacturer [OEM]), a channel distributor, and a retail channel. A virtual inventory management scenario may occur between any direct partner within such a supply chain network. Contracts between partners may be based on target service levels, bonded products and quantities, fill rates, turns, and so forth. Additionally, the upstream supplier may own the inventory in the downstream partners warehouse (consigned inventory). In the vendor managed inventory (VMI) scenario, the supplier has the additional responsibility of monitoring and triggering the replenishment request. Many of the advances in virtual inventory management have been enabled by information technologies such as electronic data interchange, e-mail with attachments (such as the ubiquitous Excel), enterprise resource planning systems, and Web portals.

There are many benefits of such inventory collaboration within the chain. Chief among these is the leveling of purchase demand and related order quantities throughout the supply chain (i.e., eliminating or reducing the bullwhip effect). Consequently, all members of the supply chain should incur increased service levels, reduced stockouts, and lower total inventory costs. Furthermore, increased visibility of demand throughout the supply chain enables better coordination of marketing events with production planning, thereby enabling the right product at the right place in the right location in the right quantity at the right time. Nonetheless, collaboration pitfalls proliferate, challenging the best efforts of firms. Crucial managerial factors to consider are the level of trust between the trading partners and the potential for losing proprietary information. Technological pitfalls include issues with EDI (content and accuracy) and new software, response to demand spikes and information distortion, inventory over stock, information disparity (supplier system says *X* quantity whereas the downstream partner says *Y*), product obsolescence, human error (downstream partner pulled the wrong part numbers), and learning curve effects of new technologies.

Virtual Inventory Fulfillment

In addition to virtual inventory management techniques, several companies have taken the notion one step further in reducing their exposure to inventory altogether via a practice of virtual inventory fulfillment. The policy of direct (or drop) shipment is when a firm takes an order from the customer, transmits this information directly to a supplier who creates the item with the original firm's logo, and ships it directly to the customer. In effect, the firm avoids carrying any inventory whatsoever by outsourcing both the manufacture and shipment of goods to a supplier. So, in determining an appropriate way to fulfill customer demand, a firm can select either a traditional structure (i.e., brick and mortar firm that maintains inventory) or a virtual structure (i.e., drop shipments).

Randall, Netessine, and Rudi (2002) highlighted the fact that product demand patterns and product attributes influence the structure choice (see Figure 3). This dichotomy, traditional versus virtual structure, represents a trade-off between inventory costs (inventory and fulfillment) and high customer service levels. The benefits of a virtual model can be alluring. For example, companies such as Spun.com have used this strategy to reduce their investment in inventory and fulfillment infrastructure and to access wholesalers (such as Alliance Entertainment) that engage in inventory pooling strategies to leverage economies of scale. Other benefits of the virtual fulfillment strategy can include better access to a wider product selection, more predictable product availability, and lower transportation costs.

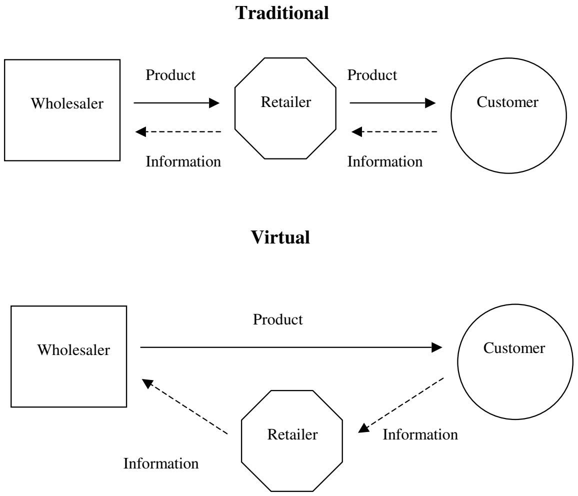


Figure 3: The two fulfillment models. Source: Randall, Netessine, and Rudi (2002). © 2003; reprinted with permission from Reed Business Information.

The pitfalls of a virtual model are numerous, however. Chief among these is the firm's loss of control over both the product quality and service levels. In addition, the strategic drawbacks are numerous because the bargaining power of the supplier is increased and the threat of forward integration is eminent because of increased knowledge of the consumer preferences. This phenomena whereby one of the members of the supply chain is essentially eliminated is widely known as the process of supply chain disintermediation. In this case, the supplier has the potential to sell directly to the customers. Interestingly enough, several prominent e-commerce firms have adopted traditional structures (i.e., similar to a brick and mortar firm that maintains inventory). To illustrate, consider Amazon.com, which has adopted a successful traditional fulfillment model with well-developed internal customer fulfillment capabilities, thereby avoiding the problems associated with a virtual structure and creating higher barriers to entry for potential competitors.

TECHNOLOGIES ENABLING INVENTORY MANAGEMENT

Enterprise Resource Planning Systems

Enterprise Software Systems, otherwise known as Enterprise Resource Planning Systems (ERPs), such as Oracle, PeopleSoft, and SAP) are a collection of application components designed to enable end-to-end business processes. Business flows enabled by ERP systems automate those business processes associated with the management of operations, customers, suppliers, partners, and employees. The technical goal of an enterprise system is to enable the disparate process flows of a firm to work within a single global database. This technical achievement provides the foundation for complete and consolidated business process information for sales, inventory, and revenue across geographically dispersed lines of business and products. A firm's inventory decision making is facilitated because of the information visibility and timeliness enabled by a common database. Additionally, enterprise systems reduce costs for firms by integrating and managing the applications for historically and organizationally disparate business processes (e.g., financials, operations, order fulfillment, etc.).

Despite the promise enterprise systems can deliver, ERP systems are not without their shortcomings. Most commonly, firms misjudge the complexity of implementing a system that models their complete enterprise. Such underestimations have led to supply chain glitches impacting the timeliness of product delivery that have torpedoed a firm's market value. For example, Hershey was unable to meet candy demand during crucial holiday seasons because of ERP implementation problems (Singhal & Hendricks, 2002).

ERP systems continue to evolve, however, and have become ubiquitous within organizations. Additionally, ERP vendors have consolidated and continue to expand their functional footprint and industry expertise. IDC (International Data Corporation) predicts that the ERP market will exceed $27 billion in 2005 (Byron, 2003).

Web Services

Cross enterprise collaboration is the greatest challenge that firms face when opting to share information between their supply chain partners. Web service solutions such as information sharing portals provide fast and low-cost alternatives for connecting partners. The costs associated with time and coordination for developing standards for information integration can be substantial. Web services technologies built on vendor neutral open standards such as TCP/IP, XML, SOAP, WSDL, and UDDI reduce or eliminate the need to develop standardized common interfaces between a partner's applications.

ERP application vendors have taken advantage of open interface standards and now provide out-of-the-box tools and application solutions that permit a firm to share information with its upstream suppliers and its downstream customers through Web portals. From the upstream side, a firm's supplier can quickly have real-time visibility into purchase orders, payment history, forecasts, performance (e.g., quality and on-time delivery), capacity management, and shipment notification (e.g., advanced ship notices, ship schedules). From the downstream side, a firm's customers can place orders, get availability dates, track shipments, review invoices and payment information, request service, return products, submit forecast, and perform myriad other customer care initiatives. Web services offer a low-cost and intuitive mechanism to share information among supply chain partners and faster integration with a firm's enterprise applications.

CONCLUSION

The advent and accessibility of Internet technologies has heightened the need for effective inventory management techniques. With the increased transparency of lead time and delivery information, customers are demanding correspondingly higher levels of service. The problems associated with poor inventory management are clear: stock-outs, product delays, loss of sales and revenues, and even a decrease in the market value of the firm experiencing the inventory related problems. Consider the case of eToys, an online retailer of toys for children. Because of the increased use of e-commerce methods as an effective way to access an increased pool of customers, eToys experienced unusually high demand during the 1999 holiday selling season. Unfortunately, eToys was unable to deliver the high volume of goods because of ineffective and costly customer fulfillment processes, and the firm eventually went bankrupt.

Existing traditional models of inventory management provide managers with a simple toolbox to plan for such situations. A detailed discussion that outlines these important decisions regarding effective inventory management is contained in the second section, Inventory Replenishment Methodologies. The first decision concerns the timing or frequency of appropriate orders to maintain an adequate supply of goods. In a continuous review system (used for costly or high-profile items), the inventory level is monitored continuously. More commonly, a period review system is used to check the inventory levels at fixed time intervals.

The second key inventory decision is the appropriate quantity of goods to order when monitoring the inventory levels. The economic order quantity (EOQ) model is one of the most commonly used inventory techniques suitable for environments with stable demand. Given estimates for demand, holding costs, and ordering costs, appropriate order quantities can be easily calculated. For environments with considerable demand uncertainty, other models take into account the cost of overstocking (i.e., having too much inventory left over after a selling season) and the cost of understocking (i.e., having too little inventory available to meet customer demand.) More specifically, the reorder point model uses information concerning both the mean and variance of demand to calculate an appropriate order quantity that ensures that customer demand can be met from current inventory levels. Although there are many variations of these two models (such as min–max and order up-to policies), the EOQ and reorder point models provide a convenient starting place for managers to think about the relative importance of several factors driving a decision concerning appropriate inventory levels.

Managers should also consider coordinating orders for multiple items to facilitate effective inventory management. For example, when ordering similar items from the same supplier, ordering them at the same time may offer the benefits of quantity discounts or lower transportation costs. Another situation warranting coordination is when the components from the bill of material (BOM) for a particular product all need to be ordered such that they arrive at the same place for assembly or fabrication at the same time. In this case, a material requirements planning system can easily determine the appropriate order quantities for individual components and stagger the replenishment according to the established lead time for the individual components.

Multi-echelon inventory systems coordinate inventory decisions across a large number of stocking points located throughout a firm or supply chain. One commonly used multi-echelon inventory system is called an installation stock reorder point policy, which essentially applies a variation of the basic reorder point model to each individual facility. Alternatively, an echelon stock reorder point policy takes into account inventory levels at all facilities downstream of a particular one to coordinate inventory levels throughout the system. A key advantage of this policy is that it minimizes the variations in orders (i.e., the bullwhip effect) that can occur between stocking points within a firm or throughout a supply chain.

More recent trends in inventory management associated with supply chain management and e-commerce initiatives are discussed in the section E-Commerce Initiatives Affecting Inventory Management. Although each of these techniques (CPFR, virtual inventory management, and drop shipping) provides certain benefits associated with minimal inventory levels, they must be carefully weighed against potential downfalls. Collaborative planning forecasting and replenishment is the technique of joint forecasting and planning for inventory management among supply chain partners. Use of CPFR techniques can yield benefits such as greater product or category sales, greater customer service through reduced stock-outs, reduced inventory held within the supply chain, compression of lead times among trading partners and increased forecast accuracy. Drawbacks to CPFR include trust issues (i.e., many companies are reluctant to share the level of data necessary due to a potential loss of competitive information), technology issues (i.e., many companies don't have the technological capabilities for this type of sharing) and high implementation costs.

Virtual Inventory Management is a common practice whereby retailers and manufacturers share their forecasted demand data with suppliers. Consequently, suppliers can monitor and maintain appropriate inventory levels for their customers. The benefits of Virtual Inventory Management include minimization of demand distortion and visibility of demand, thereby decreasing bullwhip effect and enabling better customer service. Pitfalls of such arrangements include a loss of control, access of potentially proprietary information, and technological problems. Virtual inventory fulfillment via drop shipping, another recent technique used primarily by Internet-based companies, occurs when the retailer delivers orders straight through to a wholesaler or manufacturer. In turn, the wholesaler or manufacturer ships the products directly to the customer such that the retailer holds no inventory at all. Although the retailer reaps the savings from carrying no inventory, it also has no control over the customer service and quality of goods. Moreover, the threat of the supplier sidestepping the original retailer and direct selling to the customer is a potential risk of this strategy.

Finally, several technologies are available which enable many of the inventory policies previously discussed. Enterprise resource planning systems are software that enables business process flows via a common database. Although such systems can be difficult to install, they can facilitate the planning, ordering, and monitoring of inventory levels for all of a firm's products. Whereas ERP systems are mostly used in the context of a single firm, Web service solutions such as information-sharing portals are becoming viable alternatives for connecting a firm with appropriate supply chain partners. These solutions enable real-time visibility of purchase orders, inventory levels, and supplier performance. Essentially, effective utilization of these new technologies including ERP systems and Web services can help minimize some of the ever-present risk involved with managing inventory.

GLOSSARY

Collaborative planning forecasting and replenishment (CPFR) An initiative enabling online sharing of data to establish common forecasts and production schedules among supply chain partners.

Continuous review An inventory control system whereby inventory is checked continuously.

Customer service level The probability that the inventory available during lead time will meet demand.

Direct (or drop) shipment A form of fulfillment in which a firm takes an order from the customers and then transmits this information directly to a supplier who creates the item with the original firm's logo and ships it directly to the customer.

Economic order quantity (EOQ) An inventory model appropriate for environments with stable demand that determines the optimal order size that minimizes the sum of carrying costs and ordering costs.

Enterprise resource planning systems (ERP) A collection of software application components designed to enable end-to-end business processes.

Holding (carrying) costs The costs of holding an item in inventory. These costs vary with the level of inventory.

Inventory A stock of items kept by an organization to meet internal or external customer demand.

Inventory management Decisions concerning the amount and frequency of inventory to hold and order.

Lead time The period of time between order of goods and delivery.

Min–max system A replenishment rule defined by two numbers *S* (Max) and *s* (Min). If the inventory level at a review point is *s* or less, an order is triggered for a quantity sufficient to pull the inventory level up to *S*.

Multi-echelon inventory system A supply chain with inventory stocked at multiple installations, possibly owned by distinct firms.

Order costs The costs of replenishing inventory that are independent of the order size.

Order up-to rule A replenishment rule characterized by a single number *S*; at any review point, an order sufficient to raise inventory level to *S* is triggered.

Periodic review An inventory control system whereby inventory is checked at regular intervals, not continuously.

Reorder point A predetermined inventory level below which a replenishment order is triggered.

Safety stock Inventory carried specifically to tide over variable demand during lead time; it takes into account standard deviation of demand during lead time rather than merely average demand

Underage (shortage) costs The loss of sales when demand cannot be met.

Vendor managed inventory (VMI) An agreement whereby the manufacturer is responsible for maintaining the supplier's inventory levels; the manufacturer has access to the suppliers inventory data and is responsible for generating purchase orders.

CROSS REFERENCES

See *Enterprise Resource Planning; Supply Chain Management; Web Services.*

REFERENCES

Aberdeen Group (1999, July). Strategic procurement: The next wave of procurement automation. White Paper. Boston: Aberdeen Group.

Axsater, S. (2000). *Inventory control.* Boston: Kluwer Academic Publishers.

Byron, D. (2003, March). Preliminary Worldwide ERP Market Forecast, 2003–2007 and 2002 "Top 10" ERP Players. Framingham: IDC.

Fisher, M. (1998). What is the right supply chain for your product? *Harvard Business Review, 75,*105–116.

Hallowell, R. (2000, August). Service and value in eCommerce. *Harvard Business Review Case,* 1–15.

Hopp, W., & Spearman, H. (1996). *Factory physics.* Boston: Irwin/McGraw Hill.

Johnson, E. (1998). Giving 'em what they want. *Management Review, 87,* 62–67.

Lee, C. (2002). Demand chain optimization: Pitfalls and key principles. Supply Chain Management Seminar White Paper. San Fransisco: Nonstop Solutions.

Lee, H. (2002). Aligning supply chain strategies with product uncertainties. *California Management Review, 44,* 105–119.

Lee, H., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43,* 546–558.

Nahmias, S. (2001). *Production and operations analysis.* Boston: Irwin McGraw-Hill.

Porteus, E. (2002). *Foundations of stochastic inventory theory.* Stanford, CA: Stanford Business Books.

Singhal, V., & Hendricks, K., (2002, January/February). How supply chain glitches torpedo shareholder value. *Supply Chain Management Review,* 18–24.

Randall, T., Netessine, S., & Rudi, N. (2002, November/December). Should you take the virtual fulfillment path? *Supply Chain Management Review,* 54–58.

Homepage of the Collaborative Planning, Forecasting and Replenishment (CPFR) concept. Retrieved April 1, 2003, from www.cpfr.org

J

Java

Judith C. Simon, *The University of Memphis*
Charles J. Campbell, *The University of Memphis*

INTRODUCTION

Java is a programming language that was developed at Sun Microsystems in 1991. According to the Java Web site (http://java.sun.com), it was intended originally to be used for embedded programs in consumer devices and computers. Possibilities were expected to include various household products such as entertainment devices, appliances, and security systems. Although the expected main target, digital cable television, did not turn out to be a viable business, the Internet was becoming very popular and proved to be a feasible alternative. By May 1995, Java technology had been released for use on the Internet, and it was being incorporated into the Netscape Navigator browser. Listed below are some of the major milestones shown on the java.sun.com "time-line" Web page (http://java.sun.com):

May 1995	Official launch of Java technology
September 1996	83,000 Web pages were using Java technology
February 1997	JDK 1.1 released (JDK = Java Development Kit)
January 1998	Two-million downloads of the JDK had occurred
May 2000	Five-year anniversary; 400 Java Users Groups had been established around the world

BUSINESS USES OF JAVA

Java has become an important programming language used by businesses, especially for its features that work well for Internet applications. According to the java.sun.com site, over two-million developers have "embraced" the Java platform.

The Carnegie Mellon Software Engineering Institute's (SEI) Software Technology Review (Software Technology Review, 2001) describes Java as an object-oriented programming language that does a good job of distributing software over a network by addressing such issues as security and portability. Early Java programs were focused primarily on client-side applications, such as applets embedded in Web pages and viewed by Internet users. In contrast, Java 2 Enterprise Edition (J2EE) is a newer release with options expanded to include server-side applications, e.g., the code that a server uses to generate Web pages.

An *InformationWeek* article (Greenemeier, 2002) indicated that Java is expected to continue to be important for many information technology (IT) developers. The article reported on a study of 14,000 IT developers, which looked at the ways companies are deploying Java-based applications. Results indicated that 69% of the respondents were using Java-based applications for database management and access, and 66% were using it for Web-based data entry and retrieval.

Examples of business uses of the Java programming language continue to increase. Many are due to the features described above that make Java a good fit with many Internet Web-based activities.

The java.sun.com site contains several examples of business applications, in a range of industries. One example involves a fitness company that decided to change to Java for development of its enterprise-level Web applications, because they did not want to continue to invest in the maintenance of their proprietary software and liked the reusability of the Java technology. Another example involves the gaming industry, where it was decided to use Java to write games once that could then be deployed across many platforms, making the new games accessible worldwide very quickly.

An article in *Software Magazine* (Harding, 2002) described DaimlerChrysler's move toward Web-based applications in their deployment of client-side Java applications to a large number of users. The article reported that DaimlerChrysler had deployed six different applications so far, which are accessed by approximately 4,000 users.

A large consulting organization decided to use Java for some of its business applications and indicated the following as the primary reasons (Dennis and Rubin, 1999):

ability to support platform-independent applications,

dynamically downloaded and installed applications requiring no user assistance,

secure application download through networks such as the Internet, and

improved application quality over C and C++ development through a simpler programming language and improved error handling.

Some of the benefits identified include

use of platform independence to reduce computing platform costs by 50%,

application development productivity gains of 10 to 20%, and

approximately 75% fewer application errors in newly released production applications.

EXPECTED TRENDS RELATED TO JAVA

The continuing increase in business uses of the Internet is expected to cause a corresponding increase in the use of Java for business applications. Java technology continues to expand, which will aid in these activities. Articles at the Sun developer site (http://developer.java.sun.com) provide descriptions of the numerous options available and an idea of the trends. For example, there is now a Java Message Service for accessing enterprise messaging systems, Java APIs for various new applications, including Remote Procedure Call, messages, document processing, a wireless toolkit, a Mobile Information Device Profile, Personal Java for consumer devices with a touch sensitive screen, and Java Card Technology.

One of the interesting trends is that Java is moving into areas that were part of its initial intended market, including consumer devices such as next-generation telephones, TV set-top boxes, and smart cards that fit in your wallet. In the case of the Java Card platform, applications in the form of bytecode are loaded into the memory zone of the smart card's microprocessor, where they are run by the Java Virtual Machine (JVM). The executable code is platform independent so that any card incorporating a Java Card technology-based interpreter can run the same application (see http://java.sun.com/java2/whatis/).

It is impossible to predict what new technologies await us that might be enhanced through the use of Java technology.

OVERVIEW OF JAVA FEATURES

Included in this section are further explanations of some of the general features of the Java programming language, including its object-oriented design, its portability, its use of multithreading, and its method of handling errors. A more detailed identification of Java features and options is provided in the next section, followed by sample program code to illustrate the concepts that are discussed.

Object-Oriented Design

The object-oriented design concept involves the use of classes and objects to model the characteristics of an abstract or real object. A *class* represents a general identification of an item of interest. A class describes details common to all members of that class. An *object* is an instance of a class. A software object in an object-oriented program is a representation of a real-world object. It has some characteristics of interest to the situation. When writing an object-oriented program, the focus is on objects and their components. By way of contrast, programming languages such as C are considered "procedural" and focus on writing functions, i.e., the actions or procedures desired, rather than being based on objects.

Inheritance is an object-oriented design feature that allows a class to automatically contain the variables and methods that have been defined in its superclass. The superclass is the class from which another class (a subclass) is derived. This capability provides the desirable feature of reusability, since a subclass will automatically have the variables and methods of the superclass. Another feature about inheritance is that subclasses are not limited to what is inherited. Additional attributes and methods can be added as needed. Inherited methods can be overridden by new methods when appropriate. Multiple layers of inheritance are allowed. Therefore, a subclass can serve as a superclass to a lower-level subclass.

Encapsulation is another term associated with the concept of object-oriented design. Encapsulation refers to the feature of grouping or "bundling" knowledge about an object into a module. A program can then continue to use this object without having to be concerned about changes that might occur related to its variables and methods.

Because Java is totally object oriented, everything in a program is performed by invoking a method. (See Java programming example code for further information, as well as specific books with more extensive examples and explanations.)

Applications determined to be appropriate for object-oriented design can be written in conventional languages such as C but are normally considered easier to write in a language designed specifically for object-oriented programming. Languages such as C++ are sometimes referred to as "hybrids," because they are based on a non-object-oriented language (such as C) but have been enhanced with concepts of object-oriented design. Java is considered one of the "pure" object-oriented languages, i.e., based entirely on object-oriented design principles.

Portability

Portability refers to the capability of running a program on many different types of hardware platforms. When using many other programming languages, it is necessary for a developer to write the program for a specific destination hardware platform where the program is to be used. However, the Java platform allows an application written in Java to be used on many different kinds of computers. Java programs can run on different operating systems, allowing them to be used on a PC, a Macintosh computer, or across networks. This capability is sometimes referred to as "platform independence."

Multithreading

A "thread" in programming refers to one sequence of executable statements; i.e., all the statements are executed or performed in order, one after the other. Multithreading refers to the capability of subdividing a program so that different threads of the program code can be executed concurrently; i.e., the central processing unit (CPU) can alternate very quickly among the threads. Java is a multithreaded language. C and C++ are considered to be single-threaded languages.

The Java Virtual Machine (JVM) described in one of the following sections is an example of the use of multithreading. It contains threads that are needed for the programs to be executed properly. For example, the "garbage collector" thread automatically reclaims allocated memory no longer needed. In other programming languages, the programmer must handle memory management procedures as part of the program. Therefore, if something goes wrong with the program, memory availability could become a problem that might affect the successful completion of the program.

Error Handling

Error handling refers to ways of taking care of unexpected events during program execution. Error handling in Java is provided through "exceptions." Exceptions are events that occur during program execution that cause a disruption in the typical flow of instructions. In contrast to many other programming languages, Java's error handling methodology allows the error-handling code to be separated from the main part of the program code, making it easier to read and modify programs.

The purpose of handling exceptions is to try to catch and handle program errors or other unusual occurrences rather than just waiting to see if they occur and then trying to handle the results. In general, Java code should be (and sometimes must be) written to "throw" an exception when an error is encountered. The exception "handler" is a complementary piece of code that will "catch" the exception and handle it (assuming that it is something for which error-handling code has been written).

A MORE IN-DEPTH LOOK AT JAVA'S FEATURES

Some specific features of Java that further distinguish it from other programming languages are described briefly in this section. The use of applets is discussed, followed by descriptions of the platforms currently available. Overviews of some of the key components that make Java work are provided as well. These paragraphs are designed to be introductory in nature. The java.sun.com site (as well as numerous others) should be accessed if a more in-depth coverage of these features is needed.

Availability of Java Platforms

Most people using a computer today already have access to a Java platform. Web browsers typically include the Java platform, and most desktop computers also provide the capability of executing standalone Java code that does not require the use of a browser. Anyone who needs the Java platform can download it directly from the java.sun.com site. The entire Java Development Kit (JDK) (referred to as SDK—Software Development Kit—in newer versions) can be downloaded from Sun, typically at no expense. Some restrictions related to incorporating the Java platform into commercial products may exist. Additional software to support developers in writing Java programs is available for purchase from separate vendors.

Applet Development

An applet is a Java program that is referred to in an HTML document in the same way that an image file is specified as part of the page to be displayed. The applet is executed from within the Web browser when the HTML file is used. The popularity of applets became the first large use of Java programs. They are used often to provide executable content, such as pop-up windows that are displayed when some specified event occurs. For example, a small window with additional information might be displayed when a Web user clicks on a particular location on the page. Applets are used within Web pages, in contrast to standalone applications written in Java now widely used in businesses and executed independently of a Web browser.

The number and variety of applets available has continued to expand at a very fast pace. Many of these applets can be downloaded free to individual Web sites. Any Internet search engine can be used to locate applets. In addition, the java.sun.com site provides links to a variety of applet resources.

Specific Java Platforms

The current version of Java is generally referred to as Java 2, and several platforms are available. Each platform contains various tools for developers. The paragraphs below provide brief overviews of the platforms; extensive details are available at the java.sun.com Web site.

Java 2 Platform, Standard Edition (J2SE), is generally intended for building and deploying client-side applications in an organization. For example, it is widely used for developing corporate intranets to be used by employees of an organization, as well as for Web sites to be used by customers/clients of the organization. According to the java.sun.com Web site, J2SE provides the compiler, tools, and Application Programming Interfaces (APIs) needed for using the Java programming language for developing, deploying, and running applets and applications. J2SE also includes the facilities needed for working with XML documents. (XML is a technology expanding significantly in its use in the area of data interchange over the Internet, so the ability to use XML and Java together makes it easier and faster to create more sophisticated Web applications.)

Java 2 Platform, Enterprise Edition (J2EE), is designed to manage the infrastructure and support Web services for development of interoperable business applications. According to the java.sun.com Web site, J2EE takes advantage of many features of J2SE, such as its portability and use with XML. J2SE was used as the base from which other components were added, e.g., those related to server support.

Java 2 Platform, Micro Edition (J2ME) is intended to be used for small consumer products, such as smart cards, pagers, set-top boxes, and embedded devices. This type of use is not entirely new, as it is somewhat of a return to the original expectation of using Java for consumer devices. J2ME, however, is much more extensive and is intended as a technology solution for manufacturers of a variety of small ("micro") devices.

Java Virtual Machine (JVM)

The Java Virtual Machine (JVM) is defined as a computing device that contains a set of instructions for decoding and executing "bytecode," as well as other components needed in the background, such as areas for storing data and methods. The bytecode is a key part of this process. It is machine-independent code generated by the Java compiler when a program is developed. The program is later executed on the destination computer by the interpreter. This process provides the platform independence and portability that is a major feature of this programming language. The JVM is a part of the Java Runtime Environment.

Java Runtime Environment (JRE)

The Java Runtime Environment (JRE) refers to the subset of the JDK/ SDK that can be distributed to computer users. It contains the JVM, a bytecode verifier to check the code before execution, the bytecode interpreter for execution of the program, and Java class libraries. Class libraries are existing, previously developed classes available to programmers, described further in the section below on Java Application Programming Interface. The JRE is available from the Sun site. For example, the Java 2 Runtime Environment Standard Edition is available from Sun's J2SE homepage.

Java Application Programming Interface (API)

The Java programming language contains numerous predefined classes that have been grouped into related classes called "packages." Together, these packages are often referred to as Java Application Programming Interface (API) packages. Described briefly below are some well-known and widely used packages, many since the first release of Java. Details about these and other classes are available on many Web sites.

java.applet—Enables applets to be created and to interact with a browser.

java.awt—Refers to the Abstract Window Toolkit, which includes the classes and interfaces needed for creating graphical user interfaces (GUIs) and which has been expanded in Java 2 through the Swing component discussed below.

java.beans—Allows the development of reusable software components that can interact with both Java and non-Java components.

java.io—Provides the capability for programs to input and output data.

java.lang—Automatically imported into all programs by the compiler because it has basic classes and interfaces needed by many Java programs.

java.text—Enables a Java program to work with numbers, dates, strings, and characters.

java.util—Provides a variety of additional capabilities, such as random number processing, resizing of array-like data structures, and bit manipulation.

Java Foundation Classes (JFC) refers to API software that is available as part of the J2SE SDK. It includes the original Abstract Window Toolkit (AWT) but has extended it by adding other features, especially an extensive set of GUI class libraries. In addition to the AWT, some of the features of the JFC software are:

Swing—A GUI component kit that provides user interface elements allowing a customizable look and feel; these elements have been written entirely in Java, which has simplified the deployment of Java applications without reliance on any specific windowing system.

Pluggable Look and Feel—Gives users the capability to change the look and feel of an application without the need to restart it.

Accessibility API—Provides assistive technologies that include screen readers, screen magnifiers, and speech recognition so that applications developed using the JFC software will help developers meet federal regulations regarding accessibility.

Java 2D API—Extends the AWT graphics and imaging classes by allowing developers to incorporate two-dimensional graphics, text, and images in Java applets and applications.

JavaBeans

JavaBeans is a component architecture that is platform neutral, which means that it uses the Java capability of "Write Once, Run Anywhere" for the development of reusable components. This capability allows developers to create and then use these components in combination with an organization's existing applications, regardless of the operating system or application environment involved, whether it is within the organization or across the Internet.

Java Web Start

Java Web Start is available for download as part of the J2SE. It provides a technology for simplifying the deployment of Java applications. Applications can be launched from a Web browser by clicking on a link on the Web page. For example, a spreadsheet program could be launched without going through installation procedures. If the computer does not already have the needed application, Java Web Start will download the needed files automatically and cache them on the computer so that the application can be run again at any time. Java Web Start runs independently of a Web browser, so a computer user can continue using an application whether or not the Web page is still in use. The application itself can be relaunched from a desktop icon without returning to a browser page.

A SAMPLE JAVA PROGRAM

Program Description

A Java program for a simple banking activity is described and explained below. This banking program is designed to allow the computer user to enter customer information such as an identification number, the customer's initial balance, and the annual interest rate paid for this balance. The program will allow the user to maintain an up-to-date balance while entering deposits or withdrawals. In addition, the program will determine and display the annual balance until the original balance has doubled. (For purposes of this programming demonstration, it is assumed that the customer wants to know how long it will take for the original balance to double in value.)

In this object-oriented program, one object of interest is the bank's customers. An individual customer object is an instance or occurrence of the customer class. A program related to customers would determine the particular items needed about the customers. In the banking example with a customer object, specific items of data (variables) and methods (behaviors) would be identified.

A complete Java program related to the banking example is presented below, shown in small segments preceded by explanations. This program develops an applet that could be used for this banking example. When an applet program file is developed, at least one other file is required. Since an applet must operate within a browser, a separate HTML file that calls (refers to) the applet file and causes it to be displayed on the screen is needed. In this example, another file is needed for the complete application, in the interest of good programming practice, to use multiple small files and separate the different activities into their own programs.

This additional file is a separate Java program that is not an applet. It is a class definition file designed to perform the actual calculations of the bank balance. (As suggested above, this calculation could have been performed within the applet file.) This separate program is a public class named Customer. It contains variables associated with each of the objects, constructors ("pseudo-methods") used to initialize the objects, and the methods common to the objects, such as "calc_deposit" and "calc_withdraw," that will be seen in some of the code.

For the purposes of this presentation of code, the second Java program and the HTML file are not displayed here but would be needed for a complete application of this banking example.

Program Content and Explanation

The segment of code shown below brings in (imports) existing classes so that this applet program can be developed without the need for the programmer to develop some aspects needed in all similar programs. (A standalone program, i.e., not an applet, would have some differences, such as no need for the java.applet class, but this particular program is intended to create an applet.) These statements are telling the Java compiler to give the program access to the specified class libraries. Some formatting features required in Java shown in the code in this section include the use of a semicolon at the end of each statement and the use of braces to enclose each class definition:

```
import java.applet.*;
import java.awt.*;
import java.awt.event.*;
import java.text.NumberFormat;
```

The next statement declares that a subclass named CustUser is being created. It will extend java.applet.Applet (that is, it will inherit all the characteristics and capabilities available from the Applet class), including being an applet itself, and will implement the methods available in java.awt.event.ActionListener, which is a class in the java.awt.event library. This "interface" allows the program to perform an action when a button is clicked or the Enter key is depressed on a specified location on the screen. The "public class" portion of the statement indicates that CustUser grants public access and that it is a class. Classes may be declared to be "private," "protected," or "public." If a class is declared to be "public," it is accessible, along with its public attributes and methods, to any other class. If it is "private," it cannot be accessed directly from another class. If it is "protected," it may be accessed directly by classes in the same package but cannot be accessed from other classes.

```
public class CustUser extends Applet
  implements ActionListener
{
```

The next four statements declare and instantiate (create) five TextField objects. The TextField class has been made accessible with the "import java.awt.*" statement above giving access to java.awt.TextField. This is an example of inheritance since the instance variables and methods in java.awt.TextField are available to this class. That is, java.awt.TextField is a superclass of class CustUser. This "TextField" code is used to create boxes that will be used to input text on the screen.

```
private TextField inID = new TextField(10);
private TextField inBAL = new TextField(10);
private TextField inAIR = new TextField(10);
private TextField inAMOUNT = new
  TextField(10);
private TextField out = new TextField(30);
```

The next four statements create and instantiate four new Label objects. The java.awt.Label class is accessible due to the "import java.awt.*" statement above. This is another example of inheritance. The prompts being developed will appear on the computer user's screen.

```
private Label prompt1 = new Label("Please
  enter your 4-digit Customer
  Identification Number: ");
private Label prompt2 = new Label("Please
  enter your account balance: ");
private Label prompt3 = new Label("Please
  enter the Annual Rate of Interest
  for your account (format 0.00): ");
private Label prompt4 = new Label("Please
  enter a deposit or withdrawal amount: ");
```

The next two statements create and instantiate two Button objects named deposit and withdraw. The java.awt.Button class is accessible due to the "import java.awt.*" statement above. This again is an example of inheritance.

```
private Button deposit = new
  Button("DEPOSIT?");
private Button withdraw = new
  Button("WITHDRAWAL?");
```

The following statement creates and instantiates a TextArea object of whatever size is specified. The methods of the java.awt.TextArea class are made possible with the "import java.awt.*" statement above.

```
private TextArea display = new TextArea
  ("This is how many years it would take
  for your balance to double", 20, 60);
```

The next statement creates but does not instantiate a new Customer object named user. (That is, space for the instance variables and methods associated with object user would be allocated in memory but not initialized.) In this banking example, the Customer class defines the variables and methods that are characteristics and capabilities of any bank customer. They are gathered together in one class and demonstrate encapsulation. In addition, the attributes are declared "private." They are therefore "hidden" from all other classes. Note that when the compiler sees this statement, it will seek out the Customer class and will automatically compile it. This code is defining variables of a particular type needed for this banking application.

```
private Customer user;

private int year  = 1,
            id    = 0,
            flag1 = 0,
            flag2 = 0;
private double twicebal = 0,
               bal      =  0.,
               air      =  0;
private String cid,
               inbal,
               annIR;
```

Every method can return a value, or it can return nothing. In the statement shown below, the method does not return a value, as indicated by "void." All applets have a method named "init," but the subclass, e.g., CustUser, may have its own method init() that will overwrite the method init() from the superclass Applet. The method below overwrites the inherited init method. Program execution begins with the first statement in method init. (Similarly, every standalone application program has a method named "main.")

```
public void init()
{
```

The following statements build the GUI interface. The objects specified in this code were created from the TextField and Label input (see code shown above). The code below is used to actually add the items to the window being created, in the order listed:

```
add(prompt1);
add(inID);
add(prompt2);
add(inBAL);
add(prompt3);
add(inAIR);
add(prompt4);
add(inAMOUNT);
add(deposit);
add(withdraw);
add(out);
add(display);
```

The following two statements tell the program to "listen" for mouse clicks on the buttons named deposit and withdraw:

```
deposit.addActionListener(this);
withdraw.addActionListener(this);
```

This next statement sets the size of the GUI interface. The values specify the number of horizontal and vertical pixels to be used:

```
  setSize(600, 500);
}
```

The actionPerformed method below is required due to the actionListener statements above. The ActionEvent argument is the click of a mouse on one of the buttons (i.e., an "event"). When ActionListener has been included in the program, as it was in this example, the program must also include what to do when the user clicks on an object, which is the purpose of actionPerformed.

```
public void actionPerformed(ActionEvent e)
{
      if(flag1 = = 0)
      {
         flag1 = 1;
         cid = inID.getText();
         id = Integer.parseInt(cid);
         inbal = inBAL.getText();
         bal = convertStringToDouble(inbal);
         annIR = inAIR.getText();
         air = convertStringToDouble(annIR);
```

The following statement instantiates the Customer object, which was declared and named user previously in the program. It sends the customer ID, the customer's current balance, and the annual percentage rate to the constructor in the Customer class, where this information is stored for this object:

```
  user = new Customer(id, bal, air);
}
```

The next series of statements identify what the computer is to do if the computer user clicks the deposit button, followed by what to do if the withdrawal button is clicked. In the first statement below, if the deposit button is clicked, the program will read in the string from inAMOUNT TextField, which will be converted to a double value and stored in in_amount1.

```
if (e.getSource() = = deposit)
{
   String amo1 = inAMOUNT.getText();
   double in_amount1 =
     convertStringToDouble(amo1);
```

The following statement passes the double value in_amount1 to the method calc_deposit in the Customer class and stores the value returned (the new current balance) in newBal1.

```
double newBal1 = user.calc_deposit
  (in_amount1);
```

The next statement uses the setText method to print the new balance in the TextField named out.

```
   out.setText("Your new Account Balance
     is: $" + newBal1);
}
```

If it wasn't a click on the deposit button when the "if" statement above was tested, it must have been on the withdrawal button. The "else" section handles this situation.

```
else
{
   String amo2 = inAMOUNT.getText();
   double out_amount2 =
     convertStringToDouble(amo2);
```

The next statement creates a new double value newBal2. It sends the value out_amount2 to the method calc_withdrawal in Customer class. The method then determines whether the current balance is large enough for the withdrawal. If it is not, then it returns a –1, which is stored in newBal2. Otherwise, it returns the new current balance, which is stored in newBal2.

```
   double newBal2 = user.
     calc_withdraw(out_amount2);

   if(newBal2==-1)
   {
      out.setText("Insufficient funds for
        this withdrawal");
   }
   else
   {
      out.setText("Your new Account Balance
        is: $" + newBal2);
   }
}
```

The following section creates and prints to the screen the annual total of a principal compounded daily at a rate of the annual interest rate (air):

```
if (flag2 = = 0)
{
    flag2 = 1;
    NumberFormat dollars = NumberFormat.
      getCurrencyInstance();
    NumberFormat percent = NumberFormat.
      getPercentInstance();
    percent.setMaximumFractionDigits(2);

    display.append("\n at the rate of "
      +percent.format(air/100.0) +"\n \n");

    twicebal = bal;

    while(2*bal >= twicebal)
    {
```

This next statement sends the current balance and annual interest rate to a method named Doubleit for calculation and then returns the new balance. Note that the series of } symbols are needed to indicate that some statements begun earlier in the program, indicated with { symbols, are now being completed.

```
       display.append("Year "+ year +"\t" +
         dollars.format(Doubleit(bal, air))
         + "\n");
      }
   }
}

private double Doubleit(double b,
  double rate)
{
   twicebal = b * Math.pow(1 + rate, year);
   year++;
   return twicebal;
}
```

The method below will convert a string to a double value, a procedure necessary for converting anything read into the program as a string:

```
   private double convertStringToDouble
     (String s)
   {
      return Double.valueOf(s).doubleValue();
   }
}
```

Note that the code presented in this section is not necessarily the most efficient but was developed simply to show the concepts clearly to the reader.

GLOSSARY

Note: The Sun Microsystems' Java and the Carnegie Mellon Software Engineering Institute online glossaries were used as the primary sources for these descriptions.

Abstract Window Toolkit (AWT) A set of graphical user interface components used to provide a subset of functionality common to all native platforms, but largely replaced by the Swing component set developed more recently (described separately below).

API (application programming interface) Serves as a virtual interface (exchange) between two functions; e.g., it specifies how a Java programmer writing an application accesses the behavior and state of classes and objects.

Applet A component usually executed from within a Web browser but can be executed in other applications or devices.

Bean A software component that is reusable and can be combined with other beans to create a complete application (see JavaBeans).

Bytecode Machine-independent code generated by the Java compiler, for execution by the Java interpreter.

Class A type that defines the implementation of an object; i.e., a class defines or identifies the common properties of the object, such as variables and methods, as well as the interfaces that the class implements.

Compiler A program that translates source (original) code into code that a computer can execute; e.g., the Java compiler translates Java source code into bytecode.

Component A unit of software that can be configured at the time of deployment; e.g., J2EE (see below) has several components, including enterprise beans and applets.

Constructor A "pseudo-method" used to create an object; an instance method in Java that has the same name as its class.

Encapsulation A bundling of object knowledge (e.g., data and implementation) into a module, which allows a program to continue to use an object without being affected by changes that might occur to its instance variables and methods.

Error handling A method of taking care of events that are unexpected during program execution (exceptions).

Exception An event that could occur during execution of a program that prevents the program from continuing normally.

Field An item of data in a class; sometimes referred to as an attribute.

Inheritance Allows classes to automatically contain the variables and methods defined in their superclass (see below).

Instance An object of a class.

Interpreter A module that decodes and executes every statement, one at a time, in a program; e.g., the Java interpreter decodes and executes bytecode for the "Java virtual machine."

Java Sun Microsystems' trademark for a set of technologies involved in developing and using software programs in either a standalone or networked environment; Java 2 is the current platform, with multiple editions including J2EE, J2SE, and J2ME.

JavaBeans A portable and platform-independent reusable component model.

Java Development Kit (JDK) The original environment for developing software by using Java; some change involving use of the "SDK" acronym (Software Development Kit) to refer to the environment for developing software applications in the Java 2 platform has occurred.

Java Remote Method Invocation (RMI) A distributed model in which methods of remote objects written in Java can be invoked from other Java virtual machines.

Java Runtime Environment (JRE) A subset of the Java Development Kit that can be distributed to end users; it contains the Java virtual machine, core classes, and supporting files.

Java Virtual Machine (JVM) A computing device that contains a set of instructions for decoding and executing bytecode, as well as other needed components such as registers, a stack, and an area for storing methods.

J2EE The Java Enterprise Edition, targeted for multitiered server-side Web-based applications.

J2SE The Java Standard Edition, considered the core platform and targeted for cross-platform, general-purpose applications.

J2ME The Java Micro Edition, targeted for small consumer and embedded devices such as smart cards and mobile devices.

Method A function or capability defined in a class.

Multithreaded A way of subdividing programs so that different parts of the code can be executed concurrently.

Object A unit of a program that consists of data (instance variables) and functionality (instance methods); considered a main building block of object-oriented programs.

Object-oriented design Use of classes and objects to model the characteristics of an abstract or real object.

Package A group of "types" (see below).

Polymorphism Capability of writing a program in a fairly general form so that it can eventually have multiple forms as needed, e.g., to be able to redefine methods for new subclasses.

Portability Capability of running a program on many different types of hardware platforms.

Private A keyword in Java used in declaring a method or variable so that it can be accessed only by other elements of its class.

Protected A keyword in Java used in declaring a method or variable so that it can be accessed only by other elements residing in its class, subclasses, or classes in the same package.

Public A keyword in Java used in declaring a method or variable so that it can be accessed by elements residing in other classes.

Scope A characteristic that determines where an identifier can be used; e.g., a variable declared within a method can be used only within that block and has local scope; others may have class scope.

Software Development Kit (SDK) Updated components for developing software applications that became available with Java 2, replacing the JDK that had been used with the first release of Java.

Subclass A class derived from another class, which could have one or more classes in between.

Superclass A class from which another class is derived, which could have one or more classes in between.

Swing A collection of graphical user interface (GUI) components written in Java that support the Java virtual machine and provide more functionality than the Abstract Window Toolkit.

Thread A unit of program execution; one process could have multiple threads running concurrently and doing different jobs.

Type A class or interface.

Variable A data item that has been named by an identifier; it has a type, such as integer, and a scope, such as local (i.e., known within its block but not accessible outside the block); a class variable is associated with a class as a whole, while an instance variable is associated with a particular object.

CROSS REFERENCES

See *JavaBeans and Software Architecture; Software Design and Implementation in the Web Environment.*

REFERENCES

Dennis, G. C., & Rubin, J. R. (1999). *Mission-critical Java project management.* Reading, MA: Addison Wesley Longman.

Greenemeier, L. (August 5, 2002). Demand for Java begins to percolate. *InformationWeek*. Retrieved August 26, 2002 from http://www.informationweek.com/story/IWK20020801S0006

Harding, E. U. (2002). DaimlerChrysler automates Java app development. *Software Magazine,* Spring.

Software Technology Review. (2001). *Java*™. Retrieved May 26, 2002 from Carnegie Mellon Software Engineering Institute Web site: http://www.sei.cmu.edu/str/descriptions/java.html

The source for Java technology. Retrieved August 26, 2002, from http://java.sun.com

FURTHER READING

Deitel, H. M., & Deitel, P. J. (2002). *Java: How to program* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Eckel, B. (2002). *Thinking in Java* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Flanagan, D. (2002). *Java in a nutshell* (4th ed.). Sebastopol, CA: O'Reilly & Associates.

Kayl, K. (2002). *Java tech review*. Retrieved May 16, 2002 from http://developer.java.sun.com/developer/technicalArticles/RoadMaps/techreview/

Lambert, K. A., & Osborne, M. (2000). *Java Complete course in programming & problem solving*. South-Western Educational.

Morelli, R. (2000). *Java, Java, Java! Object-oriented problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Polymorphism (2002). Retrieved May 7, 2002 from http://whatis.techtarget.com/definition/0,,sid9_gci212803,00.html

The Java tutorial (n.d.). Retrieved May 7, 2002 from http:// java.sun.com/ docs / books / tutorial / java / TOC.html

What is Java technology? (2002). Retrieved May 16, 2002 from http://java.sun.com/java2/whatis

JavaBeans and Software Architecture

Nenad Medvidovic, *University of Southern California*
Nikunj R. Mehta, *University of Southern California*

INTRODUCTION

A number of techniques have recently emerged to address the problem of consistently engineering large, complex software systems. The three most widely embraced efforts are component-based software development standards such as COM (Brockschmidt, 1994; Sessions, 1997) and JavaBeans (Hamilton, 1997; DeMichiel, 2002), middleware platforms such as CORBA (Orfali, Harkey, & Edwards, 1996) and .NET (Richter & Richter, 2002), and software architecture (Perry & Wolf, 1992). These three approaches are complementary to each other, and a judicious mix of the three is often employed in developing large, complex systems such as those targeted for the Internet.

Java is a popular programming language increasingly being used for component-oriented and architecture-based software development for the Web. The acceptance of Java technologies in both industrial and academic circles can be attributed to both Java language features and the Java platform. While the Java platform has been designed to support the paradigm of “write once, run anywhere,” the Java programming language has been designed to reduce sources of common programming mistakes and provide greater control over the organization of software systems (http://java.sun.com).

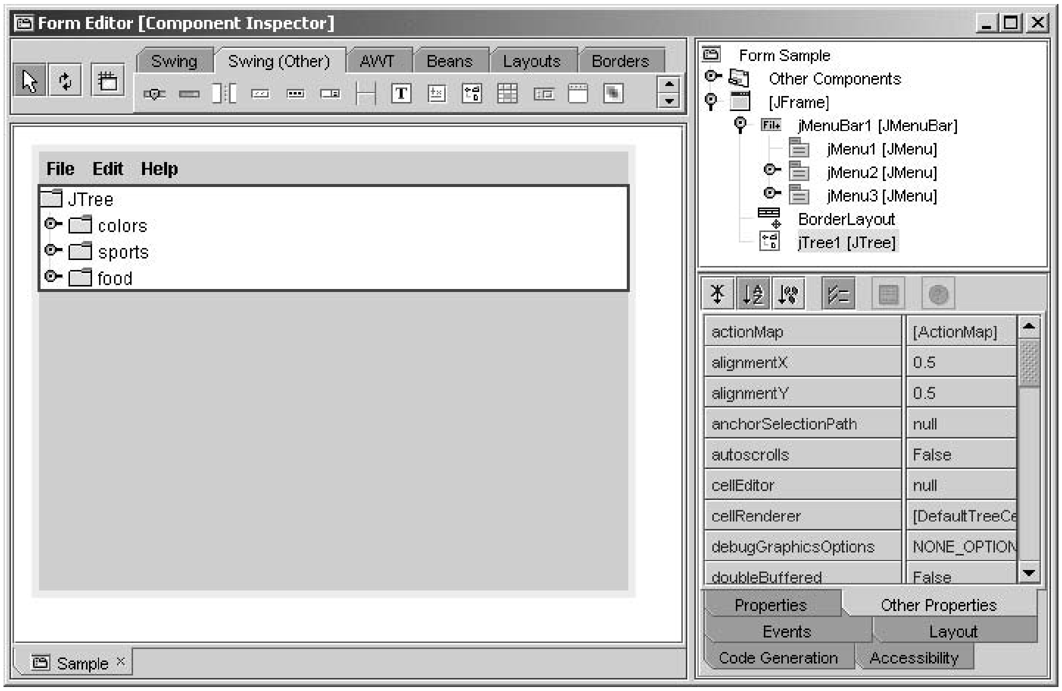
Java’s support for organizing an application from coarse-grained components has an added benefit. It results in greater acceptance and use of the Java programming language and platform in Web-based applications. This chapter focuses on the support for component technology in Java—JavaBeans—and various accompanying technologies, such as a Web server and an application server, that utilize JavaBeans to fulfil the needs of Web application development.

Support for component-oriented software development in Java has gradually improved from the early days of Java to the present. The earliest manifestation of Java’s suitability for organizing applications around components was the JavaBeans framework, first released in late 1996, as a software component model for Java (Hamilton, 1997). JavaBeans are aimed at creating components that can be reused and integrated into larger applications. Beans follow specific conventions and provide a loosely coupled interaction mechanism based on events. Beans also provide design and run time facilities such as introspection, customization, persistence, and security. Support for JavaBeans is provided in the form of standard libraries packaged by the Java run time environment in the *java.beans* package. Beans are commonly used for developing graphical user interface (GUI) widgets as well as nongraphical elements (e.g., data structures) that provide processing support for the GUI widgets.

The rest of this chapter is organized as follows: First, we introduce the syntactic and semantic constructs of JavaBeans and its use in developing and delivering components. Second, we focus on the use of JavaBeans for distributed applications in the form of Enterprise JavaBeans (EJB). JavaBeans are also widely used in other Java technologies, and we discuss major Java application programming interfaces (API) that are built on top of JavaBeans for developing Internet applications. Next, we present a discussion on the use of JavaBeans in component-based software development and its relationship to software architecture. Finally, we summarize the JavaBeans technology and its role in software architectures.

JAVABEANS

Java has emerged as a widely used object-oriented programming language, and its successful use can be attributed to its support for a variety of technologies that are used to solve specific infrastructure and application requirements. The JavaBeans technology has enabled large-scale, coarse-grained reuse of components and has influenced a number of other Java technologies. JavaBeans



actionMap	[ActionMap]
alignmentX	0.5
alignmentY	0.5
anchorSelectionPath	null
autoscrolls	False
cellEditor	null
cellRenderer	[DefaultTreeCe
debugGraphicsOptions	NONE_OPTION
doubleBuffered	False

Figure 1: NetBeans IDE 3.4 Visual Form Editor.

is the technology that enables the construction of visual components as well as the binding of those components into a GUI application using a visual design tool such as the one shown in Figure 1. This tool supports "drag-and-drop" design of a GUI application composed from smaller visual components called GUI widgets. Such a tool is capable of generating Java source code for the application based on the visual design of the application. An application programmer may then modify the source code to provide additional details of the application logic.

The JavaBeans API was originally proposed as a software component model for Java (Hamilton, 1997) chiefly to allow users to manipulate and connect widgets in a GUI. However, applications in many other domains can also benefit from its loosely coupled architecture and support for introspection, customization, and persistence. For example, JavaBeans are used in mail clients to manage connections to a mail server and manipulate the contents of one's mailbox.

In the original JavaBeans specification (Hamilton, 1997, p. 9), a JavaBean is defined as "a reusable software component that can be manipulated visually in a builder tool." This definition matches the goal of JavaBeans, which is to create a software component model. To realize the goal, Java defines an API, programming conventions, and naming patterns to promote component-oriented application development. Nevertheless, JavaBeans adhere to the same rules as those followed by any Java class at the level of source code as well as the byte code produced. A single JavaBean can be implemented as a collection of classes all working together to provide a unified set of functions. This enables the creation of coarse-grained components out of a set of Java classes.

Usage and Applications

JavaBeans provide separate interfaces for design time and run time capabilities. The design time configuration and customization capabilities supported in JavaBeans set them apart from regular Java classes and class libraries. Some JavaBeans are GUI elements such as windows and buttons, some are more complex visual elements such as database viewers or spreadsheets, still others are nonvisual components such as database connections and algorithms. Java Abstract Windowing Toolkit (AWT) is the earliest library of GUI widgets that employ JavaBeans. AWT was aimed at supporting the development of richer user interfaces for the Web using applets. AWT only provides simple widgets such as a button, a panel, and a window, and has serious performance drawbacks due to its use of heavyweight components for every AWT widget. To overcome the limitations of AWT, the Java Foundation Classes (JFC) library was introduced with a much richer set of widgets (Geary, 1999). Moreover, JFC, also known as Swing, uses an improved lightweight architecture to deliver better performance by using far fewer components per widget.

Beans are primarily targeted for use in visual builder tools such as Web page builders, integrated development environments (IDEs), visual application builders, GUI layout builders, and server application builders, although JavaBeans are entirely human-programmable as well. In this context, it is necessary to understand the distinction

```
public class Circle extends java.awt.Component implements java.io.Serializable {
   ...
   // Paint the circle on the screen
    public void paint(Graphics g)
   ...
   // Get the boolean Shown property
   public boolean isShown()
   ...
   // Get the Radius property
   public int getRadius()
   ...
   // Set the Radius property
   public void setRadius(int radius) throws PropertyVetoException
   ...
   // Get the Border Size property
   public int getBorderSize()
   ...
   // Set the Border Size property
   public void setBorderSize(int size)
   ...
   // Add an event listener for change of radius
   public void addRadiusChangeListener(RadiusChangeListener l)
      throws java.beans.TooManyListenersException
   ...
   // Remove an event listener for change of radius
   public void removeRadiusChangeListener(RadiusChangeListener l)
   ...
   // Add an event listener for change of any property
   public void addPropertyChangeListener(PropertyChangeListener l)
   ...
   // Remove an event listener for change of any property
   public void removePropertyChangeListener(PropertyChangeListener l)
   ...
   // Add an event listener for a vetoable change of any property
   public void addVetoableChangeListener(VetoableChangeListener l)
   ...
   // Remove an event listener for a vetoable change of any property
   public void removeVetoableChangeListener(VetoableChangeListener l)
   ...
}
```

Figure 2: Circle—a JavaBean.

between the two sets of interfaces offered by JavaBeans. Design time capabilities allow a JavaBean to run inside a visual builder tool and provide design information about itself to the tool so that a designer can customize the bean. Run time capabilities allow a JavaBean to manifest its behavior in an application so that the end-user can use its functionality. The separation of design time from run time capabilities enables richer visual editing of the design, while not burdening the eventual application with redundant design logic.

JavaBean Characteristics

A JavaBean does not extend any standard class, nor implement any standard interface, although visual components always extend java.awt.Component or its subclasses. However, JavaBeans typically share a set of characteristics in the form of a standard set of facilities that enable beans to be easily combined into applications. These facilities in order of usage from design to run time are *introspection, customization, properties, events, persistence, packaging,* and *methods*. Introspection and customization are used at design time; properties, events, persistence, and packaging are used at both design and run time; methods are used exclusively at run time. We discuss these facilities in the reverse order, that is, from run time to design time, to allow the reader to grasp the significance of the design time facilities after understanding the utility of the run time facilities. Figure 2 shows a sample JavaBean called *Circle*, which is used as an example for the rest of this section.

Methods

A JavaBean provides methods for services it exposes. Methods are defined for use at run time when the application wishes to use the bean's services. They provide a low-overhead, synchronous communication mechanism for components in the JavaBeans' component model. JavaBean methods follow the same lexical and syntactical rules as those followed by Java methods and have the same

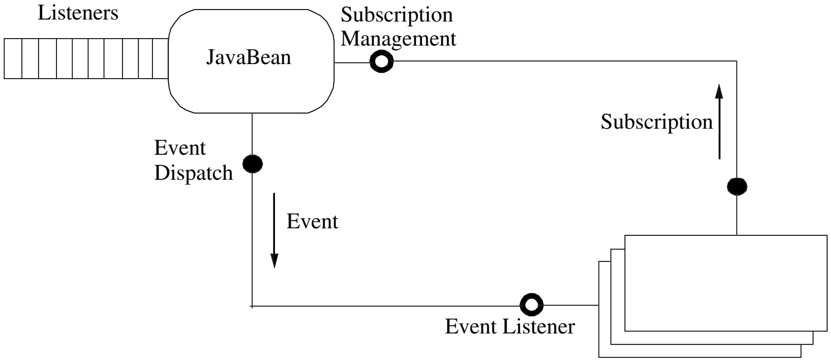


Figure 3: JavaBeans event model.

invocation semantics and mechanisms as those available to regular Java methods. JavaBean methods are defined with the public modifier that makes them available to other components in the application. Other method modifiers such as *final* and *synchronized* can be used with JavaBean methods. However, not all methods defined in a JavaBean are considered to be JavaBean methods. This distinction is necessary to distinguish high-level services offered by the bean from low-level implementation mechanisms such as Java declarations, without burdening the language with unnecessary keywords and concepts.

Any public method defined with a name that does not start with the reserved prefixes *get, is, set, add,* and *remove* is considered to be a JavaBean method. The reserved keywords are used in conjunction with JavaBean properties and events as explained in subsequent sections. As shown in Figure 2, Circle is a visual component that can be used as a widget for drawing circular shapes of all sizes on a graphical window using the method called paint. Other Java methods defined in this class are not considered JavaBean methods.

Events

Events are the occurrences of incidents of interest to components. Events are used in JavaBeans as a publish–subscribe based communication mechanism. Figure 3 graphically represents the JavaBean event architecture, which is based on the Java event model (Englander, 1997). Any object, not necessarily a JavaBean component, registers with beans that generate events that are of interest to it. Each interested object, called an *event listener,* makes a call to the *subscription management* port of a JavaBean for registering an interest in events. The JavaBean records all listeners for each event category. Later, when an event occurs, the bean sends the event message, an instance of *java.util.EventObject* or one of its subclass, to all registered listeners of that event as a regular Java method call.

A single JavaBean may provide a number of events defined through methods that follow a certain naming pattern: subscription for an event is performed by calling methods of the form *public void add <EventName>Listener(<EventName>Listener l)* and unsubscription is performed using methods of the form *public void remove<EventName>Listener(<EventName> Listener l)*. These methods together define the subscription management port shown in Figure 3. An example of this naming pattern can be seen where the JavaBean *Circle* defines a *RadiusChange* event.

This subscription management port allows multiple subscribers to be registered for a single event. Sometimes it is necessary to constrain the number of listeners, and a common use is to perform unicast event notification, i.e., notification to only one listener. To support this, a bean keeps the number of listeners for the event to 1 or less. While one listener is registered for the event, if another listener tries to register for the same event, the bean can throw a *java.beans.TooManyListenersException* indicating that the requested event subscription cannot be allowed. An example of unicast events is the *RadiusChange* event as shown in Figure 3.

Properties

Properties provide access to the state of a JavaBean that can be read and/or changed and have a name and a type. Properties allow a builder tool to access the state of a JavaBean for customization, as well as for an application to modify the behavior of a bean by changing its state. Thus properties are used at design time as well as run time. Properties are defined as a set of methods in a JavaBean instead of being defined as public fields. This enforces encapsulation by hiding away the fields defined in the Java class implementing the JavaBean, and instead exposing only the bean's high-level properties.

A naming pattern is used to identify properties of a JavaBean. Since all properties are accessed through method calls, special prefixes *get, is,* and *set* are used to identify the methods used for property access. For example as shown in Figure 2, the *Circle* bean has a property named *BorderSize* of type *int* that is accessed through methods *getBorderSize* and *setBorderSize* for reading and changing the property, respectively. Boolean properties such as *Shown* from are read using the *is* prefix, i.e., as *public boolean isShown()*.

JavaBeans support both single- and multivalued properties. Single-valued properties are accessed through simple accessor methods, as illustrated above by the *BorderSize* property, whereas a multivalued property, also called *indexed* property, requires that an index be provided in the accessor methods. For example, *public Color getColor(int index)* would be the signature of a getter for the indexed property *Color*, whereas *public void setColor(int index, Color value)* would be the signature of a setter for the same property. A property can be either read only, write only, or read/write, depending on which of the accessor methods are defined in the bean. Moreover, based on how the bean responds to a change in the value of the property, three kinds of properties—*simple*, *bound*, and *constrained*—are possible.

In the case of a simple property, the JavaBean unilaterally changes the property value without alerting other components that depend on the property. Bound properties are used when the bean is required to notify any interested components of the change in the value of one or more properties. Interested components are notified through an event of class *java.beans.PropertyChangeEvent*. Such components may respond to the event in a coordinated manner to maintain a stable application state, e.g., by updating their own properties, changing the properties of other beans, or calling a method on a bean. Although a change made to any property can result in an event, those properties for which a change in value causes the bean to generate a PropertyChangeEvent are considered as bound properties. A bean should change the value of its property before firing a PropertyChangeEvent to ensure that components receiving the events do not find an inconsistency between the event and the state of the bean. It is possible to register interest for changes to any property of a bean or to individual properties. This is accomplished by providing two sets of methods for registering listeners, one called *addPropertyChangeListener(PropertyChangeListener p)* for registering interest in all bound properties of the bean, and another of the form *add<Property Name>Listener(PropertyChangeListener p)* for registering interest in a specific bound property of the bean.

Simple and bound properties do not constrain changes to their values to ensure consistency between beans. Sometimes, it is necessary to do so when beans are combined so that those property values that are unacceptable to listeners monitoring these properties can be rejected. In the case of constrained events, the bean notifies listeners using a *java.beans.VetoableChangeEvent* to request their approval for a change in the value of a property before the change can be applied. If even one such component rejects the change by throwing a *java.beans.PropertyVetoException* while processing the event, the new property value is disallowed. Accordingly, the accessor methods for a constrained property follow a slightly different naming pattern in that the setter method is defined to throw a *java.beans.PropertyVetoException*. A bean persists the changed value only if it receives the approval of all interested components, and then sends a *PropertyChangeEvent* once the property value is changed. This two-phase protocol allows for distributed approval to take place without causing transient side effects. It is possible to register interest for changes to any vetoable properties of a bean as well as to a specific property. This is accomplished by providing two sets of methods for registering listeners, one called *addVetoableChangeListener(VetoableChangeListener v)* for registering interest in all constrained properties of the bean, and another of the form *add<Property Name>Listener(VetoableChangeListener v)* for registering interest in a specific constrained property of the bean. Figure 3 shows an example of a specific constrained property called *Radius*. When placed with other widgets on a window, the *Radius* of a *Circle* might need to be acceptable to other widgets present in the same window. The vetoable property *Radius* allows for a check to be made with the neighbors of a *Circle* object in the window to ensure that a change in its *Radius* does not consume space occupied by another widget.

Persistence

A JavaBean may store information internally in order to perform its behavior. Some of this information may be accessible as properties and other information may be private to the bean. To re-create an instance of a bean, it is necessary to record and restore this internal information. This information management behavior is called persistence. JavaBeans can persist their state so that a customized bean can be re-created later with the set of previously customized properties. JavaBeans supports persistence through three mechanisms *serialization*, *externalization,* and *encoding*.

The first mechanism, serialization, is based on the use of a standard Java interface namely, *java.io.Serializable*. Implementing this interface in any Java class will generate additional byte code for persistence when the class is compiled. This special byte code is capable of recording all the primitive attributes as well as the serializable nonprimitive attributes of the object being serialized. The Java virtual machine (JVM) serializes any super class objects that implement the Serializable interface in the same manner. A bean can exert more control over the manner in which its declared fields are persisted by implementing special methods *readObject* and *writeObject*. An object can be serialized to any stream format such as a disk file or a network connection and deserialized when the stream is read, thus communicating the design information to an application at run time. Deserialization offers a more flexible and preferable way of instantiating beans compared to regular constructor-based instantiation as the exact set of properties of a customized bean can be restored.

In addition to the standard serialization, Java supports externalization, which gives a JavaBean complete control over the amount of and manner in which internal information of the bean is serialized using standards such as compound documents (Brockschmidt, 1994). A Java class must implement *java.io.Externalizable* interface to use externalization, which in turn requires the class to provide definitions of methods *readExternal* and *writeExternal,* both used with binary object streams.

With Java Development Kit 1.4, Java also provides a more resilient persistence mechanism in the form of *java.beans.Encoder*. This mechanism only persists properties of a bean instead of all the internal fields of the bean's Java class. This mechanism is especially used with XML as provided in the *java.beans.XMLEncoder,* which records the properties textually instead of in the binary formats that are used in serialization and externalization. Since this technique does not depend on any private class

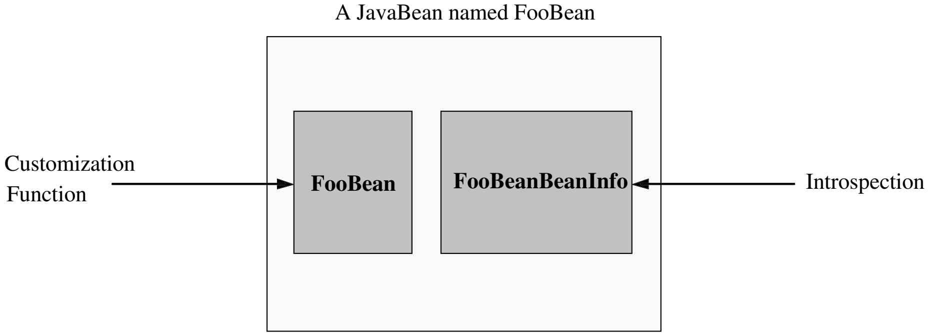


Figure 4: Separating JavaBean behaviors.

information, this technique is more resilient as it can survive changes in the JVM as well as versions of the bean; it is more robust as a partial damage to the API of the bean can still allow remaining properties to be read; and it is more compact as default properties of a bean can be safely removed from the persistent form of the bean. The encoder is responsible for traversing the entire graph of properties used in a bean as well as structuring and recording the information about each bean in this graph into a single stream. A counterpart of this encoder, the *java.beans.XMLDecoder,* is available to reconstruct this graph of beans from its textual persistent form. While the binary representation is more useful for interprocess communication, the textual form is preferred for archival purposes.

Packaging

JavaBeans use the Java archive (*Jar*) file mechanism for packaging multiple bean classes, other associated classes, and serialized beans that can be distributed with the application for use in a visual builder tool. The Jar file uses the ZIP format, which has the benefits of compression as well as support on a large number of platforms. Jars also contain a manifest file, which is used to record additional information about the contents of the Jar such as the versions and names of JavaBeans. Jar files are often used to record other information in the form of graphics files, textual resources and documentation about the bean, thus providing a manageable means of distributing JavaBeans. At design time Jar files can be created from the beans used in the design, which can then be deployed at run time to produce executing instances of the same beans using deserialization mechanisms discussed in the previous section.

Introspection

Introspection is the ability to acquire design information about the bean so that a visual builder tool can assist a designer in using the bean to compose applications or a composite JavaBean. In the process of introspection, a visual builder tool can discover the customization and configuration capabilities of the bean. Introspection involves providing explicit high-level information about the bean that can help a designer use it. Each JavaBean that supports introspection provides a class that implements the *java.beans.BeanInfo* interface. With beans that do not provide or omit certain high-level introspection information through the BeanInfo interface, a builder tool derives this information by using the Java reflection APIs and applying the JavaBeans naming conventions to discover properties, events, and methods.

As shown in Figure 4, the bean customization and functional behavior is separated from the introspection behavior. Since introspection behavior is only required during design, such a division of behaviors results in lightweight beans that can be employed at run time without any need to manage large amounts of introspection data about the bean. The names of the two classes (the bean and its BeanInfo) are coordinated by a *naming pattern*. According to this naming pattern, the names of these two classes should only differ by the suffix BeanInfo used for the introspection class. Thus a JavaBean by the name *FooBean* will provide at least two classes, *FooBean* and *FooBeanBeanInfo*. The BeanInfo interface provides the following information about the bean:

- A set of descriptors for the events emitted by the bean,
- A set of descriptors for the properties supported by the bean,
- A set of descriptors for the methods supported by the bean,
- An icon for the bean,
- A descriptor for the bean specifying the class for the bean as well as the class for the customizer of the bean (JavaBean customization is explained in the next section), and
- Introspection information about other beans that are relevant to this bean.

A default implementation of this interface is provided in the Java API class *java.beans.SimpleBeanInfo,* which defines a default behavior for each of the methods in the BeanInfo interface. This allows a developer of the JavaBean to avoid implementing those methods in the interface that are not relevant while specifying relevant introspection behavior in a class that extends the *SimpleBeanInfo*. Figure 5 gives an example of a BeanInfo implementation for the Circle JavaBean introduced in Figure 2. More information about the JavaBeans API for introspection can be found in several elaborate references on this topic (Hamilton, 1997; Englander, 1997).

```
public class CircleBeanInfo extends java.beans.SimpleBeanInfo {
   public PropertyDescriptor[] getPropertyDescriptors() {
      PropertyDescriptor[] result = null;
        try {
         PropertyDescriptor pd1 = new PropertyDescriptor(“radius”, int.class);
         PropertyDescriptor pd2 = new PropertyDescriptor(“borderSize”,
                  int.class);
         result = {pd1, pd2};
      }
      catch (Exception e) {
         System.err.println(“CircleBeanInfo: unexpected exception” + e);
      }
      return result;
   }
   public EventSetDescriptor[] getPropertyDescriptors() {
      EventSetDescriptor[] result = null;
      try {
         EventSetDescriptor ed1 = new EventSetDescriptor(Circle.class,
            “radiusChange”, RadiusChangeListener.class, “radiusChanged”);
         result = {ed1};
      }
      catch (Exception e) {
         System.err.println(“CircleBeanInfo: unexpected exception” + e);
      }
      return result;
   }
}
```

Figure 5: Circle JavaBean introspection.

Visual builders use a JavaBeans utility, *java.beans. Introspector,* to locate the introspection information about a bean. The Introspector then locates the BeanInfo class for the given bean using the BeanInfo naming pattern. The Introspector then constructs the BeanInfo using the beans introspection class, and fills in the missing aspects of the introspection through reflection-based analysis of the bean. Thus, introspection allows users of a bean to obtain high-level information for using the bean.

Customization

A visual builder is used to design an application that uses JavaBeans. The introspection facilities of a JavaBean discussed in the previous section enable the builder to understand the structure of beans. A builder uses this information to create suitable tools for manipulating the properties of beans as well as for providing visual techniques for connecting various beans and defining their interaction. For simple properties such as numbers and text, simple text editors are sufficient tools for specifying the values of properties. One of the immediate benefits of using a visual builder is the ability to specify sophisticated properties of a bean, such as colors and font types, more naturally using intuitive techniques present in rich *property editors* such as a color palette. A bean can also provide its own property customization user interface to give it complete control over how a designer provides values for its properties.

Property editors are used only at design time to capture properties of a bean. The property editors to be used for manipulating each property can be specified in the introspection information about that property. Since it is not necessary for the bean to provide introspection information about all its properties, nor to provide a custom editor for any of the properties, the development tool automatically selects an editor for properties based on the type of the property. Custom property editors can be developed and packaged with the beans, and specified for use in the introspection information. A custom property editor implements the *java.beans.PropertyEditor* interface, which ensures consistent behaviors across all property editors, both tool-provided and custom.

Property editors are used for editing one bean property at a time. On the other hand, a customizer may be provided for customizing the entire bean. A bean customizer should implement the *java.beans.Customizer* interface. When a customizer is provided for a bean, it takes over all aspects of customization of the bean, which could be a super set of the properties exposed by the bean. Introspection of a bean reveals the customizer to be used for a bean. More details about customization may be found in Englander (1997).

DISTRIBUTED JAVABEANS

When an application consists of JavaBean components distributed over a network, serialization is used in combination with Java remote method invocation (RMI) (Sun Microsystems, 2001) to transport methods, events, and JavaBean instances over the network. Java RMI is a protocol that allows methods of a Java object to be remotely invoked from another Java virtual machine, possibly residing on a different host across the network. RMI is used in conjunction with Java naming and directory interface (JNDI) to look up registered instances of JavaBeans that are hosted remotely. Once the required instance of a JavaBean is located using JNDI, a client of that instance can make a method call using RMI. Parameters passed in the method are serialized using the serialization mechanism described earlier in this section. The method is executed remotely, and results of the execution as well as any exceptions are returned to the caller. RMI enables object-oriented invocations of methods without the need to implement low-level data and network communication logic.

Enterprise JavaBeans

JavaBeans are commonly used in graphical application development, and this has led to the development of other component models. Another prominent Java component model is Enterprise JavaBeans, which is targeted at the development and deployment of component-based distributed business applications (DeMichiel, 2002). EJB derives its name from JavaBeans although EJBs are used for a very different purpose. An EJB is an "enterprise" component, i.e., a component used in large, distributed systems that can be concurrently accessed by a large number of users over a network. The demands of a distributed system used at an enterprise scale requires that special attention be paid to the architecture of the components.

The EJB component model was developed to address the need for providing data access and information processing to large groups of users. In an enterprise application, data is stored in large databases, with a large number of users simultaneously accessing the data. For business critical applications, high availability and security are extremely important. Moreover, such applications consist of components that are distributed over a network, further requiring management of resources such as processors, memory, and network communication. The EJB component model aims at simplifying the task of developing, deploying, and maintaining such applications. In the EJB component model a number of lower level infrastructure services are provided by EJB middleware called an EJB *container* in which the EJB components are deployed. It should be noted that, unlike JavaBeans, EJBs are not visual components and need not be visually manipulated by users. Moreover, EJBs do not require customization of behavior, but instead require configuration of resources required by the EJB. The goal of EJB—to simplify distributed application component development—is achieved by providing a number of built-in services in the EJB container.

Characteristics

All Enterprise JavaBeans extend *javax.ejb.EJBObject* and provide implementations of all the abstract methods defined in this class in a *template method pattern* (Gamma, Helm, Johnson, & Vlissides, 1995). A standard set of facilities are required in enterprise applications to ensure efficient usage of resources as well as scalability. These facilities, provided in the EJB container that supports the deployment and execution of EJBs, arranged sequentially in order of usage from deployment to run time, are

Deployment: Classes of an EJB component are packaged in the form of an *enterprise archive (EAR)* file, building on the Jar file format, which contains a *manifest* for describing the contents of the EAR file, and *deployment descriptors* that define the configuration of facilities required by the component.

Memory and instance management: The EJB container performs memory management for EJB components, over and above JVM garbage collection, in order to reduce overheads involved in creating and initializing instances of components, as well as to keep an optimum number of instances available to serve average and peak demands.

Thread management: The EJB container manages the run time environment of EJB components including the Java threads available for execution.

Communication management: The EJB container provides services over the network to EJB clients and dispatches calls received to EJB components.

Security: The EJB container enforces security by controlling access to restricted components as well as performing authentication of clients requiring access to restricted components.

Location: An EJB container enables the distribution of components across the network, with run time binding of services to their clients, thus enabling transparent load balancing and failure management.

Messaging: Asynchronous communication between components is supported through message queues that store and dispatch messages with varying levels of delivery guarantees to network destinations.

Invocation: The EJB container invokes available instances of EJB components for services requested by EJB clients.

Persistence: The EJB container ensures durability of component state by entrusting it to external storage systems such as a database management system.

Transaction management: The EJB container supports grouping together multiple database accesses corresponding to a single transaction by defining and managing transaction boundaries.

Internal Architecture

Figure 6 shows the containment relations between the various constituents of the EJB architecture.

The EJB architecture consists of three kinds of EJB components, namely session beans, entity beans, and message-driven beans, and the EJB container. Session

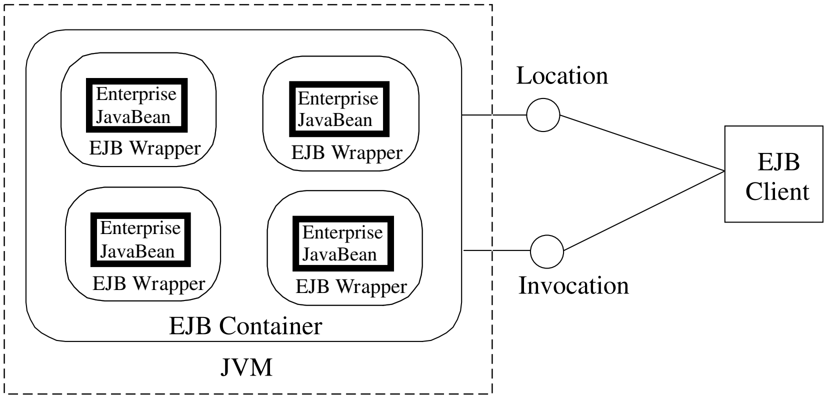


Figure 6: Enterprise JavaBeans container architecture.

beans provide business logic to clients; entity beans provide an object-oriented representation of persistent application data; and message-driven beans process asynchronous messages. The EJB container is responsible for providing infrastructure services identified in the previous section whereas EJB components provide application services. The EJB container provides two RMI-based entry ports to EJB clients, one each for locating an EJB and for invoking a method on the EJB. In the EJB architecture, a third constituent, an EJB *wrapper,* plays an important though invisible role. The EJB client cannot access an EJB instance directly, but can only do so through the EJB wrapper. The EJB wrapper is a *connector* (Mehta, Medvidovic, & Phadke, 2000) between an EJB and its container generated from the deployment descriptor of the bean. This wrapper provides the deployed instances of an EJB with their required services. The EJB wrappers can be generated either within the containers when the bean is deployed or using special tools when the EJBs are packaged for deployment.

Figure 7 shows the EJB invocation model where EJB clients communicate with the EJB through interfaces rather than directly invoking the implementation class. This separation of interfaces from implementation results in greater encapsulation of the business logic inside the EJB. EJBs have two interaction ports, i.e., access interfaces: *home interface* and *remote interface*. The home interface provides methods for creating new instances of the EJB as well as for locating required EJB instances. In the case of a session bean, the home interface is only used to create instances of the EJB, whereas in a message-driven bean this port is not used. In the case of an entity bean, the home interface is used to load specific instances of an entity from a persistent data store, such as a relational database, based on the properties of those instances.

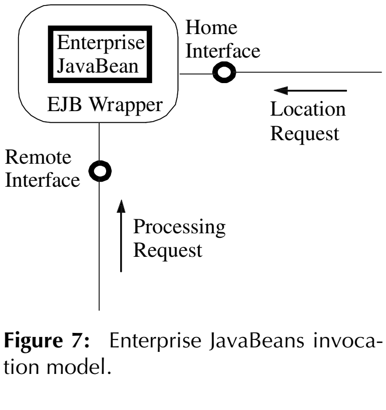


Figure 7: Enterprise JavaBeans invocation model.

The remote interface is used by an EJB client for performing business services using a specific EJB instance. A message-driven bean does not require a remote interface since it does not expose any business service except the standard message consumption service. Session and entity beans may expose specific methods in the remote interface to meet application needs.

EJB wrappers generated in an EJB container intercept calls made to the home and remote interfaces and pass the calls to the EJB for processing after fulfilling security, transaction, and instance management. Certain calls can directly be processed by the wrapper including those that involve *container-managed persistence,* where the persistence is directly managed by the container based on the deployment descriptor of an entity bean. Others are delegated to the implementations provided in the EJB by its designer.

Comparison with JavaBeans

JavaBeans and Enterprise JavaBeans have emerged as the primary component models of architecture-based software development, each in its own niche. JavaBeans are used for developing desktop applications that provide rich graphical interfaces to users. Enterprise JavaBeans, on the other hand, are used in *back office* applications as distributed components providing business logic, data processing, and persistent data access. JavaBeans serve to simplify development of applications through the use of visual builders that can provide a visual means to designers for composing applications. Enterprise JavaBeans, on the other hand, are not necessarily aimed at supporting a visual development environment, although wizard-like tools are often provided to “jump start” their development.

JavaBeans do not require inheritance from a specific class, although visual components are expected to extend *java.awt.Component*. On the contrary, EJBs are always expected to extend *javax.ejb.EJBObject*. JavaBeans can execute in standard JVM environment, whereas EJBs require a container typically provided in a Java2 Enterprise Edition (Shannon, 2002) server. JavaBeans require customization before they can be integrated into an application; the customized instances are recorded as serialized instances that can then be re-created in the application. EJBs on the other hand are customized by configuring the deployment descriptor and repackaging the beans. In both cases, certain customizations can be performed even at run time. Both types of components require standards-based packaging of classes to ensure manageable deployment units, and interoperability with various vendor implementations of JVM and EJB containers.

JavaBeans support an architectural style of loose component integration and decoupled interaction in the form of events, whereas EJB interaction takes the form of asynchronous messages and synchronous method calls. JavaBeans provide two ports for interactions: subscription management and event dispatch. EJBs also provide two ports but with different purposes: home interface for location and remote interface for business services.

JAVABEANS IN OTHER JAVA TECHNOLOGIES

JavaBeans define a generic component model that can be used in different domains where Java is used. Since JavaBeans is suited for use in a client environment as explained previously, most uses of JavaBeans and derivative technologies employing them have emerged for the client environment. We present a brief overview of these technologies in the remainder of this section.

Java Abstract Windowing Toolkit (AWT)

Java AWT is part of the standard API for producing GUI applications for Java (Sun Microsystems, 2002a). Java AWT was originally introduced in Java version 1.0.2, and subsequently reengineered to use JavaBeans. AWT beans satisfy the goal of JavaBeans—reusable software components that can be manipulated visually in a builder tool. The *beanified* version of AWT presents GUI widgets (such as buttons, text input controls, and check boxes) as JavaBeans replete with properties, methods, and events. AWT widgets can be customized for use in an application using a bean customizer, properties can be edited, and events can be managed both at design time and run time. An example of the use of constrained events in AWT is when various beans are placed on a *panel;* if a single widget is resized, then every widget in the panel is alerted to a resulting modification in its size and/or location; if any widget disapproves such a change, the change in the original widget is disallowed. AWT also uses the same event model as employed in JavaBeans to communicate the occurrence of user actions on the GUI widgets as events to the application class handling them.

Java Database Connectivity (JDBC)

JavaBeans can be persisted on file systems using any of the three standard persistence mechanisms: serialization, externalization, and encoding. However, these standard mechanisms do not offer the same levels of durability and integrity required in many business systems. Such requirements can be met by storing data in a database that is supported in Java with the help of JDBC technology. JDBC provides classes for establishing connections, performing queries, and manipulating result sets of data returned from the database (Ellis & Ho, 2002). JDBC version 2.0 introduced JavaBeans to the existing API by adding events and properties to the low-level data access model.

Two JDBC interfaces primarily employ JavaBean mechanisms, *javax.sql.RowSet* and *javax.sql.PooledConnection*. A RowSet can be configured at design time and used at run time to retrieve and persist information to and from the database. The RowSet is responsible for creating its own database connections and performing data access using lower level JDBC APIs based on specified configuration. It generates a RowSetEvent to indicate the occurrence of three kinds of events: *cursorMoved, rowChanged,* and *rowSetChanged*. A PooledConnection is used to make intermittent access to the database more efficient by pooling the use of a lower level database connection. A ConnectionEvent can be emitted to indicate the occurrence of two kinds of events: *connectionClosed* and *connectionErrorOccurred*. The additions of JavaBeans mechanisms to JDBC enables database connectivity to be used at the level of components manipulated through visual builder tools.

JavaMail

JavaMail is designed to create rich mail client components based on standard mail protocols that can be integrated into applications (Sun Microsystems, 2000). It can be used at design time to record properties for establishing a connection to the mail store and at run time to retrieve information from the mail store. A set of closely related JavaBeans are available from this API: *Message, Folder, Store,* and *Transport*. Seven types of events are generated by JavaMail components: *ConnectionEvent, FolderEvent, MailEvent, MessageChangedEvent, MessageCountEvent, StoreEvent,* and *TransportEvent*. These events are emitted in response to user actions and data received from mail servers.

Java Management Extensions (JMX)

The Java management extensions API defines an architecture for managing applications and networks (Sun Microsystems, 2002b). Application components and complete applications can be instrumented to obtain information about their functioning, as well as to control their run time behavior. Components that can be managed need to follow certain naming patterns and implement a *management interface*. An *MBean* (managed bean) is a Java object that implements the management interface for the component being managed. MBeans can define *meta data* that provide introspection information about the management interface. The MBean instance also needs to register itself with an *MBean server* thus making the bean manageable even from outside the JVM. The management interface contains properties, methods, and notifications available for management in the component. Properties and methods follow the same naming pattern as JavaBeans, while notifications behave identically as JavaBean events.

JAVA AND ARCHITECTURE-BASED SOFTWARE DEVELOPMENT

Software architectures provide high-level abstractions to represent and reason about software systems (Perry & Wolf, 1992). The focus of software architecture is to represent the structure of a system without exposing all the complex details of its implementation. JavaBeans and various Java technologies based on it have developed programmatic means of implementing conceptual software architectures. In this section we provide a brief synopsis of software architectures and discuss how JavaBeans and other Java APIs come together to realize software architectural concepts.

Concepts of Software Architectures

Complementing other approaches to developing large, complex systems, software architectures provide support for composing software systems from coarse-grained *components*. A software component in this sense is an element of a software architecture that performs processing and records state. Another important aspect, particularly magnified by the emergence of the Internet and the growing need for distribution, is *interaction* among components. Component interaction is embodied in the notion of *software connectors*. Components and connectors can be linked together in specific *configurations* to create the architecture of a software system and provide a higher level of abstraction as compared to classes and methods. In a study of current architecture description languages (ADLs), Medvidovic and Taylor (2000) proposed that an ADL describe applications in terms of at least the following: software *components* with their *interfaces, connectors,* and *configurations*.

Architectures of a family of systems contain many similarities to each other, and such similarities of organization can be represented as *architectural styles* (Shaw & Garlan, 1996). An architectural style defines a set of rules that describe or constrain the structure of architectures and the way in which their components interact. Architectural styles improve the efficiency of application development since individual applications can be designed using a similar set of rules and, potentially, underlying infrastructure.

JavaBeans and related Java technologies therefore lend support to software architectures in two main ways: support for architectural abstractions and support for architectural styles.

Support for Architectural Abstractions in JavaBeans and Related Technologies

JavaBeans define a software component model for Java. Every technology described in this chapter defines a set of *components:* EJB provides session, entity, and message-driven beans; AWT provides GUI widgets; JDBC provides pooled connections and row sets; JavaMail defines folder, message, store, and transport; and JMX provides MBeans. Moreover, each technology defines *connectors* that can be used for interaction between components: JavaBeans, AWT, JDBC, JavaMail, and JMX use properties, methods, and events for interaction; EJBs use RMI, method invocation, and asynchronous messages for communication and coordination; EJB containers provide a number of facilities in the EJB wrapper, which acts as a connector between the EJB and its container. Finally, an architectural configuration is achieved in one of two ways: instantiation and registration.

In instantiation-based configuration, the client of a component instantiates the component that provides services and maintains a reference to it, thus creating a configuration in which components hold references to every component they use. This technique is used in JDBC, AWT, and JavaMail. This technique allows for efficient interaction among components. However, the drawback of the technique is that components are highly coupled: any run time component additions and removals need to be reflected in all affected components. Furthermore, a configuration created in this manner is only implicitly represented in the dependency information distributed across the components, hampering system understandability.

In registration-based configuration, a component is instantiated by its container to which the instance of the component is registered. The container then offers a lookup service which helps the required instance of a component to be located by its client. Once located, the component is referenced by the client, thus setting up the configuration of components in that architecture. This technique is used in EJB as well as JMX. JavaBeans also use registration of listeners as a variation of this technique to locate the clients of their events. The registration-based technique results in more flexible systems than the instantiation-based technique in that run time changes to a system are handled more naturally via the lookup service. Of course, this flexibility is achieved at the expense of performance, which is impacted by the lookup service.

Support for Architectural Styles in JavaBeans and Related Technologies

JavaBeans support a publish–subscribe architectural style, in which interaction between components takes place through events, and components are producers or consumers of events. EJB defines a distributed component style, whereby two sets of interfaces are exposed, one for locating the components and another for providing business services. Both these styles are extensively used for developing consumer and business applications.

It has also been shown that JavaBeans can be adapted to support more elaborate styles such as the C2 architectural style (Taylor et al., 1996). C2 is a novel architectural style that is highly flexible and lends itself to a variety of application domains as well as to dynamically changing architectures of systems. The C2 style consists of components and connectors that interact through messages. Rosenblum and Natarajan (2000) showed how a seamless integration can be achieved between the creation of individual components and the architecture-based construction of a system to satisfy the system's requirements. They provided a C2-style aware visual builder tool for JavaBeans that can be used to develop the architecture of an application, and then to design and implement the architecture of the application. Components created using the tool are JavaBeans, and their interaction takes place through JavaBean events which form the connectors in this style. The configurations of and interactions among the JavaBeans in Rosenblum and Natarajan's system adhere to the topological rules of the C2 style.

CONCLUSION

Component-based development of software systems has shown a lot of promise in alleviating the many problems experienced by traditional software development approaches. Similarly, the Java programming language has demonstrated a number of characteristics that make it particularly useful in developing large, complex software systems distributed across the Internet. The natural solution, then, is to provide a component-based solution for the Java setting. JavaBeans are just such a solution. In tandem with a suite of accompanying technologies including EJB, AWT, JDBC, JavaMail, and JMX, JavaBeans are capable of effectively addressing numerous challenges common in software development today.

This chapter has provided an overview of JavaBeans and the suite of component-based development technologies anchored around Java. The chapter has also related these technologies to an important recent direction in software engineering—software architectures. While the software development marketplace is very dynamic and new techniques and technologies are frequently introduced, this chapter has tried to focus on the principles underlying JavaBeans and its companions. These principles reflect some of the best software development practices and are likely to persist through future technological advances.

GLOSSARY

Architecture-based software development An approach to software development where the architecture of the software system forms the blueprint for its development and evolution.

Architecture style A set of rules that describe or constrain the organization of components and connectors and the manner in which components interact.

Asynchronous A mode of interaction where the initiator continues with its processing while the target of interaction is processing the interaction request.

Container A software program or run time environment that provides necessary infrastructure facilities and services such as communication, multithreading, security, and transactional isolation to other software systems.

Deployment The process whereby software is installed into an operational environment.

Design time The state and behavior of a software component while it is being composed in a system.

Enterprise application An application that comprises an enterprise's systems for handling companywide information.

Event An occurrence of an incident of interest to a software system.

Listener A software component that receives and processes events.

Middleware Software used to provide a standard interface to low-level operating systems and network services and a run time environment for deploying components.

Naming pattern A naming convention for classes, fields, or methods followed in order to enable the systematic discovery of functionality.

Notification The communication of an event of interest to a listener.

Observer A software component that expresses interest in one or more events.

Property An attribute of a software component that can be read and/or modified.

Run time The state and behavior of a software component when the component is executing.

Software architecture The overall design of a software system expressed in terms of coarse-grained computational and data components, connectors used for interaction between the components, and their configurations.

Software component An element of a software architecture that performs processing and records state.

Software connector An element of a software architecture that performs communication, coordination, facilitation, and conversion services between interacting software components.

Synchronous A mode of interaction where the initiator pauses processing while the target of interaction performs processing.

Wizard A tool that simplifies software development by guiding a designer through a well-defined sequence of steps.

CROSS REFERENCES

See *Java Server Pages (JSP); Middleware.*

REFERENCES

Brockschmidt, K. (1994). *Inside OLE 2.* Redmond: Microsoft Press.

DeMichiel, L. (Ed.). (2002). *Enterprise JavaBeans^{TM} specification, version 2.1.* Palo Alto: Sun Microsystems.

Ellis, J., & Ho, L. (Eds.). (2002). *JDBC 3.0 specification.* Palo Alto: Sun Microsystems.

Englander, R. (1997). *Developing JavaBeans.* Sebastopol: O'Reilly & Associates.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* Reading: Addison–Wesley.

Geary, D. M. (1999). *Graphic Java 2: Mastering the JFC* (Vol. II). Palo Alto: Sun Microsystems.

Hamilton, K. (Ed.). (1997). *JavaBeans API specification, version 1.01.* Palo Alto: Sun Microsystems.

Medvidovic, N., & Taylor R. N. (2000). A classification and comparison framework for software architecture description languages. *IEEE Transactions on Software Engineering, 26*(1). Piscataway, NJ: IEEE.

Mehta, N. R., Medvidovic, N., & Phadke, S. (2000). Towards a taxonomy of software connectors. *Proceedings of 21st International Conference on Software Engineering.* New York: ACM Press.

Orfali, R., Harkey, D., & Edwards, J. (1996). *The essential distributed objects survival guide.* New York: Wiley.

Perry, D. E., & Wolf, A. L. (1992). Foundations for the study of software architectures. *ACM SIGSOFT Software Engineering Notes.* New York: ACM Press.

Richter, J., & Richter, J. (2002). *Applied Microsoft .NET Framework Programming.* Redmond, WA: Microsoft.

Rosenblum, D., & Natarajan, R. (2000). Supporting architectural concerns in component interoperability standards. *IEE Proceedings on Software.* London: IEE.

Sessions, R. (1997). *COM and DCOM: Microsoft's vision for distributed objects.* New York: Wiley.

Shannon, B. (Ed.). (2002). *Java2 enterprise edition, version 1.4.* Palo Alto, CA: Sun Microsystems.

Shaw, M., & Garlan, D. (1996). *Software architecture: Perspectives on an emerging discipline.* Upper Saddle River, NJ: Prentice-Hall.

Sun Microsystems (2000). *JavaMail API design specification, version 1.2.* Palo Alto, CA: Author.

Sun Microsystems (2001). *Java remote method invocation specification.* Retrieved January 19, 2003, from http://java.sun.com/j2se/1.4/docs/guide/rmi/spec/rmiTOC.html

Sun Microsystems (2002a). *Java foundation classes: Cross-platform GUIs & graphics.* Retrieved November 13, 2002, from http://java.sun.com/products/jfc/index.html

Sun Microsystems (2002b). *Java management extensions instrumentation and agent specification, version 1.1.* Palo Alto, CA: Author.

Taylor, R. N., Medvidovic, N., Anderson, K. M., Whitehead, E. J., Robbins, J. E., Nies, K. A., et al. (1996). A component and message-based architectural style for GUI software. *IEEE Transactions on Software Engineering, 22*(6). Piscataway, NJ: IEEE.

JavaScript

Constantine Roussos, *Lynchburg College*

INTRODUCTION

JavaScript is most commonly used to enhance the functionality of Web pages. It enables the Web page developer to create a variety of dynamic responses to user actions. These responses include changing colors, moving images, performing computations, and making components appear and disappear. The JavaScript language has evolved with the World Wide Web and has established itself as a popular and important technology. It will continue to evolve in response to increasing user demand for greater Web functionality.

THE INTERNET, HTML, AND THE NEED FOR A SCRIPTING LANGUAGE

Prior to 1995 HTML documents were mostly plain text in paragraph format. In response to user demand for more functionality in displaying Web pages, in April 1995 Netscape Version 1.1 added tables and many more elements and attributes. Users, however, were accustomed to programs on their individual computers that interacted with them in dynamic ways. Web pages could do this in only the most minimal fashion. At the core of the problem was HTML. Based upon a commercial art markup paradigm, it was not designed to be a programming language. HTML "tags" specify where text and graphic images are to be placed, how large they should be, and what their relationship is to other elements in a document. HTML did not possess the features of the programming languages that were used to create the sophisticated applications of the day. In order to meet the demand for more dynamic interactivity Web pages needed some kind of embedded programming language.

Programming the Web—Java

By 1995 the World Wide Web was ubiquitous and the need for a supporting programming language to manage Web content was painfully evident. The Java language, a precursor to JavaScript, was designed in the summer of 1992. Given the widespread acceptance of the C++ programming language it was no surprise that Java employed C++ syntax and its object-based approach to programming. Java's first commercial application was to control an interactive, handheld home-entertainment device controller with an animated touchscreen user interface. On May 23, 1995 John Gage of Sun Microsystems and Marc Andreessen of NetscapeCommunications announced that Java programming language support would be incorporated into Netscape Navigator. In October 1995 Netscape Navigator version 2.0B1 included support for Java applets, small Java programs that ran within a browser window. The days of purely static Web pages had ended. Java incorporated methods to make Web pages dynamic in numerous ways. Java applets could run almost independently in their own windows using the Web page as an enclosure. HTML elements could be modified with Java code, and Java could manage add-in components (plug-ins) that users could download. Except for restrictions due to security concerns the Web programming environment was now nearly as robust as that of programs running locally on a computer.

Unfortunately, Java did not provide solutions for all needs. Java applets were often slow. Many users were reluctant to download add-in components due to security concerns, inconvenience, and lengthy download times. Furthermore, Java was more complicated to use than simply creating standard HTML documents. Nonprogrammers were now creating Web pages and many who were able to use HTML found Java difficult to master. Java appeared to be more of a tool for professional programmers. Also, even programmers were not entirely satisfied with the additional time required to create Java applets to perform relatively simple tasks like determining whether a user had filled an input box on a Web page. There continued to be a demand for a simpler means of creating and controlling sophisticated Web content. This demand set the stage for the development of a scripting language specifically designed for the Web page environment. A scripting language is a type of programming language that can be utilized without much of the complexity associated with traditional "full-strength" programming languages. Scripting languages had been utilized by programmers for many years for numerous purposes. Even many technically competent nonprogrammers had discovered that they could write useful code using a scripting language since its use does not require knowledge of compilers, linkers, object libraries, or most of the other technical roadblocks to writing computer programs.

HISTORY OF JAVASCRIPT

The JavaScript Scripting Language

Java was designed to be a complete programming language, and Java applets were typically created by professional programmers. The popularity of the Web derived in large part from the fact that nonprogrammers were able to deliver Web content. The introduction of Java brought forth a new level of Web page functionality that was not accessible to most nonprogrammers. Furthermore, much of what Java enabled programmers to do was quite simple in concept. For example, those using HTML saw no reason why operations like changing colors and images, modifying font size, and performing calculations interactively should require a sophisticated programming language. Netscape concluded that what was needed was a relatively simple language, embeddable in Web pages, that could perform these tasks and also enable the Web page designer to easily incorporate Java applets written by programmers. In early 1995 Netscape hired Brendan Eich to take charge of the design and implementation of such a new language. Eich decided that a loosely typed scripting language suited the environment and the audience. Since the purpose of the language was to bring life to static Web pages it was dubbed "LiveScript." For the same reasons that Java employed C++ syntax and principles so did LiveScript. Due to marketing considerations the language was soon renamed JavaScript. On December 4, 1995 Netscape and Sun jointly announced the new language in Netscape Navigator version 2.0B3, calling it a "complement" to both HTML and Java.

Java vs. JavaScript

The similarity of the names still causes confusion, as the perception is that there is a strong link between the two languages. There is not. The common C++ syntax shared by the two languages belies the significant differences in purpose, principals, and implementation.

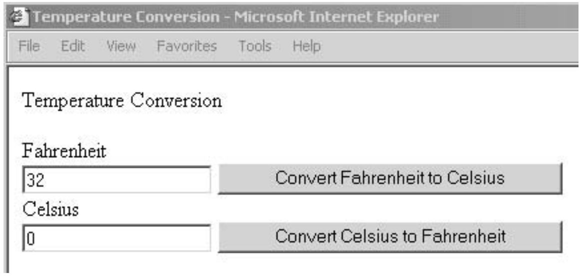
Java is a fully functional object-oriented programming language. Java programs may be run at the command line as most other languages. Java programs may also be written as "applets." These applets run in a window embedded in a Web page and have little direct access to Web page elements. JavaScript may only run embedded in an environment (normally a Web page) but was designed to directly interact with Web page elements. Java is often utilized by programmers for its graphics capabilities. JavaScript has no inherent graphic capabilities. It may, however, utilize the graphic capabilities of the environment in which it is running (the Web page). JavaScript code only exists as plain text. Java is compiled to a lower-level language that is interpreted by the Java Virtual Machine. This means that Java source code is not normally viewable to the user but that JavaScript code is.

Despite their significant differences Java and JavaScript have one important similarity. The driving force behind their popularity is that each is commonly and successfully used to enhance the functionality of Web pages.

Listing 1: JavaScript temperature conversion.

```
<HTML>
<HEAD><TITLE>Temperature Conversion</TITLE>
  </HEAD>
<BODY>
Temperature Conversion
<form name=TempConv>
Fahrenheit<BR>
    <input type=text name=Fahrenheit value=
      32 >
    <input type=button name=ConvertF2C
      value="Convert Fahrenheit to Celsius"
        onClick="Celsius.value = Math.round
          ((Fahrenheit.value - 32)*5/9);">
          <BR>
Celsius<BR>
    <input type=text name=Celsius value=0 >
    <input type=button name=ConvertC2F
      value=
      "Convert Celsius to Fahrenheit"
        onClick="Fahrenheit.value = Math.
        round (Celsius.value * 9/5 + 32);
        "><BR>
</form>
</BODY>
</HTML>
```

In Figure 1, we see a simple Web page that enables the user to convert from Fahrenheit temperatures to Celsius


Temperature Conversion - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Convert Fahrenheit to Celsius
Convert Celsius to Fahrenheit

Temperature Conversion

Fahrenheit

32

Celsius

0

Figure 1: Temperature conversion Web page.

and vice versa. We also see the HTML and JavaScript code that created the Web page in Listing 1. The user of this Web page simply types a temperature into either text box and clicks on the appropriate button to perform the temperature conversion and write the computed value into the other text box.

Understanding the Code. The first line, <HTML>, tells the browser that this is an HTML document. The second line creates a title for the document and the third line defines the beginning of the body of the HTML code.

The next line of code simply displays the text "Temperature Conversion" <form name = TempConv> creates a form named TempConv into which the two text boxes and two command buttons are placed.

The first line beginning with <input creates a text box named Fahrenheit to hold a Fahrenheit temperature. The user may enter whatever numeric value he/she wishes into this text box.

The next line creates a button named ConvertF2C that the user should click on to convert Fahrenheit to Celsius.

Up to this point all of the code has been pure HTML.

In the next line of code onClick = is followed by JavaScript code (in quotes) that is executed when the user clicks on the ConvertF2C button. The action of clicking on the button is referred to as an event. The code executed in response to this event is called an event handler. The formula embedded in the code is the standard formula for converting Fahrenheit to Celsius. Fahrenheit.value represents the value in the Fahrenheit textbox and the round method simply rounds a number to the nearest whole number (integer). In this case the whole number represents degrees in Celsius. Celsius.value = indicates that the new computed value is now to be placed into the Celsius text box, which has been created with another <input statement. The final <input statement creates the button that converts Celsius to Fahrenheit.

The JavaScript code (that enclosed in quotes after onClick =) may be immediately preceded by JavaScript: to identify it as JavaScript code. However, the browser assumes that the statements after onClick = are JavaScript from the context and from the fact that JavaScript is the default scripting language for HTML.

As one can readily see, the Web page functionality described above is made possible through the interaction between HTML and JavaScript. Dynamic HTML or DHTML is sometimes defined to be "the ability of the browser to alter a Web page *after* the document has loaded." By this definition the above is an example of DHTML. Some require greater activity such as animation or pop-up windows before the DHTML label is applied. More information on and examples of DHTML are given in the chapter on DHTML.

JavaScript's Success

JavaScript became an immediate success because of its ease of use and its ability to increase Web page functionality and visual appeal. Its popularity continues to grow. Microsoft responded to the Sun/Netscape alliance and the popularity of JavaScript in a two-pronged strategy. In May 1996, with the release of Internet Explorer 3.0B1, Microsoft introduced its own Web scripting language called VBScript, which was based upon Microsoft's popular Visual Basic programming language. At about the same time Microsoft also introduced its own version of JavaScript called "JScript." At that time, however, Netscape was the dominant Web browser and would not execute VBScript code. JavaScript managed to stay one step ahead by adding significant additional features. With the final release of Internet Explorer 3.0 in August 1996 the browser wars began in earnest. Incompatibilities between the two languages and their interaction with browser environments were problematic for programmers and users alike.

Standards for JavaScript

Eventually programmers and users brought about significant pressure for standardization. The European Computer Manufacturers Association (ECMA) created ECMAScript, a standard for the core of Web scripting languages (ECMA, n.d.). The standardization effort began in November 1996 and a standard was adopted in June 1997 by ECMA and by the International Standards Organization (ISO) in April 1998. Although the core of the language was now standardized, the environments in which it lived (e.g., the Web browser) were not. Microsoft, Netscape, and dozens of other companies worked with the Word Wide Web Consortium (W3C) to try to lay the groundwork for a truly universal document object model (DOM) that would be a compromise between an ideal standard and one that would be backward compatible with existing technologies to protect the billions of dollars invested in existing scripts. The document object model is a standard managed by the W3C. It defines the logical structure of HTML documents. More information about the DOM is given below.

ADVANTAGES AND DISADVANTAGES OF JAVASCRIPT

The Popularity of JavaScript

Probably the single most important reason for JavaScript's, now immense, popularity is that its ease of use allows the greatest number of individuals to productively utilize this product. Just as the numbers of computers sold skyrocketed when IBM brought computing to the masses with the IBM-PC in 1981, so HTML and JavaScript have now enabled millions of individuals to produce the most popular form of computer-based, information presentation: Web pages. Although designed to be easy to use, JavaScript has the functionality to be employed in very complex and powerful ways in the hands of professional programmers. The scripting language has established itself as a core technology of the Web and JavaScript users now span the range from personal homepage developers to Fortune 500 developers.

Low "Cost of Entry"

The term "cost" here refers both to money and time. The JavaScript interpreter comes with the Web browser so there is no additional cost to the programmer to write and run JavaScript code. JavaScript source code and executable code are the same so no development environment or compiler is required. An ordinary text editor is all one actually needs to produce executable JavaScript code.

JavaScript Functionality

Due to the fact that JavaScript programs download quickly and are interpreted and executed on the client computer, Web pages utilizing JavaScript are much more responsive than Web pages that only utilize Java applets or computer programs that must execute on the Web server. Even in cases where JavaScript cannot perform all of the programming tasks required in an application it is often able to significantly reduce the amount of processing that must take place on the Web server. JavaScript's functionality has steadily increased as the language has developed. For example, add-in components (plug-ins) may now be detected and loaded with JavaScript. Due to its pragmatic approach it has incorporated many of the best features of several popular programming languages.

Although JavaScript has many powerful features that expert programmers might utilize, the language was designed to be easy to use for beginning programmers and even nonprogrammers. A search of the Web will reveal numerous Web sites promoting literally millions of lines of JavaScript code that may simply be copied into a Web page. This code is often packaged nicely within functions with an explanation of the purpose of each function. Additionally, JavaScript tutorials, reference materials, and commentary abound on the Web. A Web search or walk through the local bookstore will reveal the fact that more than one hundred books on JavaScript are now available. With the establishment of standards such as ECMA-262 and DOM, JavaScript is now more portable than ever.

What JavaScript Can Do

Dynamically Manipulate Web Page Content

JavaScript, as a scripting language embedded in HTML documents, is a means of dynamically controlling Web content. JavaScript can be used to modify almost any attribute of any object or element of the Web browser. For example, JavaScript can be used to change the color, size, and position of HTML elements in response to user events such as moving the mouse or pressing a key. These elements may include images, tables, input boxes, and blocks of text. It can be used to make elements visible or invisible. It can cause the browser to navigate to another Web page or to move forward or back in its history list. It can even create new browser windows with content that JavaScript defines. That is, JavaScript can dynamically create HTML to be executed by the browser.

The ability of JavaScript to dynamically change Web content is referred to as dynamic HTML. Some of the most popular dynamic uses of JavaScript are the following:

Create crude animation by loading images sequentially to simulate movement similar to how the sequential individual frames of a movie tape simulate movement. JavaScript may also change the position of individual images in the document window.

Cause text to change color and size or cause images to change in response to mouse movements to indicate the action that may be invoked by clicking the mouse over a particular object.

Cause text to appear or disappear in order to make information available to the user or to reveal additional links to other documents when the mouse travels over a specific area on the screen. This facility is quite useful in producing drop-down menus.

Additionally, JavaScript is adept at creating and managing the content of frames in Web pages. JavaScript's ability to modify style attributes in cascading style sheets (CSS) accounts for much of its power to modify Web page elements.

For example, the following JavaScript code causes the background color of a text box named *area* to turn yellow. The input box is contained in a form named CircArea:

```
<script>javascript:CircArea.area.style.
  backgroundColor='yellow';</script>
```

Validation of Data. JavaScript is particularly powerful and useful when used in conjunction with forms. JavaScript can determine the content of text boxes, which radio buttons are turned on, which check boxes are checked, and what item(s) the user has selected from a list of choices from a list box. It can then check the data it has acquired for correctness, accuracy, and consistency and communicate its finding to the user.

Computation. JavaScript can perform arithmetic calculations using data it has acquired from the user and other sources such as internal program code. For example, JavaScript can compute the total cost of an order based upon the cost and quantity of the items selected, sales tax rates for various states, shipping costs for different destinations and item weights, and other factors without having to access information outside the HTML page. Numerous general and special purpose calculators have been programmed entirely in JavaScript.

Example: Data Validation and Computation. The following JavaScript function, ComputeArea, returns the area of a circle if its argument, radius, is a valid floating point number and returns the string "Invalid radius" if the radius is not valid:

```
function ComputeArea(radius) {
       if (parseFloat(radius)) return Math.
         PI*radius*radius;
       else return "Invalid radius";
}
```

Use Methods

JavaScript may also invoke methods that have been defined for HTML objects and elements. In simplest terms a method is an operation that an object can perform. The method may take parameters. One example of a method is displaying the alert pop-up dialog boxes we often see. The browser window is the object that displays the alert pop-up dialog box. The text displayed in the box is the parameter of the alert method. A JavaScript statement to display an alert box containing the message "hello" is the following:

```
window.alert('hello');
```

Note that the format of the above statement is the *object name* (window), followed by a *dot* followed by the *method name* (alert), followed by *parameters in parentheses* ('hello').

Additionally JavaScript can invoke and, to some degree, control other components that have been loaded into the browser. These include Java applets and plug-ins such as controls that play music or display video. (See the example under Control of Java and Plugin Components.)

Respond to Events

When JavaScript programs change the content of a Web page they normally do so in response to user-initiated events. Events typically utilized by JavaScript include the following:

Clicking on a button;

Passing the mouse over an element such as a link or image; and

Leaving or entering a Web page.

Communicate with the Server and Other Web Pages

JavaScript can dynamically create and pass information to programs that reside on the Web server. It does this through the use of forms and query strings. Web pages that contain forms can use JavaScript to cause the contents of the elements of a form to be passed to a program on the server. Although ordinary HTML can also cause form data to be passed to the server via the SUBMIT button element, JavaScript has much more robust capabilities. For example, JavaScript can analyze data entered by the user, perform computations using the data, and determine the appropriate server to which to send the data. Additionally, a programmer can send information that may not be contained in form elements to the server by using a query string. The query string normally consists of the address or URL (uniform resource locator) of a program on the server followed by the question mark character followed by the information to be sent to the server. This same method may be used to send information to other Web pages that may be invoked or even created with JavaScript.

Example. The following line of JavaScript code loads a Web page into the browser window and passes it information:

```
window.location=http://GlobalU.edu/
  admissions/Info.htm?frosh=500
```

Understanding the Code. The destination Web page is located on the main Web server at GlobalU.edu. It is in a folder named admissions. The Web page is named Info.htm. The information being passed to Info.htm is named "frosh" and its value is 500.

Save Server-Side Communication and Processing Time

Delays in the presentation of Web page information are usually due to the slowness of Web browser/Web server communications. Some operations, however, such as changing the color of a Web page link as the mouse passes over it appear to happen almost instantaneously. The reason is that the code to perform that operation is executed locally on the client machine. No communication with the Web server is required. JavaScript can be used to reduce communication with the server in a number of ways. For example, a Web page that contains a registration form typically contains some required fields like, perhaps, the registrant's name. If the form is submitted to the server with a blank name field the server will have to send back a Web page requesting the user to fill in the name field before the form can be processed. This communication with the server can be eliminated. When the user clicks on the submit button JavaScript can determine whether all of the required fields have been properly completed. If they have not, then it can notify the user. This requires no communication with the server. If JavaScript determines that all required fields have been properly completed, it can then submit the form information to the Web server.

JavaScript Shortcomings

Several of the items listed below under the category of "shortcomings" are intentional restrictions on the JavaScript language. They are listed because the missing functionality is present in many other languages.

Because it is an interpreted language JavaScript program source code cannot be effectively hidden from the user or other programmers. JavaScript has no object libraries although source libraries may be used equally effectively for some purposes. Additionally, interpreted languages typically run much more slowly than compiled languages and JavaScript is no exception.

JavaScript is a loosely typed language. The JavaScript interpreter often makes decisions about how to interpret the content of variables depending upon the context. This fact makes it easier for novices to use the language but can occasionally hide subtle and serious programming errors. Thus, in critical applications where errors may result in dire consequences JavaScript would probably not be the language of choice.

Primarily due to the fact that JavaScript executes within a host application it does not have its own development environment or debugger. It is dependent upon the host application to provide these. Various vendors including Netscape (Netscape JavaScript Debugger) and Microsoft (Visual Studio Interdev) provide such environments but, typically, they are not as comprehensive as those supplied for C++, Visual Basic, and other languages.

Client-side JavaScript executes solely on the client computer. It has no access to server resources such as databases, files, or information about the client that may be stored on the server. JavaScript must rely on programs located and run on the server to retrieve such information. This protects the server from being compromised by malicious JavaScript code. Only programs that reside on the server may access server resources. These programs may validate the source of requests for data if needed.

For technical and practical reasons, JavaScript has no direct access to the underlying hardware (such as registers or I/O ports) of the computer on which it is executing. Contrast this to C and C++ where the programmer potentially has almost complete control. JavaScript cannot be used for "real-time" applications such as monitoring sensors and controlling switches. Without this restriction the possibility would exist for malicious JavaScript code to damage the client machine through actions like reformatting the hard drive or overwriting the system BIOS.

ROLES AND APPLICATIONS OF JAVASCRIPT IN THE INTERNET AND E-COMMERCE

The World Wide Web has proven itself to be important if not essential for companies to present and market their wares. Consumers today look for the comparative shopping information they need on the Web. Shoppers expect to be able to browse through catalogs, be advised of current sales and place orders—all online. JavaScript has and will continue to play a significant role in this environment. Because JavaScript executes much faster than server-based programs programmers try to place as much functionality in the JavaScript code as possible. Although entire e-commerce applications have been built using JavaScript the norm is that JavaScript is used as a supplement to other technologies.

JavaScript is also used extensively in intranets, Web-based networks that are internal to a company. These may be used for distributing personnel policies, coordinating company projects, and numerous other purposes. Intranet information content tends to be maintained by departments rather than by a company's IT personnel. Thus, nonprogrammers are often tasked with maintaining the information. JavaScript's ease of use makes it a good candidate for bringing greater functionality to intranet Web pages.

Finally, JavaScript's programming language capabilities together with its ability to control Java applets and plug-ins means that functionality, special to a particular application, may be incorporated into Web pages with a minimum of effort. For example, a lending institution can easily place a loan calculator on a Web page to enable potential customers to calculate monthly payments. A music store can utilize a music player control to allow potential customers to hear samples of CDs. Other technologies can be used to implement such functionality but each requires some special knowledge on the part of the programmer. Only JavaScript has demonstrated the ease of use and versatility required of a one-size-fits-all tool.

PRINCIPLES OF THE JAVASCRIPT LANGUAGE

Scripting Languages Defined

"A scripting language is a programming language that is used to manipulate, customize, and automate the facilities of an existing system" (ECMA-262 standard document, ECMA, n.d.). The scripting language is a mechanism for exposing already available functionality to program control and is intended for use by both professional and nonprofessional programmers. The existing system provides a "host environment" of objects and facilities.

A General Purpose Scripting Engine

JavaScript is a general purpose, cross-platform, object-based, scripting language. The most common use of JavaScript is, of course, as an interface to HTML objects and elements in Web pages. However, JavaScript may be embedded in any number of applications. For example, a C or Java application may be made scriptable using JavaScript just as Netscape makes its Web browser and Web server scriptable using JavaScript. To facilitate this the JavaScript engine including a JavaScript interpreter is available from http://www.mozilla.org for licensing by third parties for inclusion in their client and server products and various tools. The engine includes an application programming interface (API) for developers to expose their own objects to the JavaScript programming environment. JavaScript statements may reference these objects and elements. For example, one might consider defining objects such as paragraphs and tables in a text-processing system and enable a scripting language like JavaScript to manipulate these objects. A method called changeFont might be exposed to the scripting language. If P1 is the name of a paragraph then the statement P1.changeFont(size:12) would change the size of paragraph P1's font to 12. The core scripting language would understand the format of statements of the form Object.Method(Parameters) and would know that this requires a function call. Creation of computer code to actually record and display a change of font size would be the responsibility of the programmer implementing the embedding.

An Object-Based Language

The object-based programming paradigm has proven to be a very powerful one and JavaScript's implementation of objects is quite effective, especially for its intended environment. JavaScript supports classes (in the form of object prototypes), objects, attributes, and methods.

Because JavaScript's syntax is so similar to C++ and Java, programmers skilled in these languages generally find no difficulty in utilizing JavaScript's somewhat different implementation of objects, attributes, and methods.

JavaScript Security

The core JavaScript language has no facility for interacting with the outside world and so has no security issues. However, JavaScript is most commonly utilized as an HTML scripting language and security issues abound in the World Wide Web environment.

Inherent Security

Once the Web evolved beyond purely static HTML documents the threat of security breaches was introduced. A program executing on a user's computer has the potential to do damage or appropriate confidential information such as passwords or credit card numbers. To address these issues and gain users' confidence strict security restrictions were placed upon Java and JavaScript programs. In particular, file access was severely limited. JavaScript may write cookies and read cookies written from scripts originating from the same server. Cookies are special files that a Web server, through a script, can write to the user's disk. The location, size, and content of the cookie file are restricted. Only one cookie file per Web server is allowed on a user's system. The Web server may later, through a script, read the information it placed in the cookie earlier. Additionally, the World-Wide-Web Consortium's HTML standard contains the *type = file* input control. This allows cooperating users to upload local files to a Web server through a Web page.

With the exception of loading URLs and sending form data to a server and to e-mail addresses, JavaScript has no network capabilities and so cannot be an agent in most of the common Internet security exploitations. Still, numerous, complex, and/or subtle security holes have been found in JavaScript under certain software configurations. Known flaws have been fixed but no one can predict when others may surface. Experience has shown that diligence in employing security mechanisms can reduce the likelihood that a security hole can be exploited.

Same Origin Policy

JavaScript employs the *same origin policy*, which specifies that when loading a document from one origin a script loaded from a different origin cannot get or set certain *predefined* properties of certain browser and HTML objects in a window or frame. That is, the policy prevents a script that originated from one server and is running in a browser window from accessing potentially confidential data located in a different browser window whose content originated from a different server. The policy also applies to cookies since cookies written by one server may contain passwords or credit card information that an intruding script should not be able to access.

Security Zones

Microsoft has implemented a security mechanism called *security zones*. Briefly, security zones allow the user to group Internet sites by specified levels of trust. Unfortunately, the technology does not enable the user to specify in complete detail what privileges may be granted or revoked. Additionally, security zones are not supported by Netscape, making it difficult for programmers to utilize this security mechanism.

Signed Digital Certificates

Through Netscape, JavaScript can utilize *digital certificates* and *signed scripts* to address identification, authentication, and privacy concerns. A digital certificate is an electronic identification that the creator of a JavaScript program attaches to a signed script (program). Digital certificates employ advanced cryptographic methods to enable the recipient of a message to confirm the identity of the message sender and ensure that the message content was not altered in transit.

The *signed script* policy for JavaScript is based upon the Java security model, called *object signing*. To make use of the new policy in JavaScript, the programmer must use the new Java security classes and then sign the JavaScript scripts. Signed scripts can request, from the user, privileges and the lifting of certain security restrictions. The digital signature allows the user to identify the author of a JavaScript program. By signing the JavaScript program the signer acknowledges itself as the author and accepts responsibility for the program's actions. Unfortunately Microsoft's Internet Explorer does not support signed scripts. Ease-of-use issues for programmers and users have also played a role in preventing signed scripts from being widely employed.

Downloaded Software

Perhaps the greatest potential for security compromise is downloaded software. Once a program is running on a user's computer there is little or no protection from malicious code. Unfortunately some devious individuals are very clever about tricking users into loading software onto their systems. Some viruses, worms, etc. come disguised as e-mail messages or image files. Others may masquerade as ActiveX controls, plug-ins, or even security patches. JavaScript can be used to attempt to trick users into downloading and running such damaging code. The safe and simple rule is *do not download* unless you are sure of the origin and purpose. The rule not only applies to the Web but also to programs on CD or floppy diskette.

OVERVIEW OF JAVASCRIPT SYNTAX

JavaScript is an interpreted, high-level, object-based language with syntax similar to C and Java. As is C and Java, JavaScript is case-sensitive. It follows naming conventions similar to C and Java as well. Significant particulars about JavaScript syntax are given below.

Competing versions of JavaScript and incompatibility in the way browser objects were exposed to JavaScript fueled an outcry from programmers, vendors, and users for standardization of the core language and the browser environment. A standard (ECMA-262) for the core language was adopted in June 1997 by the ECMA and governs current implementations of JavaScript (ECMA, n.d.). In June 1998 the ECMA General Assembly approved the second edition of ECMA-262 to keep it fully aligned with the

ISO/IEC 16262 standard of Joint Technology Committee 1 of the International Standards Organization and the International Electrotechnical Commission.

Although some inconsistencies exist, nearly all current implementations of the JavaScript interpreter for the core language are ECMA-262 compliant. In order to ensure compatibility between browsers, however, browser developers must ensure that their products are compliant with the current DOM standards of the W3C as this standard affects how JavaScript references objects in an HTML document. Incompatibilities still exist in the browser environment. Both the ECMA-262 and the DOM standards have been updated over the past several years and will continue to be into the foreseeable future. Thus JavaScript interpreters must also be updated to remain conformant.

Distinguishing Features of JavaScript Syntax

JavaScript programming constructs are very similar to those of C++ and Java. For example, comments in JavaScript begin with two forward slashes (//). All text after // on a line are ignored by the JavaScript interpreter.

Listing 2: JavaScript comments.

```
//This entire line is a JavaScript comment
  and is ignored by the interpreter
x = x+1;  //Only the text past the first
  // is ignored on this line
```

We will now look at a few distinguishing features of the language.

Primitive Data Types

Primitive data types are Number, Boolean, and String. All numbers in JavaScript are floating points. Objects are complex data types. JavaScript supports built-in objects and user-defined objects. Objects are comprised of one or more components called properties or attributes that may be of any data type. Objects may also contain methods, which are functions specific to a particular object.

Literals

JavaScript supports the usual literals for primitive data types as well as special character sequences and named constants. Additionally, JavaScript has special constants such as NaN (a value that is not a number), undefined (the content of an uninitialized, declared variable), MAX_VALUE (the largest representable number), and POSITIVE_INFINITY.

Variable Declarations

Variable declarations are recommended but optional in JavaScript. One may assign a value to an undeclared variable but may not read the value of an uninitialized, undeclared variable.

Loops

JavaScript supports the common loop constructs found in C++. Additionally it supports the for-in loop construct. The following example displays the name and value of each property of the document object.

Listing 3: A for-in loop.

```
<script language=javascript>
        //write the name and value of each
          property of a document
        for (var prop in document) {
                document.write(prop + " = " +
                  document[prop] + " <BR>");
        }
</script>
```

Note that the *document* object is contained in *window* so technically we should refer to window.document. However, it is common practice to omit the reference to window.

Arrays

Arrays in JavaScript are more flexible than in most other programming languages. Array elements may contain any data type. Arrays are accessed by reference. That is, the accessing program code is given the starting location of the array elements. Individual data elements are accessed by value; the accessing program code is given a copy of the value contained in that element.

JavaScript supports several different types of array declarations.

Listing 4: JavaScript array declarations.

```
var eArray = new Array(); //declare an
  array with no elements

var uArray = new Array(12); //declare an
  array containing 12 undefined elements

//declare an array and initialize it with
  5 integers
var oddNums = new Array(1, 3, 5, 7, 9);

//declare an array and initialize with
  5 elements of differing data types
var mixedData = new Array("Hello World",
  1.5, 2, false, "bye");

//Array literals may be defined using
  square brackets

var evenNums = [0, 2, 4, 6, 8]
```

Array elements are indexed/accessed with integers beginning with 0. So, for example, the following code places the number 5 into the zeroth element of array numArray:

```
numArray[0] = 5;
```

The length of an Array in JavaScript is determined in an unusual manner. In the example declaration of uArray above, we declared an array containing 12 undefined elements so the length of uArray is 12. If we wish to increase the number of undefined elements to 25 we may simply change the length property of uArray to 25:

```
uArray.length = 25; //change the length
  of the array to 25
```

Additionally we may add a new element anywhere in the array, even past its current length:

```
//put a value in the 49th element of
  uArray, increasing its length to 50 (0, 1,
  .., 49)
  uArray[49] = "dog"
//now uArray[0] through uArray[48] contain
  undefined values.
```

Conversely we can truncate an array by setting its length to a value smaller than its current length:

```
uArray.length = 40; //array elements
  40 - 49 no longer exist
```

We may determine which values of an array are undefined by comparing each array element to the special *undefined* constant:

Listing 5: Writing the defined elements of an array.

```
for(var indx = 0; indx < uArray.length;
  indx++) {
if (uArray[indx]! = undefined) document.
  write(uArray[indx]);
}
```

JavaScript supplies the programmer with several useful array methods. These include the following:

concat()—Appends data elements and/or arrays to the end of an array and returns a new array;

sort()—Sorts the elements of an array;

join()—Creates a string containing the elements of an array;

slice()—Returns a new array that is a subarray of the specified array; and

splice()—Inserts and removes array elements of an array.

Functions and Properties of the Global Object

Functions and properties of the global object are few but useful. The global object serves as a home for special functions and properties as well as programmer-defined global variables. For example, *isNaN* and *isFinite* are functions of the global object. These functions allow the programmer to test for valid and finite numbers respectively. Another interesting and quite useful JavaScript function of the global object is *eval*, which takes a JavaScript expression or statement as a parameter, executes it, and returns the result. Other "top-level" objects such as Array, Number, and Math are considered properties of the global object. Other commonly used global functions include parseFloat() and parseInt(), which convert strings to numbers, and toString(), which converts various objects including numbers to strings. For client-side JavaScript code the window object serves as the global object.

User-Defined Objects

User-defined objects in JavaScript are similar to but defined somewhat less rigorously than Java objects. Instead of classes JavaScript utilizes object prototypes, which perform roughly the same function as classes. In fact the similarities are so great that we often refer to JavaScript classes. In keeping with JavaScript's great flexibility we can define properties and methods that pertain only to an individual object or to all objects with the same prototype (i.e., same class).

JavaScript does not support pointers as C and C++ do. Pointers commonly provide semistructured access to the memory addresses of objects. Such access can expose an opportunity for malicious code to access memory that contains data that are private to other applications. JavaScript does support the use of nested (recursively defined) objects and the null object. Judicious use of this facility can be very powerful, eliminating the need for pointers in many circumstances.

JAVASCRIPT AS A CLIENT-SIDE HTML SCRIPTING LANGUAGE

The Document Object Model (DOM) Defined

"The Document Object Model is a platform- and language-neutral interface that will allow programs and scripts to dynamically access and update the content, structure and style of documents" (Document Object Model, 2002). The DOM is a standard managed by the W3C and defines the logical structure of HTML documents. The DOM is language-independent. Its specifications allow programmers to build and modify documents. The W3C describes the DOM as a programming API for documents based on an object structure that closely resembles the structure of the documents it models. With such a standard, programs that manipulate documents should work correctly independent of the hardware or software that hosts the document. This means, for example, that a JavaScript function that changes the background color of a table in a particular HTML document should work consistently within Netscape Navigator and Internet Explorer. It should also not matter whether the Web server is Apache or Internet Information Server (IIS) or if the computer on which the document is located is running the Linux or Windows 2000 operating system. Although 100% compatibility has yet to be achieved, the DOM has greatly improved the portability of Web software from the earliest attempts at DHTML.

In recent years the DOM has grown from its *Core* to cover HTML, XML, CSS, events, and other features. Due to its increasing size and broader scope the DOM has been broken into modules to make it more manageable. Consequently most Web browsers and other software currently only fully support the DOM modules they deem essential.

Objects are, by nature, tree-structured and the DOM can be used to describe these objects down to the finest granularity. For example, there are DOM descriptions for the Web page window, for the document contained in the window, and for each individual element of the document. Although a complete DOM description of the Web page environment would be impractical to illustrate, a high-level picture of a few of the most important elements is useful (see Figure 2).

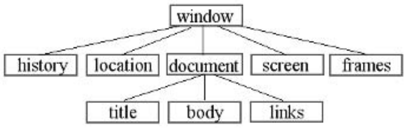


Figure 2: Some important elements of the DOM description of Web pages.

DOM—Web Page Object Hierarchy

Embedding JavaScript Code in Web Pages

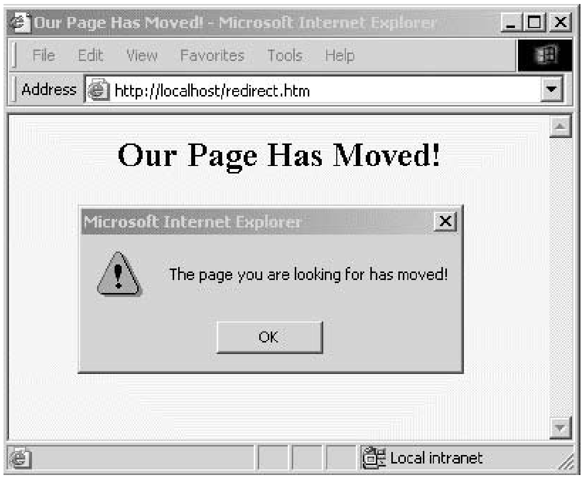
Perhaps the most straightforward means of embedding JavaScript into a Web page is to simply put it in line with ordinary HTML statements. The JavaScript must be enclosed in <SCRIPT> and </SCRIPT> tags so that the browser knows that it is JavaScript code.

Web pages are often moved from one location to another. This can be frustrating for Web surfers unless the page's author leaves a Web page in the original location to redirect them to the new page. In the following example we create a Web page that redirects the browser to a new page. (See Figure 3 and Listing 6.)

Listing 6: Redirect the browser to another Web page.

```
<HTML>
<HEAD><TITLE>Our Page Has Moved!</TITLE>
  </HEAD>
<BODY BGCOLOR="#E8F4FF" >
<CENTER><H2>Our Page Has Moved!</H2>
  </CENTER>
<script language="javascript">
        alert("The page you are looking for
          has moved!");
        window.location.href="http:
          //localhost newpage.htm";
</script>
</BODY>
</HTML>
```

Understanding the Code. The first line, <HTML>, tells the browser that this is an HTML document. The second line creates a title for the document, and the third line defines the beginning of the body of the HTML code and sets the background color to light blue.


Our Page Has Moved! - Microsoft Internet Explorer

Address http://localhost/redirect.htm

Our Page Has Moved!

Microsoft Internet Explorer

The page you are looking for has moved!

Figure 3: Web page redirect.

The fourth line of code displays the text "Our Page Has Moved" as a number 2 size header and centers it.

The <script> tag tells the browser that what follows is JavaScript code.

The first JavaScript statement displays an alert box informing the user that the page has moved. After the user clicks OK the next line of JavaScript will direct the browser to a page named newpage.htm.

The </script> tag ends the JavaScript code.

The most common means of embedding JavaScript is by creating JavaScript functions and placing them in the HEAD of the HTML document. These functions may then be called from event handlers (see Events in the DOM) or other embedded JavaScript code.

Below is an html document that contains a JavaScript function that computes the area of a circle given the radius. The function is embedded in the HEAD section of the HTML document:

Listing 7: CircArea.htm computes the area of a circle.

```
<html>
<head>
<!-- A JavaScript function to compute the
  area of a circle -->
<SCRIPT LANGUAGE = "JavaScript">
function ComputeArea(radius)
{return (Math.PI * radius * radius)}
</SCRIPT>
</head>
<body>
        <script>alert(ComputeArea(2));
          </script>
</body>
</html>
```

Understanding the Code. The ComptuteArea function is embedded in the HEAD section of the HTML document above. The function must also be enclosed in <script>, </script> tags.

The *function* key word tells the browser that a function is being defined. After the function name, enclosed in parentheses is the function's input (parameter), the radius of the circle. The function executes the standard formula for computing the area of a circle and then *returns* the answer. Note that the instructions contained in a function are enclosed in curly braces: { and }.

In the body of the HTML document is a JavaScript statement that displays an alert box containing the computed area of a circle with radius equal to 2. ComputeArea(2) causes the function to execute (perform its instructions) using the number 2 as the radius parameter.

HTML Element Control

Perhaps the most common and important use of JavaScript is to modify attributes of HTML objects and elements. These objects and elements include tables, images, forms, frames, the current Web page, and the status bar to name a very few. Attributes they possess that may be programmatically controlled include color, text content, size, and position. JavaScript may also be used to send to the server data that users have entered into forms. It is also commonly used to store and retrieve data stored in cookie files.

Events in the DOM

JavaScript functions that control HTML objects and elements as well as plug-ins are usually activated by events. In the context of Web pages events may be defined as actions detectable by the Web browser. Examples of events include clicking on a button, moving the mouse over an element, navigating to a new page, or pressing a key on the keyboard. For each event to which a programmer wishes to respond, he/she can define a JavaScript program that executes whenever the event occurs. Such a program is called an event handler.

Example—Event Handling

In this example we define a table element called CoursePB that contains the text "Course" and three event handlers. The event handlers are simply single JavaScript statements. When the user clicks on element CoursePB the browser navigates to a page named course.htm. When the mouse is moved over the table element the word Course changes color to red and when the mouse is moved off the element the color of the word Course is changed to black.

```
<TD ID=CoursePB
        onclick=javascript:document.
          location="course.htm"
        onmouseover=javascript:style.color=
          "RED"
        onmouseout=javascript:style.color=
          "BLACK">
        Course
</TD>
```

Control of Java and Plug-in Components

JavaScript may be used to detect, define, and load Java applets and plug-ins. Plug-ins are software components that allow the processing of audio, video, and other data not directly supported by the browser. For example, JavaScript can detect whether the RealPlayer plug-in has been loaded onto your computer. If not, it can load the plug-in to play a streaming audio file from the Web.

JavaScript may also exercise limited control over such plug-ins. For example, it can start and stop the playing of the streaming audio in response to user action. Similarly, JavaScript can load, start, and stop Java applets.

External Scripts

Full-strength programming languages have sophisticated code-management capabilities. Due to the fact that JavaScript is interpreted and that it executes within the context of a host application (most commonly a Web browser) its code management features are minimal. The Web browser does, however, support the use of included files, a simple, yet effective, form of code management.

Skilled programmers write general-purpose functions that can be reused in many different Web pages. A JavaScript function that determines whether a required input box on a form has been completed and notifies the user if it has not is an example. Another similarly useful function is the example below. Function isBadNumber takes one parameter, a text box, and returns true if the content of the text box is not a valid number; otherwise, isBadNumber returns false.

```
function isBadNum(textBox) {
       if (isNaN(textBox.value) || (textBox.
         value == ")) return true;
       else return false;
}//end function isBadNum
```

Copying such a function into every Web page in which it is used can be very inconvenient for the programmer, make the page unwieldy, and waste resources. Additionally, as the programmer, over time, makes improvements to the function he/she will find that many older and less effective versions of the function still exist. Retaining a single copy of the function and having the newest version accessible to all Web pages that use it is preferable.

The following line of code when placed in an HTML file causes the JavaScript functions contained in the file named CalFuncs.js to become accessible to the HTML document:

```
<Script language=JavaScript src="CalFuncs.
  js"></Script>
```

JavaScript functions stored external to the Web page in which they are used are often referred to as *external scripts*. Only one copy of such functions need exist. When the function is updated, the most recent version is used in every Web page that calls that function the next time the Web page is accessed. Programmers do have to take care that the updated function will work in every previously created Web page that uses it.

JavaScript I/O

In most languages there are three separate types of I/O (input/output):

Keyboard/mouse input and monitor/printer output to interact with the user;

File I/O that reads and writes files to disk; and

Network I/O that connects to and communicates with other computers.

Core JavaScript has no I/O capabilities. However, client-side JavaScript, through interaction with a Web browser, has a minimal set of I/O capabilities.

File and Network I/O

File and Network I/O in client-side JavaScript is severely restricted due to security concerns. JavaScript may write and read cookies and, by using the *file*-type HTML input control, initiate file uploads with user permission. With the exception of loading URLs and sending form data to a server and to e-mail addresses, JavaScript has no network I/O capabilities. (See the section JavaScript Security.)

Interactive I/O through Dialog Boxes

JavaScript can interact with the user in two ways. The first is through the use of the window methods, alert, confirm, and prompt, all of which present dialog boxes to the user with increasing functionality.

Alert. The purpose of the *alert* method is to present a message to the user. After the alert dialog box containing the message is displayed, the user must press the OK button to clear the dialog box and resume the program.

Confirm. The *confirm* method presents a query to the user together with an OK and a Cancel button in a dialog box. The user must click one of the two buttons to clear the dialog box. The confirm function returns either true or false, depending upon whether the user clicked on OK or Cancel.

Prompt. The *prompt* method presents to the user a query, a text input box, and two buttons labeled OK and Cancel in a dialog box. The text input box may be blank or contain a default response. The user may type into the input box and must finally click on OK or Cancel to clear the dialog box and resume the program. Clicking OK will pass the data typed into the input box to the program; clicking Cancel or closing the prompt box will return null to the program.

Interactive I/O through Web Pages

JavaScript may also communicate with the user through the Web page that hosts the script. JavaScript may write directly to the Web page using the *write method* of the *document object*.

For example, the JavaScript statement

```
document.write('Today is ' + Date());
```

writes the host computer's system date to the Web page.

JavaScript may also use HTML elements to communicate with the user. For example, JavaScript may write to a text input box to communicate to the user and it may read what the user has typed into a text input box to receive communication from the user.

SERVER-SIDE SCRIPTING AND CGI

Server-Side JavaScript

Core JavaScript is a general purpose scripting language that can be embedded in most applications. Netscape's Web servers include a JavaScript interpreter, which makes possible the use of JavaScript as a server-side scripting language on these Web servers. Server-side JavaScript is composed of core JavaScript and additional objects and functions for accessing databases and file systems, sending e-mail, etc. Microsoft supports the use of JScript as a server-side scripting language. Although having a nearly identical language available on the Web server and the client is a convenience for programmers, server-side JavaScript has not enjoyed great popularity. Instead several other technologies are commonly employed to interact with the server to access databases and other resources on the server. These technologies include ASP, Perl, and PHP. Client-side JavaScript programs may interact with these technologies.

ASP and VBScript from Microsoft

Microsoft's Web server, IIS, supports a number of technologies including the Microsoft-specific ASP (Active Server Pages) and VBScript. ASP files are programs that run through IIS on the server, are able to access databases, files, and other resources, and communicate with the browser on the client computer. VBScript is a client- and server-side scripting language. It is used most extensively for programming ASP files. After collecting information from the server the ASP program may send information and client-side programming code to the client in the form of a Web page. Typically client-side programs are written in JavaScript. JavaScript may also be used to invoke ASP programs and send information, such as the contents of forms, to an ASP program.

PHP

PHP is a widely used, general-purpose scripting language especially well suited for Web development. PHP is a project of the Apache Software Foundation. JavaScript code can be created using PHP code just as in ASP programs. As with ASP, PHP has access to server resources such as databases and environment variables. For example, the environment variable $HTTP_USER_AGENT holds the name of the Web browser the client is using to display the Web page. PHP files have an extension of .php. <?php denotes the beginning of a PHP script and ?> denotes the end. Note that the abbreviated <? may also be used to denote the beginning of a PHP script. The echo statement is used to display data.

Example

The program PHPTest.php below writes a Web page containing the name of the Web browser (e.g., Microsoft Internet Explorer 6.0), the name of the Web server, and the system date on the client computer.

```
<html><head><title>PHP Test</title></head>
<body>
<?
echo $HTTP_USER_AGENT, "<BR>";
echo $SERVER_NAME, "<BR>";
echo "<script language=\"JavaScript\
  ">\n"
echo "document.write('Today is ' + Date());
  \n";
echo "</script>";
?>
</body></html>
```

Perl

Perl is an acronym for "practical extraction and report language." Created by Larry Wall, it is an interpreted language optimized for string manipulation, I/O, and system tasks. It is widely used by Web programmers due to its functionality and flexibility. It is able to access databases and files on a server and send information back to the client via a Web page as PHP and ASP do. The Perl interpreter is located on the server. Perl scripts are executed by running the Perl interpreter through the CGI (common gateway interface—see below) and passing to the interpreter the name of the file containing the Perl script.

When executing a Perl script from the operating system command line one simply invokes the Perl interpreter and passes it the name of the Perl script with any parameters.

For example, the command

```
perl myscript.pl
```

will cause the Perl interpreter to execute the Perl program named myscript.pl.

In order to run a Perl script from a Web page the script is normally specified as the action object of a form. That is, the Perl script will be run when the form is submitted and the form data will be passed to the script. The example below illustrates this:

```
<FORM METHOD=POST ACTION="/cgi-bin/
  myscript.pl">
    <!-- Input data elements -->
    <INPUT TYPE=SUBMIT VALUE="Submit Data">
</FORM>
```

Common Gateway Interface (CGI)

The common gateway interface (CGI) is not a programming language. It is a standard for interfacing external applications such as Web browsers to information servers such as Web servers. Most programming languages that run on a server can utilize CGI to implement two-way communication with Web browsers. CGI programs implement the CGI standard, run on the Web server, and, potentially, have access to all resources located on the server such as databases, files, and server-based environment variables. CGI programs are normally executed in response to requests from an HTML document (Web page). Normally, CGI programs retrieve information from the Web server and package that information into an HTML document that is then displayed on the client browser. The Web page created by the CGI program may contain JavaScript that can execute when the Web page is loaded into the browser or in response to events such as the user clicking on a button element or moving the mouse over an image.

THE FUTURE OF JAVASCRIPT

A Core Web Technology

JavaScript has become a core Web technology. It is used by programmers and nonprogrammers alike. Millions of Web sites rely on it. These facts alone ensure that it will remain viable for a number of years. Additionally, JavaScript has evolved to meet the ever-increasing demands of Web programmers. By all indications it will continue to evolve, improve, and become more powerful. As JavaScript gains more sophisticated programming capabilities in the future, developers will be able to write full-featured applications in JavaScript. It has demonstrated its resiliency by weathering bad publicity over a number of security breaches and has made some progress in developing standards to address security issues, although security will continue to be a concern. The establishment of the ECMA-262 standard has addressed browser incompatibility and future development issues. The ECMA technical committee is currently working on significant enhancements to the specifications and is also working with other standards groups to coordinate the development of the language. More programmers now code to the ECMA and DOM standards, thus putting pressure on browser manufacturers to ensure that their JavaScript interpreters conform to the standards.

New Technologies

As rosy as the future looks for JavaScript, there are no guarantees. Netscape's JavaScript interpreters are now "open source" hosted by mozilla.org, and many products have faltered when ownership appeared to be in question. Microsoft now owns the bulk of the browser market, and it has a reputation for going its own way. New technologies spring up regularly in the Internet world and these could have an impact on the viability of any current product. Microsoft has long promoted VBScript without great success but the expected popularity of Microsoft's .NET technology raises new questions for the future of Web-based programming. JavaScript interpreters must be continually updated to keep current with changing standards. Faced with competing technologies those standards may need to change rapidly. Keeping "free" software up to date in this environment may be a challenge.

Corporate Pacts

Agreements between major vendors are commonplace and have an impact on browser and other product development. On November 24, 1998, AOL announced that it had purchased Netscape (the creator of JavaScript) and consummated a separate but almost simultaneous strategic development agreement with Sun Corp. (creator of Java). The final impact of this has yet to be felt. Finally, it bears note that the World Wide Web is less than 10 years old. Ten years ago no one could have predicted the present state of network computing. Its state 10 years hence may be just as unknowable.

GLOSSARY

Attribute A component of an object (same as a property).

Cookie A computer file that may be written to or read from a specially designated location on a user's disk drive by a Web browser. The cookie typically contains information that a Web application will use at a future time.

Compiler A computer program that reads programming language statements, translates them into instructions that the computer hardware can execute, and writes the translated instructions to a file for later execution.

Debugger A computer program that aids a programmer to locate errors in program code.

Element A component of a Web page such as a text box, table, link, or command button.

Event An action, such as a mouse click, that can be detected by a computer program.

Function A sequence of programming language instructions executed in response to a call.

Hypertext markup language (HTML) A language used to specify the content of Web pages.

Hypertext transport protocol (HTTP) A protocol that specifies how Web page information is communicated over a network.

Integrated development environment (IDE) A computer program that provides a programmer-friendly environment in which a programmer can write, test, and debug computer code.

Interpreter A computer program that reads programming language statements and causes them to be executed while the interpreter is running.

Method A computer language function specific to an object. See function.

Object A complex data type that contains components that may be primitive data types such as numbers or strings of characters or complex data types such as other objects.

Property See attribute.

Prototype A model for a collection of one or more objects having a common structure.

Scripting language A programming language used to manipulate, customize, and automate the facilities of an existing system.

Source code Computer programming language statements that are readable by humans. Compilers and interpreters translate source code into code that a computer can execute.

Tag A command in an HTML document that defines an element or specifies an attribute of text or elements of a Web page. Tags are enclosed in angle brackets.

CROSS REFERENCES

See *Active Server Pages; Common Gateway Interface (CGI) Scripts; HTML/XHTML (HyperText Markup Language/ Extensible HyperText Markup Language); Java; Perl.*

REFERENCES

Document Object Model (DOM) (2002). Retrieved April 11, 2003, from http://www.w3.org/DOM/

ECMA (n.d.). Retrieved March 31, 2003, from http://www.ecma-international.org

FURTHER READING

Cohen, Y. (1997). *JavaScript cookbook*. New York: Wiley.

DevEdge Online Archive (n.d.). *Technologies*. Retrieved December 27, 2001, from http://developer.netscape.com/tech/

Dickson, P. (2001). *Sputnik—The shock of the century*. New York: Walker.

Flanagan, D. (2002). *JavaScript: The definitive guide* (4th ed.). Sebastopol, CA: O'Reilly.

Free On-Line Dictionary of Computing (n.d.). Retrieved May 14, 2002, from http://wombat.doc.ic.ac.uk/pub/darpa

Gosselin, D. (2002). *JavaScript*. Boston, MA: Course Technology.

JavaScript language resources (n.d.). Retrieved December 27, 2001, from http://www.mozilla.org/js/language

NCSA HTTPd Development Team (n.d.). *Common gateway interface.* Retrieved from http://hoohoo.ncsa.uiuc.edu/cgi/intro.html

O'Reilly (n.d.). *The source for Perl.* Retrieved from http://www.perl.com

PHP Group (n.d.). Retrieved May 14, 2002, from http://www.php.net

Sun Microsystems Inc. (n.d.). *Java*. Retrieved December 27, 2001, from http://java.sun.com

University of Albany Learning Technology Library (n.d.). *The Internet & the WWW: A history and introduction*. Retrieved May 14, 2002, from http://www.albany.edu/ltl/using/history.html

W3C: World Wide Web Consortium (n.d.). Retrieved December 27, 2001, from http://www.w3c.org

Wilton, P. (2000). *Beginning JavaScript*. Chicago: Wrox.

JavaServer Pages (JSP)

Frederick Pratter, *Eastern Oregon University*

INTRODUCTION: JAVASERVER PAGES IN CONTEXT

Originally, content on the World Wide Web was entirely static. Early on, however, it became clear that there needed to be some way to interact with Web content dynamically. In response, the HTML developers introduced the `<form>` element along with a variety of input modes that allowed users to develop graphical interfaces similar to those available in other tools such as Visual Basic and Delphi. The HTML `<form>` tag supports several attributes including method and action. These are used to specify the location of a script or program to be run when the form is submitted and the manner in which user-supplied form data are sent to the Web server.

In general, there are two ways to support interactive Web pages: client-side and server-side. Two of the early approaches to providing dynamic content using forms were *JavaScript*, which relies on downloading scripts that run in the client's browser, and *CGI* (see other chapters in this encyclopedia), where scripts written in Perl or another language run on the Web server.

Both of these have a number of perceived deficiencies that have led to alternative strategies for delivering dynamic content. Of these, the most common are ASP (Active Server Pages), PHP (Linux freeware; see http://www.php.net), and JSP (JavaSever Pages; see http://java.sun.com/products/jsp).

Active Server Pages only work on the Microsoft Internet Information Server. Where cross-platform applications are required (as for UNIX Web servers), the choice is between PHP and JSP. There are a great many sites running PHP, usually in conjunction with the MySQL database management system. At the same time, Sun Microsystems has done a great deal to make JavaServer Pages a powerful and scalable technology, and consequently it seems that this is an approach that will continue to be available and supported in the future.

These protocols do not rely exclusively on the HTML `<form>` tag but instead use a more sophisticated object-oriented approach. In particular Microsoft Active Server Pages use ADO objects in a COM framework to provide a distributed information environment. The equivalent in the cross-platform world is Sun's J2EE Web technology including servlets, JavaServer Pages, and JavaBeans.

JAVA SERVLETS

In order to make use of JavaServer Pages, it is necessary first to understand how servlets work (see http://java.sun.com/products/servlet). Java servlets are the server-side equivalent of Java client-side *applets*. The discussion that follows is intended as a general overview of the servlet life cycle. The details of how servlets work are described in detail in several of the references at the end of this chapter; in particular, see Hall (2000).

Like JavaScript and applets, servlets can only be run from within a Web server; there is no main method as in Java console applications. Servlets require both a Web server and a *servlet engine*. The function of this engine is to load the servlet `.class` file into the *Java virtual machine* (*JVM*) running on the server. The engine can then run the servlet.

Ordinarily, the `.class` file is not reloaded into the JVM again after the first time. Only one copy of the servlet is loaded into the Web server, and all clients receive a thread of the servlet, not a separate instance. In the early implementations of the servlet API, it was necessary to stop and restart the servlet engine manually in order to reload the `.class` file. Recent updates to the available servlet engines now include options to reload `.class` files automatically when they are updated.

The most widely available servlet engine is *Tomcat* from the Apache Foundation Jakarta Project (see Apache Jakarta Project, n.d., at http://jakarta.apache.org/tomcat).

There are also a number of other engines available; since corporate realignments seem to be a constant, however, it is necessary to check with the vendors frequently in order to keep track of the latest developments in servlet software (see http://java.sun.com/products/servlet/industry.html for an updated list of servers and engines).

Since it is the most widely employed servlet engine the examples presented in this chapter all use Tomcat. The examples were run using the Tomcat 4 Servlet/JSP container on a Linux Web server. (The only exception is Listing 9 which was run using the Windows version of the Tomcat engine because it requires ODBC, available only in a Microsoft environment.) This is not to suggest that this is only or even the best technology available, but only that since Tomcat is freely available, it is likely to be familiar to most Web developers. The reader is directed to the documentation for the specific products for details about how the following general functions are implemented there.

The canonical first program in any software tutorial is always “Hello World”; herewith an example:

Listing 1: Simple servlet program.

```
import java.io.*;
import javax.servlet.*;
import javax.servlet.http.*;

public class SimpleServlet extends
  HttpServlet
{
    public void doGet
      (HttpServletRequest request,
        HttpServletResponse response)
      throws ServletException,
           IOException
    {
      response.setContentType
        ("text/html");
      PrintWriter out =
        response.getWriter();

      out.println("<html>");
      out.println("<head>");
      out.println("<title>JavaServer
        Pages: Examples</title>");
      out.println("</head>");
      out.println("<body>");
      out.println("<h1>Simple Servlet
        Output</h1>");
      out.println("<p style=\"color:red\">
        Hello World!<p>");
      out.println("</body>");
      out.println("</html>");

      out.close();
    }

    public void doPost
     (HttpServletRequest request,
       HttpServletResponse response)
       throws ServletException,
         IOException
         {
               doGet(request, response);
         }
   }
```

In order to compile this servlet it is necessary to download the Java Servlet API contained in the file `servlet.jar` (from http://java.sun.com/products/servlet/download.html) and add it to the Java classpath. The `.jar` file contains the `javax` packages imported at the head of the listing above.

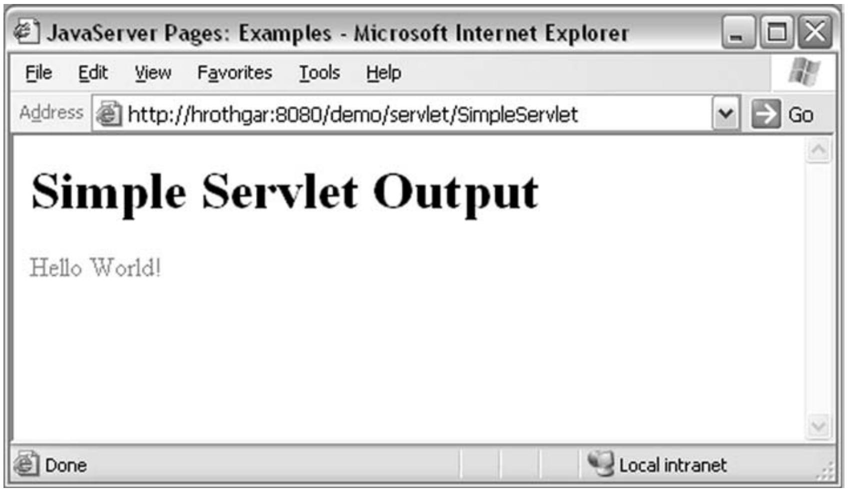
This class illustrates the most important features of servlet programming. The servlet class is derived from the parent class `javax.servlet.http.HttpServlet`. The servlet can be loaded in either of two ways, by either a **GET** or a **POST** request from an HTML form. Consequently, the two overridden methods are `doGet()` and `doPost()`. Since there is no way to know which of these is the referencing URL, the usual approach is as shown above: write the code to be executed in one of the methods, and simply call that one from the other. Both method calls result in the same code being executed. The `HttpServletRequest` parameter object is bound to the page that called the servlet, while the `HttpServletResponse` parameter is the generated page.

The servlet writes to a `PrintWriter` object that is bound to the response object by the `getWriter()` method; the HTML source is then generated directly by the `println` statements.

The Tomcat engine must be running as a process on the Web server; this is the responsibility of the system administrator, who can configure the server to start Tomcat automatically at boot up. Assuming Tomcat is started, the output of this program should appear to be as shown in Figure 1. Note that the URI shown references an instance of Tomcat running on port 8080. While the Apache Web server usually listens on port 80 by default, Tomcat uses a different TCP/IP connection. It is of course possible to set up Apache and Tomcat to work interchangeably on the default port; since this requires the addition of the mod_jk dynamic module to Apache, the reader is referred to the Apache foundation Web site for directions on how to accomplish this on various operating system platforms [see Shachor, G. (n.d.) at http://jakarta.apache.org/tomcat/tomcat-3.3-doc/mod_jk-howto.html].

DEPLOYING SERVLETS

One important but frequently overlooked issue in implementing Web sites is the directory structure of the Web server. This is of course dependent on the specific requirements at a given site. Many of the initial difficulties with implementing a JavaServer Page Web application have to do with details such as directory names and where `.class` files and JSP files are located. As the technology matures these low-level installation details should be abstracted away. Nonetheless, at present whichever server configuration is implemented, the rules for resolving URL references must be followed exactly.



JavaServer Pages: Examples - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address
Go
Done
Local intranet

http://hrothgar:8080/demo/servlet/SimpleServlet

Simple Servlet Output

Hello World!

Figure 1: Java servlet output.

The following discussion is specific to one engine, Apache Tomcat, and is in fact specific to a single version of that server. The details presented, however, illustrate many of the issues that need to be considered regardless of the site configuration chosen.

In an Apache installation, the initial server directives are included in the file httpd.conf in the `$APACHE_HOME`/co*nf* directory (where `$APACHE_HOME` is the location of the installation root). The DocumentRoot directive in this configuration file specifies the default directory from which HTML documents are served; usually this is $APACHE_HOME/htdocs.

Tomcat uses the $TOMCAT_HOME/webapps/ directory as the server root; URIs are specified with reference to this root. What's more, in order to invoke a servlet it is necessary to call an imaginary directory called servlet. Consequently, the actual path name to the servlet `.class` file is the WEB-INF/classes subdirectory, as illustrated below:

```
URI:       http://<server-name>:8080/
             examples/servlet/SimpleServlet
Pathname:  $TOMCAT_HOME/webapps/examples/
             WEB-INF/classes/SimpleServlet.
             class.
```

The servlet reference is automatically converted by the servlet engine. This convention can be confusing initially, but it is the standard that has been adopted. As will be seen below, since JSP pages rely on the servlet engine, the engine also must redirect the HTTP URI to the correct directory on the server.

HOW TO MAKE A JAVASERVER PAGE

A JavaServer Page can be created using any text editor, or a visual development environment may be used. In either case, the process is straightforward. First, create an HTML page, but save the file with a `.jsp` extension instead of `.html`. Then add three kinds of JSP elements:

1. *Scripting elements*—Java code that becomes part of the servlet;
2. *Directives*—Control overall structure of the servlet; and
3. *Actions*—Specify existing components that should be used.

Deploy the file to the correct location on the server and open in your favorite browser. That's it! Of course, in practice it's a little more complicated than that.

JavaServer Pages combine static HTML, known as *template text*, with dynamically generated content. Within the HTML Java code is enclosed in special tags. The JSP container compiles Java code (the first time only) and creates a `.class` file in the "work" directory on the server. Thus running JSP is a two-step process: first the servlet engine generates a .java file and then the source code is compiled and executed. (In production environments, the pages are usually precompiled after they are loaded onto the server. In this way, the first user of the page is not confronted with a lengthy delay while first the source and then the byte code is generated.)

Scripting Elements

Expressions

Expressions must evaluate to a type of `java.lang.String` since they are inserted directly into `println` statements. The initial symbol must be "<% = " as shown in the example below:

```
<%= expression%>
```

Note that you cannot include Java program code in an expression. The servlet container simply translates the expression into an output String, for example:

```
JSP code: <%= java.util.Date() %>
Generated servlet code:
  out.println(java.util.Date());
```

Scriptlets

Scriptlets are in-line Java code. The initial symbol must be "<%" as shown below:

```
<% code %>
```

Code is inserted into the generated servlet's _jspService method to determine dynamically what text, graphics, links, etc. are displayed. As in CGI, parameters are passed automatically via the URI and accessed in JSP through the "request" object (see below).

Scriptlets are useful for making HTML parts of JSP file conditional. For example,

```
<%
      if (null == request.
        getParameter("txtBox"))
      {
              out.println("You need to
                enter a value!");
      }
      else
      {
              out.println(txt);
      }
%>
```

Declarations

The initial symbol for a declaration must be "<%!" as shown below:

```
<%! declaration %>
```

These are inserted into the body of the servlet class, outside of the _jspService method. Typically they do not generate any output and are used in conjunction with expressions and scriptlets:

```
<%! private int accessCount = 0; %>
```

Accesses to page since server reboot:

```
<%= ++accessCount %>
```

Predefined Variables

The JSP container provides a number of *implicit objects* for use in scriptlets, including

request—The HttpservletRequest object;

response—The HttpservletResponse object;

session—The HttpSession associated with the request;

out—The PrintWriter used to send output to the client (there is also an *in* BufferedReader object for input);

application—The ServletContext object, used to store session attributes;

config—The ServletConfig object;

pageContext—The PageContext object, used to store page attributes; and

page—A reference to the current page; a synonym for "this."

Table 1 Page Directive Attributes

Page Attribute	Value
Import	Package.class
contentType	MIME-type
isThreadSafe	True|false
session	True|false
buffer	Sizekb|none
autoflush	True|false
extends	Package.class
Info	Message
errorPage	URI
isErrorPage	True|false
language	Java

For example, the *request* object can be utilized to obtain the values of parameters sent to the page by using the request.getParameter() method.

Directives

JSP directives control the overall structure of the servlet. There are three kinds of directives: page, include, and taglib. The required initial symbol is "<%@".

The page directive is used to specify various attributes included in the code to describe the output. Some of the available page attributes and their possible values are shown in the Table 1.

The include directive in JSP is used to insert an external file into the page at the specified point:

```
<%@ include file="menu.jsp" %>
```

The taglib directive references a custom tag library, and is explained in detail later on.

Actions

JSP actions employ XML syntax to specify existing components that should be used. An action start tag must use a prefix; for the standard actions listed below the prefix must be "jsp." (For custom tags it can be any name defined by the JSP taglib directive; see below.)

```
<jsp:include page="MyPage.jsp"
  flush="true"/>
<jsp:forward page="/utils/errorReporter.
  jsp"/>
<jsp:useBean id="myBean"
  class="MyBeanClass"/>
<jsp:setProperty name="myName"
  property="someProperty" ... />
<jsp:getProperty name="itemBean"
  property="numItems"/>
```

The include **action** has an effect similar to the include **directive** shown above. The difference is that a directive is executed at *page transition* time, that is, when the page is first translated into a servlet by the JSP container. The include action, on the other hand, includes files at the time the page is requested by the client.

The forward action redirects the browser to a new page, either .jsp or .html. The `useBean`, `setProperty`, and `getProperty` actions are described in the following section on JavaBeans.

A Simple JSP Example

Here is another version of the "Hello World" servlet, rewritten as a JavaServer Page:

Listing 2: JavaServer Page example: Hello World.

```
<html>
<head>
   <title>JSP Examples</title>
   <link rel="stylesheet" type="text/css"
     href="MyStyle.css">
</head>
<body>
   <h1>Simple JavaServer Page</h1>
   <% out.print("Hello World!"); %>
   The time now is
     <%= new java.util.Date() %>
</body>
</html>
```

The output of this program is shown in Figure 2.

Clearly, JSP relieves much of the burden of formatting the page from the developer. In place of the multiple Java `println` statements shown in Listing 1, this entire page is template text except forthe two lines of JSP code. These illustrate two difference ways to include Java code in JSP. The statement

```
<% out.print("Hello World!"); %>
```

is a `scriptlet`, that is, embedded Java code. The following line

```
The time now is <%=new java.util.Date() %>
```

is just HTML with the addition of a JSP `expression`.

The example also includes a CSS style sheet reference in order to illustrate one widely used approach toward formatting Web pages. The recent XHTML standard strongly deprecates the use of <font> tags and other in-line formats in favor of style sheets.

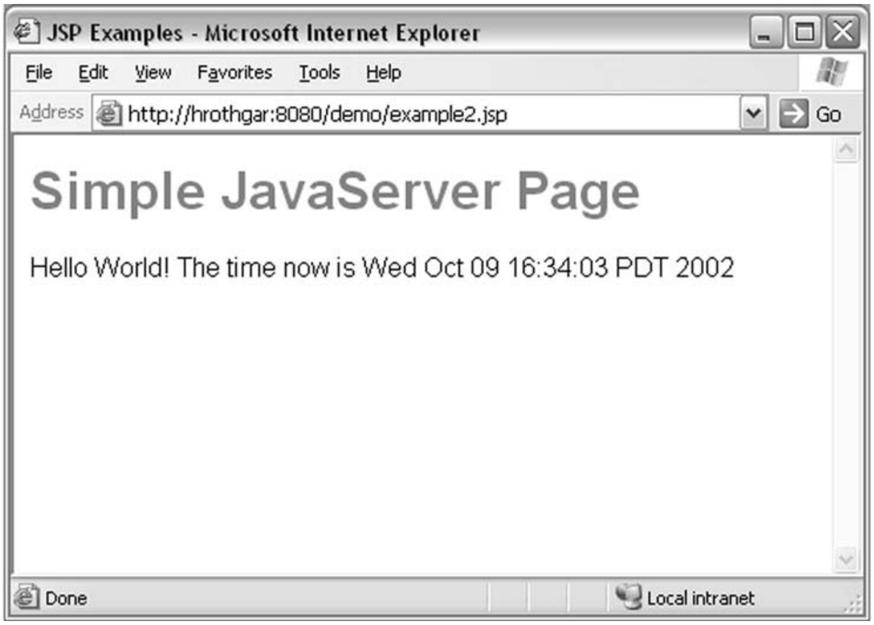
As noted above, Tomcat (along with all of the other available servlet engines) has very specific rules governing where files must be located and what they are to be named, but these rules are nearly all user-configurable. For example, the default directory for serving JavaServer Pages in Tomcat is $TOMCAT_HOME/webapps. However, the `server.xml` configuration file (located in the $TOMCAT_HOME/conf/ directory) can be modified to alter the behavior of the Tomcat JSP container, including changing the default Web directory (see http://jakarta.apache.org/tomcat/tomcat-4.0-doc/config for directions).

In any case, it is important to note that the actual path name of the JavaServer Page in the example above is not the same as the URI specified:

```
URI:       http://<server_name>:8080/demo/
             jsp example2.jsp
Pathname: $TOMCAT_HOME/webapps/demo/jsp/
             example2.jsp
```

This can lead to considerable confusion when first deploying JSP applications.

As noted above, running a JavaServer Page is a multistep process. First the JSP is parsed and a servlet is generated in the Tomcat work/localhost directory. Then this



JSP Examples - Microsoft Internet Explorer

Address http://hrothgar:8080/demo/example2.jsp

Simple JavaServer Page

Hello World! The time now is Wed Oct 09 16:34:03 PDT 2002

Figure 2: JavaServer Page output.

servlet is compiled and loaded. Finally, the servlet engine executes the resulting Java code and displays the HTML page in the client's browser.

Ordinarily, this servlet code is never seen by the user; the developer neither can nor should modify any of it. Instead the JSP container maintains all of the generated code. For illustrative purposes however, it is useful to examine the following servlet code, generated by the two lines of JSP embedded in the HTML. (The code shown has been reformatted somewhat for clarity.)

Listing 3: Servlet source code.

```
package org.apache.jsp;
import javax.servlet.*;
import javax.servlet.http.*;
import javax.servlet.jsp.*;
import org.apache.jasper.runtime.*;

public class example1$jsp extends
  HttpJspBase
{
   static {}
   public example1$jsp( ) {}
   private static boolean _jspx_
     inited = false;
   public final void _jspx_init()
      throws org.apache.jasper.runtime.
           JspException {}

   public void _jspService
      (HttpServletRequest request,
         HttpServletResponse response)
         throws java.io.IOException,
           ServletException
   {
      JspFactory _jspxFactory = null;
      PageContext pageContext = null;
      HttpSession session = null;
      ServletContext application = null;
      ServletConfig config = null;
      JspWriter out = null;
      Object page = this;
      String _value = null;
      try
      {
         if (_jspx_inited == false)
         {
            synchronized (this)
            {
               if (_jspx_inited == false)
               {
                  _jspx_init();
                  _jspx_inited = true;
               }
            }
         }
         _jspxFactory = JspFactory.
           getDefaultFactory();
         response.setContentType
            ("text/html;charset=
              ISO-8859--1");
```

```
   pageContext = _jspxFactory.
     getPageContext
      (this, request, response,"",
           true, 8192, true);
   application = pageContext.
     getServletContext();
   config = pageContext.
     getServletConfig();
   session = pageContext.
     getSession();
   out = pageContext.getOut();

// HTML
// begin [file="/jsp/example1.jsp";
      from=(0,0); to=(13,1)]
   out.write ("<?xml version=\"1.0\
     "?>\r\n
      <!DOCTYPE html PUBLIC
   \"-//W3C//DTD XHTML Transitional
     //EN\"\r\n
   \"http://www.w3.org/TR/xhtml1/
     DTD/
      xhtml1-transitional.dtd\">\
        r\n
   <html xmlns=\"http://www.w3.org/
     1999/xhtml\">\r\n
   <head>\r\n
   \t<title>JSP Examples</title>
     \r\n
   \t<style type=\"text/css\">\r\n
   \t h1 {color: blue;}\r\n
   \t body {font-family:helvetica,
     arial;}\r\n
   \t</style>\r\n</head>\r\n
   <body>\r\n
   \t<h1>Example 1. Simple
     JavaServer Page</h1>\r\n
   \t");
// end

// begin [file="/jsp/example1.jsp";
   from=(13,3);to=(13,33)]
   out.println("Hello World!");
// end

// begin [file="/jsp/example1.jsp";
   from=(13,35);to=(14,17)]
   out.write("\r\n\tThe time now
     is ");
// end

// begin [file="/jsp/example1.jsp";
   from=(14,20);to=(14,42)]
   out.print(new java.util.Date());
// end

// begin [file="/jsp/example1.jsp";
   from=(14,44);to=(18,0)]
   out.write("\r\n</body>\r\n
     </html>\r\n\r\n");
// end

}
```

```
      catch (Throwable t)
      {
         if (out != null && out.
           getBufferSize() != 0)
            out.clearBuffer();
         if (pageContext != null)
            pageContext.
                handlePageException(t);
      }
      finally
      {
         if (_jspxFactory != null)
            _jspxFactory.releasePage
              Context(pageContext);
      }
   }
}
```

The details of the generated servlet may seem arcane but in fact it's fairly easy to follow. The real work of the program is in the _jspService method The servlet creates the necessary HTML and writes it to the PrintWriter attached to the HttpServletResponse object. Thus the developer is spared the necessity of manually creating all of the write and print statements to generate the Web page.

JAVABEANS

Instead of inserting Java code directly into your JSP (as a scriptlet) it is usually a good idea to create a "bean" to implement good code reuse practice. JavaBeans are .**class** files that follow a few simple rules:

A bean class must have an empty constructor;

A bean class must have no public instance variables; and

Persistent values should be accessed through getXxx() and setXxx() methods (rather than through public variables) where Xxx is the name of a property.

Note that these all represent generally accepted coding practices.

The action for loading a bean is

```
<jsp:useBean id="name" class="package.
  class" />
```

The following JSP example is taken from the set supplied with the default Tomcat installation. It shows the code for using various Date class methods to display the current date and time:

Listing 4: JSP date bean example.

```
<html>
<!-
   Copyright © 1999 The Apache Software
     Foundation.
   All rights reserved.
→
<body bgcolor="white">
   <jsp:useBean id='clock' scope='page'
      class='dates.JspCalendar'
      type="dates.JspCalendar"/>
   <font size=4>
   <ul>
      <li>  Day of month: is  <jsp:
       getProperty name="clock"
        property="dayOfMonth"/>
      <li>  Year: is  <jsp:getProperty
       name="clock"
        property="year"/>
      <li>  Month: is  <jsp:getProperty
       name="clock"
        property="month"/>
      <li>  Time: is  <jsp:getProperty
       name="clock"
        property="time"/>
      <li>  Date: is  <jsp:getProperty
       name="clock"
        property="date"/>
      <li>  Day: is  <jsp:getProperty
       name="clock"
        property="day"/>
      <li>  Day Of Year: is  <jsp:
       getProperty name="clock"
        property="dayOfYear"/>
      <li>  Week Of Year: is  <jsp:
       getProperty name="clock"
        property="weekOfYear"/>
      <li>  era: is  <jsp:getProperty
       name="clock"
        property="era"/>
      <li>  DST Offset: is  <jsp:
       getProperty name="clock"
        property="DSTOffset"/>
      <li>  Zone Offset: is  <jsp:
       getProperty name="clock"
        property="zoneOffset"/>
  </ul>
  </font>
  </body>
  </html>
```

The JspCalendar class is not shown, but the source code can be viewed in the standard Tomcat installation at $TOMCAT-HOME/webapps/examples/WEB-INF/classes/dates/JspCalendar.java. The output of this program is shown in Figure 3.

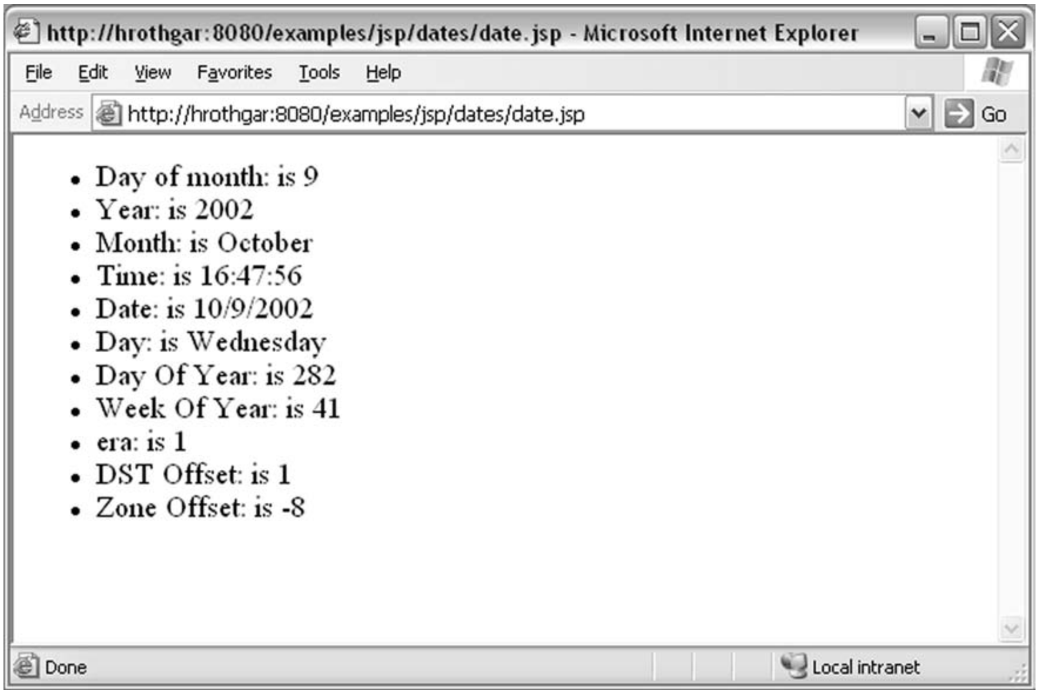
Note that selecting *View Source* on this page will not display the JSP code but instead the HTML generated by the servlet. The only way to view the original JSP is by looking at the following file:

```
$TOMCAT_HOME/webapps/examples/jsp/dates
  /date.jsp
```

Properties

There are two ways to load a property with a value, using either XML syntax or object dot notation in a scriptlet or expression:

```
<jsp:setProperty name="name"
  property="property" value="value" />
```



http://hrothgar:8080/examples/jsp/dates/date.jsp - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://hrothgar:8080/examples/jsp/dates/date.jsp
Go
Done
Local intranet

- Day of month: is 9
- Year: is 2002
- Month: is October
- Time: is 16:47:56
- Date: is 10/9/2002
- Day: is Wednesday
- Day Of Year: is 282
- Week Of Year: is 41
- era: is 1
- DST Offset: is 1
- Zone Offset: is -8

Figure 3: JavaBean output.

```
<% name.setProperty("value"); %>
```

There are also two ways to get the value of a property:

```
<jsp:getProperty name="name"
  property="propName" />
<%= name.getPropName(); %>
```

Parameters

When an HTML form calls another page, all of the data entered on the form are made available to the subsequent page as parameters. If the `GET` method is used, these are appended to the URL; if the form uses the `POST` method, the parameters are sent as a string that is not visible to the user, thus providing a small degree of security. (Unless the data have been encrypted they data can still be read easily; see http://www.itglossary.net/packetsnif.html for an example of one way to do this.)

JavaServer pages offer two ways to load a property with a parameter value:

```
<jsp:setProperty name="entry"
  property="itemID"
value='<%=request.getParameter("itemID")%>'/>

<jsp:setProperty name="entry"
  property="itemID" param="itemID"/>
```

The first approach uses the `getParameter()` method of the `HttpservletRequest` object to get the values of parameters passed to the page by the HTML get or post methods. The second way takes advantage of the `param` attribute of the JSP `setProperty` tag as a short cut to the value of the named parameter. Types are automatically converted, since all parameters are passed as text.

Scope

The <jsp:useBean> element can also have an optional "scope" attribute. This defines the scope of the bean and allows a single class to be used throughout the Web application or the HTTP session. The four possible values for scope are

1. *page*—Default;
2. *application*—Any servlet in the same Web application;
3. *session*—current HTTP session; and
4. *request*—The servlet request object (usually equivalent to the page).

The topic of scope is quite complex; the reader is referred to the References section for more detail, especially Hall (2000).

Deploying JavaBeans

The `.class` file for a bean must be accessible in the server classpath. For the Tomcat 4.0 examples this is $TOMCAT_HOME/webapps/examples/WEB-INF/classes/. The JSP container looks in this directory for the Java `.class` files. In the above example the file `JspCalendar.java` is in a package called dates so the `JSPCalendar.class` file is located in the subdirectory classes/dates. The directory structure for Tomcat is specified in the server.xml file, which by default is located in $TOMCAT_HOME/conf. By modifying this configuration file, different directories can be specified for different server applications. The directions for doing this are somewhat complex, and are covered in detail in the Tomcat documentation.

WEB ARCHIVE FILES

The examples shown so far have all been extremely simple. In the real world, most Web applications include multiple JavaServer Pages, usually linked to a package of classes. In order to make it easier to deploy complex applications, Sun has introduced the concept of Web application archive (WAR) files. A `.war` file is essentially just a regular Java `.jar` file containing all of the files and directories necessary to deploy a given Web application.

According to the *Java Servlet Specification 2.2*, Web applications can be a collection of JSP, servlets, HTML, Java classes, and other resources (see http://java.sun.com/webservices/docs/ea2/tutorial/doc/WebApp.html). Web application can be run from a **.war** file directly, or from an unpacked directory with the same structure.

The following discussion is Tomcat specific; the other servlet engines use quite different approaches. Consequently, this section should be considered as an example of one specific way to accomplish the general task of creating and deploying Web archive files.

The top level of the directory is the *document root* of the application. As noted previously, when using Tomcat the default document root is $TOMCAT_HOME/webapps. The webapps directory should contain a subdirectory called WEB-INF, which in turn can contain some or all of the following files and directories:

1. The `web.xml` file—the required Web application deployment descriptor;
2. Tag library descriptor files (see the next section);
3. A `classes` subdirectory that contains any needed servlets, utility classes, and JavaBeans components; and, optionally,
4. A `lib` subdirectory that contains JAR archives of libraries (tag libraries and any utility libraries called by server-side classes).

It is also possible to create package directories in either the document root or the `classes` subdirectory (see the Sun *WebApp* tutorial). Note that this is pretty much the default Tomcat structure described above—the **.war** file mechanism is just a way to package the contents is a fashion analogous to the familiar `.zip` or `.tar` files.

In order to create a .war file, it is first necessary to write a `web.xml` application deployment descriptor file. (The `web`.xml descriptor file should not be confused with the Tomcat configuration `server.xml` file described above.) The structure of the deployment descriptor file is fairly complex and will not be described in detail in this chapter. A good introduction is the Sun tutorial cited previously, as well as various other online sources [see Goodwill, J. (n.d.) at http://www.onjava.com/pub/a/onjava/2001/03/15/tomcat.html].

The *Java Web Services Developer Pack* available from Sun's Java Web site includes a Web application development tool called `deploytool` that can be used to generate deployment descriptors automatically. According to the Sun documentation (Web Application Deployment Tool for Java Web Services Developer Pack 1.0 EA2 Release Notes, n.d., available at http://java.sun.com/webservices/docs/ea2/deploytool/ReleaseNotes.html):

> In the previous version of the Java^{TM} Web Services Developer Pack, developers had to manually deploy a web service or web application by placing the web application archive (WAR) file into the "webapps" directory. The user also had to manually create the web.xml to configure the web application. Now the user can use the deploytool to perform the following operations:
>
> *Packaging*—The user can use deploytool to package their web application files into a WAR.
>
> *Configuration*—The user can use deploytool to configure the web application by using the New Web Component wizard in the tool to automatically generate the web.xml file.
>
> *Verification*—The user can use the Verifier in deploytool to verify that the configuration is correct.
>
> *Deployment*—The user can use Tomcat to deploy the web application to the Tomcat Server.

Clearly this is a rapidly evolving area; the developer is directed to the online resources cited in order to stay abreast of new developments.

JSP TAG LIBRARIES

JavaBeans allow the developer to encapsulate a great deal of functionality, but they cannot themselves manipulate JSP content. In order to create fully reusable distributed components, Java now provides support for custom JSP tag libraries (see http://java.sun.com/products/jsp/taglibraries.html).

The advantage of using custom JSP tags is that Web development can now be separated into two components: *presentation* and *logic*. Multimedia specialists can accomplish the work of creating the HTML template text while the components that provide the underlying functionality can be developed independently by Java programmers. By supplying the HTML coders with custom JSP tags, the entire process can be streamlined and made somewhat more robust.

The JSP specification provides for a set of standard actions to access objects and their properties. Actions are written using XML syntax (see the discussion above). In addition, however, the developer can create custom actions that have a wide range of functionality. In order to use custom tags, three components are necessary:

1. The `tag handler` class—Specifies tag's behavior;
2. The `tag library descriptor (TLD)` file—maps XML element names; and
3. One or more pages that use the tag library.

Tag Handlers

Tag handlers are a type of Java event handler, similar to the SAX-2 XML Parser interface. They are specialized Java classes that implement the `javax.servlet.jsp.tagext.tag` interface or one of its subinterfaces. The tag handler is invoked by the JSP container to implement the action specified by a custom tag. Some of the methods available in the tag API are

`doStartTag()`—Process the start tag for this instance;

`doEndTag()`—Process the end tag for this instance;

`getParent()`—Get the parent (closest enclosing tag handler) for this tag handler; and

`setParent()`—Set the parent (closest enclosing tag handler) of this tag handler.

For a simple tag without attributes or a body the developer simply overrides the doStartTag() with the code to be executed. When the JSP container encounters the tag, the generated servlet code will come from the referenced tag method. Examples of each of the three required files are given below.

The following Java program can be used to print out “Hello World” and the current date and time:

Listing 5: Custom tag handler.

```
package examples;

import java.util.Date;
import javax.servlet.jsp.*;
import javax.servlet.jsp.tagext.*;

public class SimpleTag
   extends TagSupport
{
   public int doStartTag() throws
    JspException
   {
      try
      {
           pageContext.getOut().print
           ("Hello World." + "The time
             now is " +
            new Date());
      }
      catch (Exception ex) {
         throw new JspTagException
         ("SimpleTag: " + ex.
           getMessage());
      }
      return SKIP_BODY;
   }
   public int doEndTag()
   {
      return EVAL_PAGE;
   }
}
```

The `doStartTag()` method is triggered when the start tag is encountered, sending the message to the client. The two possible return codes for `doEndTag()` are `EVAL_PAGE`, indicating that page evaluation should continue, and `SKIP_PAGE`, which causes the code on the remainder of the page to be bypassed.

Tag Library Descriptor

A *tag library descriptor* file has the extension `.tld`. A TLD is just an XML document that describes a tag library. The syntax is quite complex, however; see the online documentation for the XML Document Type Definition (DTD). There were significant changes between version 1.1 and version 1.2 of the DTD; make certain to use the correct specification (see http://java.sun.com/dtd/). The function of the TLD is to define the custom tags in the JSP. One possible TLD for the above example is shown below:

Listing 6: Tag library description.

```
<?xml version="1.0"?>
<!DOCTYPE taglib PUBLIC
"-//Sun Microsystems, Inc.//DTD JSP Tag
  Library 1.2//EN"
"http://java.sun.com/dtd/web-
  jsptaglibrary_1_2.dtd" >
<taglib>
   <tlib-version>1.0</tlib-version>
   <jsp-version>1.2</jsp-version>
   <short-name>Sample tag library
     </short-name>
   <description>Library for simple tag
     example</description>
   <tag>
         <name>SayHello</name>
         <tag-class>examples.SimpleTag
           </tag-class>
         <tei-class>EMPTY</tei-class>
         <description>Hello world
           example</description>
   </tag>
   </taglib>
```

The heading of this XML file reference the JSP 1.2 Tag library descriptor (DTD); a DTD must be included in the TLD or it will not work. The element names are as specified in the DTD; the root element is `<taglib>`, and one or more `<tag>` elements may be included. This TLD includes only one tag, called “SayHello.”

JSP with Custom Tags

A Web page that uses custom tags must include a `taglib` directive referencing this file before any custom tag is used. For example:

```
<% @ taglib uri="sample.tld"
  prefix="test" %>
```

The URI may point to a file in the same directory as the JSP, in another directory, or on any host accessible to the server.

Just as the `jsp:useBean` element uses the `id =` attribute to specify a prefix for bean properties and methods, so taglib uses `prefix =` to identify the XML namespace for custom tags. JSP custom tags are then written using this prefix

```
<test:tag>body</test:tag>
```

or for simple tags with no body, `<test:tag/>`.

Tags can and frequently do have attributes, and attribute values can be set by runtime expressions:

```
<logic:iterate collection="<%=bookDB.
  getBooks()%>"
id="book" type="database.BookDetails">
```

A page using the TLD and handler described above might look like the following:

Listing 7: JSP using custom tag.

```
<% @ taglib uri="simple.tld"
  prefix="test" %>
<html>
      <head>
   <title>JSP Examples </title>
   </head>
   <body>
      <h2>Custom Tag Example</h2>
      <test:SayHello/>
   </body>
   </html>
```

Inserting the custom tag results in the output shown in Figure 4.

Deploying Tag Libraries

Custom tags are generally collected into packages; in the above example, the class `SimpleTag` is part of the examples package. In `simple.tld`, the tag library descriptor file, the `SimpleTag` class is referenced as `examples.SimpleTag` in Listing 6. It is important to understand where each of these files is located on the server.

The JSP file is located under the Tomcat server root directory as

```
Pathname: $TOMCAT_HOME/webapps/examples/
            jsp/example3.jsp
URI:      <server-name>:8080/examples/jsp/
            example3.jsp
```

The TLD file can be anywhere the server can find it—in this case it is stored in the same directory as the JSP, so the URI in the taglib directive simply refers to simple.tld.

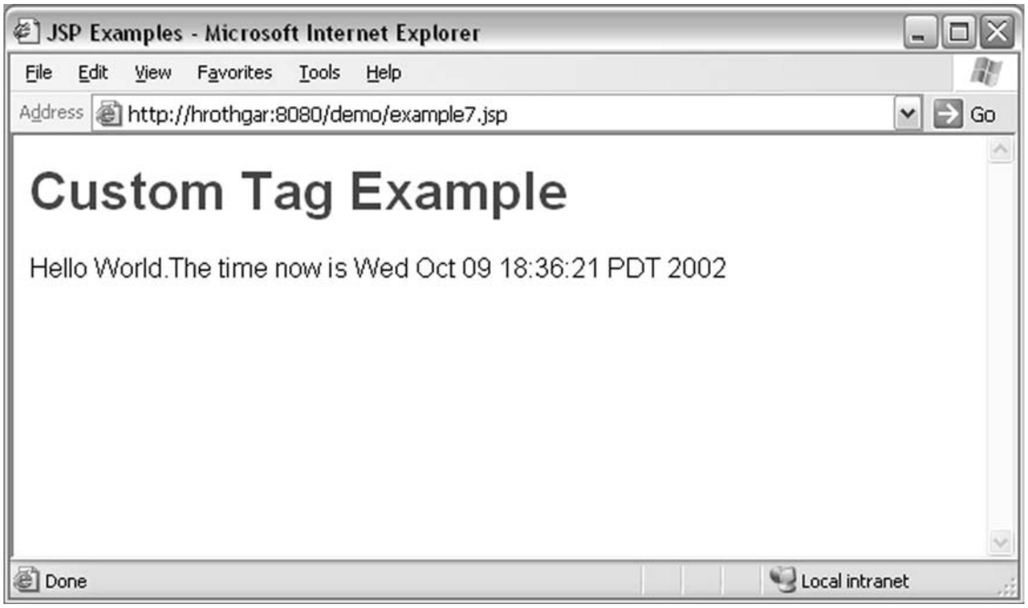
The Java `class` file must be in the servlet WEB-INF/classes subdirectory—since it is in the package examples, the path to this file is

```
$TOMCAT_HOME/webapps/examples/WEB-INF/
  classes/examples/SimpleTag.class
```

Additional custom tags can be referenced in the TLD file, and the associated `.class` files must also appear in this directory.

ACCESS TO DATABASES USING JDBC

One of the most common applications for JSP is reading and writing to databases. This is usually accomplished via *Java Database Connectivity* (*JDBC*); see http://java.sun.com/products/jdbc. Although this topic is not


JSP Examples - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://hrothgar:8080/demo/example7.jsp
Go
Done
Local intranet

Custom Tag Example

Hello World.The time now is Wed Oct 09 18:36:21 PDT 2002

Figure 4: JSP custom tag output.

strictly speaking part of the JSP specification, it is important to understand how Java database code can be implemented using JSP and particularly custom tags.

JDBC Overview

The first task in implementing a JDBC application is to choose a *driver*—these are available from various vendors, and are specific to a particular database management system (see http://industry.java.sun.com/products/jdbc/drivers). These drivers must be downloaded, installed on the server, and added to the Java classpath. There is also a JDBC–ODBC bridge driver for accessing database systems such as MS Access for which ODBC drivers are available. The JDBC–ODBC bridge is included with the Java JDK 1.4 (see http://java.sun.com/j2se/1.3/docs/guide/jdbc/getstart/bridge.doc.html). The JDBC API comprises two packages: `Java.sql` and `javax.sql`, which adds server-side capabilities. This review will only consider the classes and methods of the first package, `java.sql`, since this illustrates most of the concepts required for using JDBC with JSP. The developer is referred to the API documentation available at the Sun Web site.

Obtaining a Connection

In order to implement JDBC in a program, the first step after binding to a driver is to create a connection object.

The following example illustrates how to open a connection to an MS Access database; the principles are the same for Oracle, DB2, Sybase, MySQL, or any other RDBMS. The code is the same, only the driver and connections specifics are different.

Listing 8: JDBC-ODBC bridge.

```
import java.sql.*;

public class ODBCTest
{
   public static void main(String[] args)
   {
     Connection con = null;
     try // open connection to database
     {
        Class.forName("sun.jdbcOdbc
          JdbcOdbcDriver");
        con = DriverManager.
        getConnection("jdbc:odbc:
        Northwind");

        //print connection information
        DatabaseMetaData dma = con.
          getMetaData();
        System.out.println("\nConnected
          to " + dma.getURI());
        System.out.println("Driver " +
          dma.getDriverName());
        System.out.println("Version " +
          dma.getDriverVersion());
     }
     catch(Exception e)
     {
        e.printStackTrace();
     }
     finally//make sure connection gets
       closed properly
     {
        try
        {
           if (null ! = con) con.close();
        }
        catch (SQLException e) {}
     }
   }
}
```

The statement `Class.forName("sun.jdbc.odbc.JdbcOdbcDriver")` instantiates a new driver object. This allows the creation of a connection to the MS Access *Northwind* sample database. For this simple example, the program just opens a `DatabaseMetaData` object for the connection and prints some diagnostics. The output of this console application should be something like the following:

```
Connected to jdbc:odbc:Northwind
Driver JDBC-ODBC Bridge (odbcjt32.dll)
Version 2.0001 (04.00.6019)
```

In multiuser systems it is important to make sure the connection gets explicitly closed, since otherwise "zombie" processes may be created on the database server.

Using SQL

In order to read from or write to the database, JDBC uses SQL syntax. There are several types of objects available, the most important of which are

1. `Statement`—Execute a simple statement with no parameters;
2. `ResultSet`—Holds the results of a SELECT query;
3. `PreparedStatement`—Optimize repeated queries; and
4. `CallableStatement`—Call database stored procedures.

A `Statement` object is used to send SQL statements to the database: `PreparedStatement` and `CallableStatement` are actually both kinds of `Statements`.

Executing an SQL SELECT Statement creates a result set, as shown in the following example:

```
Statement stmt = con.createStatement();
ResultSet rs = stmt.executeQuery
      ("SELECT a, b, c FROM Table2");
```

There are three different methods for executing SQL statements:

`executeQuery`—For statements that produce a single ResultSet;
`executeUpdate`—For INSERT, UPDATE, or DELETE statements; and
`execute`—For complex queries.

A ResultSet object is the result of executing an SQL SELECT query. Database fields are returned by various getXXX methods, as shown below:

```
ResultSet rs = stmt.executeQuery
  ("SELECT a, b, c FROM Table1");
while (rs.next())
{
 int i = rs.getInt("a");
 String s = rs.getString("b");
 float f = rs.getFloat("c");
 System.out.println
   ("ROW = " + i + " " + s + " " + f);
}
```

If a ResultSet is empty, the return value of the Boolean method rs.next() is false; otherwise the effect of the method is to move the result set cursor to the next row. When the end of the set is reached, the value of rs.next() is set to false.

Using JDBC with JSP

The Apache Jakarta project has created a number of publicly available tag libraries, including the DBTags library for SQL databases (see http://jakarta.apache.org/taglibs/doc/dbtags-doc). The following example illustrates a page that uses this library:

Listing 9: Using JDBC with custom tags.

```
<?xml version="1.0"?>
<!DOCTYPE html PUBLIC
   "-//W3C//DTD XHTML 1.0
     Transitional/EN"
   "http://www.w3.org/TR/xhtml1/DTD/
     xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/
  xhtml">
<%@ taglib
  uri=http://jakarta.apache.org/
    taglibs/dbtags
  prefix="sql"%>
<head>
   <title>JSP Examples - JDBC</title>
   <style type="text/css">
     h1  {color: maroon;
       text-align:center; }
     h2  {color: blue;
       text-align: center;}
     th  {background-color: cyan; }
     caption {font-weight: bold; }
     body {font-family: helvetica,
       arial;}
   </style>
</head>

<body>
   <h1>Example 5. JavaServer Page</h1>
   <h2>JDBC Example</h2>
   <table border="1" align="center">
      <caption>
         Category = 1
      </caption>
      <tr>
         <th>ProductID</th>
         <th>ProductName</th>
      </tr>
      <%-- open a database connection
        --%>
      <sql:connection id="conn1">
         <sql:driver>
            sun.jdbc.odbc.
              JdbcOdbcDriver
         </sql:driver>
         <sql:URI>
            jdbc:odbc:Northwind
         </sql:URI>
      </sql:connection>

      <%-- open a database query --%>
      <sql:statement id="stmt1" conn=
        "conn1">
         <sql:query>
            select ProductID,
              ProductName
               from products
               where CategoryID = 1
         </sql:query>
         <%-- loop through the rows of
           your query --%>
         <sql:resultSet id="rset2">
           <tr>
              <td><sql:getColumn
                position= "1"/></td>
              <td><sql:getColumn
                position="2"/></td>
              <td>
                  <sql:wasNull>
                       No records
                         were
                         selected.
                  </sql:wasNull>
              </td>
           </tr>
        </sql:resultSet>
      </sql:statement>
      <%-- close database connection
        --%>
      <sql:closeConnection conn="conn1"
        />
   </table>
  </body>
 </html>
```

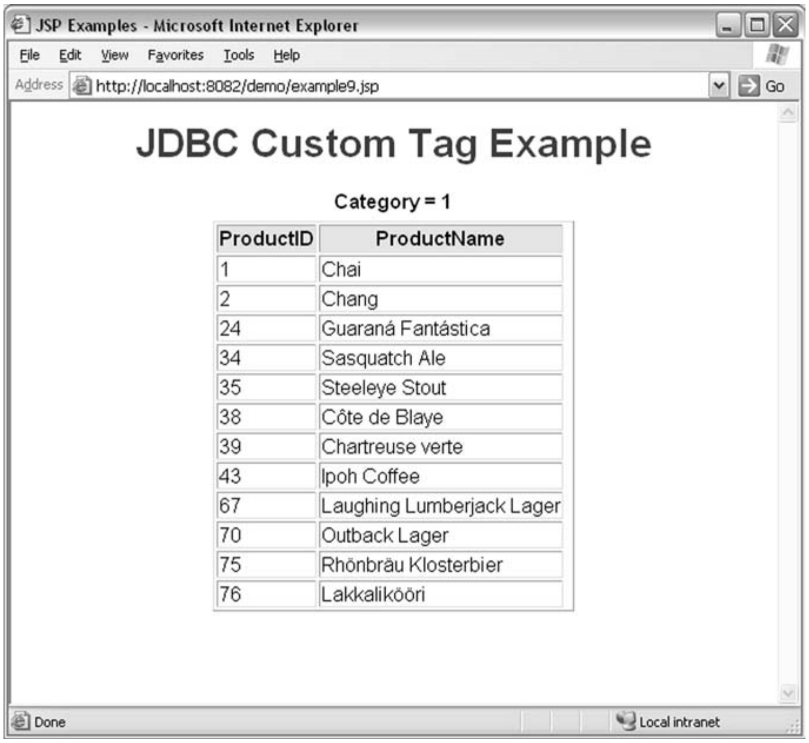
The custom tags encapsulate the JDBC database actions described above:

sql:connection—Create a JDBC connection;

sql:driver—Specify driver, subelement of sql:connection;

sql:URI—Specify database URI, subelement of sql:connection;

sql:statement—Create a JDBC statement;


JSP Examples - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Address
http://localhost:8082/demo/example9.jsp
Go
Done
Local intranet

JDBC Custom Tag Example

Category = 1

ProductID	ProductName
1	Chai
2	Chang
24	Guaraná Fantástica
34	Sasquatch Ale
35	Steeleye Stout
38	Côte de Blaye
39	Chartreuse verte
43	Ipoh Coffee
67	Laughing Lumberjack Lager
70	Outback Lager
75	Rhönbräu Klosterbier
76	Lakkaliköori

Figure 5: JavaServer Page JDBC custom tag output.

`sql:query`—Specify SQL query, subelement of sql:statement;

`sql:resultSet`—Iterate over query results, subelement of sql:statement;

`sql:getColumn`—Return field from ResultSet, subelement of sql:statement;

`sql:wasNull`—Message to print if ResultSet is empty; and

`sql:closeConnection`—Close JDBC connection.

The rest of the program is just standard XHTML, as in Listing 3 above. The resulting output is shown in Figure 5.

The SQL query "`select ProductID, ProductName from products where categoryID = 1`" results in 12 records being displayed. The table is extracted from the MS Access sample database and displayed as JSP, without recourse to any Java programming. The Web developer need only know the behavior of the tags, and the rest of the results are assured.

In order to use the DBTags library, several files need to be located in the appropriate directories on the server:

1. Download a binary distribution of the DBTags library from the Apache Jakarta Web site (see http://jakarta.apache.org/builds/jakarta-taglibs/releases/ dbtags).
2. Unzip the distribution file (currently available as jakarta-taglibs-dbtags-1.0-B1.zip).
3. Copy the tag library descriptor file dbtags.tld to a/WEB-INF subdirectory of your Web application—for this example this is

   ```
   $TOMCATHOME/webapps/examples/WEB-INF/jsp
   ```
4. Copy the tag library .jar file to the/lib subdirectory of your Web application—for this example

   ```
   $TOMCATHOME/lib
   ```
5. Edit the/WEB_INF/web.xml configuration file—in this example

   ```
   $TOMCATHOME/webapps/examples/WEB-INF/web.
     xml
   ```
6. Locate the existing <taglib> elements and add the following right after the others in the installation file:

   ```
   <taglib>
     <taglib-uri>
          http://jakarta.apache.org/taglibs/
            dbtags
          </taglib-uri>
     <taglib-location>
          /WEB-INF/dbtags.tld
          </taglib-location>
   </taglib>
   ```
7. Add the following directive at the top of each JSP:

   ```
   <%@ taglib uri="http://jakarta.apache.org
     /taglibs/dbtags"
         prefix="sql"%>
   ```

Any prefix may be used but "sql" is the convention.

CONCLUSION

The following marketing prose from the Sun JSP Technology Web site is a good summary of the advantages of this approach:

> JavaServer Pages technology uses XML-like tags that encapsulate the logic that generates the content for the page. Additionally, the application logic can reside in server-based resources (such as JavaBeans component architecture) that the page accesses with these tags. Any and all formatting (HTML or XML) tags are passed directly back to the response page. By separating the page logic from its design and display and supporting a reusable component-based design, JSP technology makes it faster and easier than ever to build web-based applications. (JavaServer Pages, n.d.)

Overstatement aside, the particular advantages of JSP are that it is platform independent and that it is highly extensible. JSP technology, as with most of the Web, is evolving so rapidly that no survey could ever attempt to be comprehensive. This chapter has only touched on the surface aspects. There are many good reference works on JSP; a partial list can be found below. Furthermore, the online references listed are an essential place for the developer to keep track of the changes and new concepts so quickly emerging in the area of dynamic, server-based Web development.

GLOSSARY

.class file A Java binary file containing machine-independent byte code.

.jar file A Java archive file containing one or more .class files.

.war file A .jar file that contains a Java Web application.

Active Data Object (ADO) Microsoft's newest interface for accessing data objects.

Active Server Page (ASP) A Web page including dynamic content provided by the Microsoft Internet Information Server (IIS).

Apache The Apache Software Foundation, which provides support for a number of open source software projects. The Apache HTTP Server is the most widely deployed Web server software.

Applet A small Java program designed to be run by a Web client from within a browser.

Application Program Interface (API) A set of specifications and tools for building and running software applications.

Bytecode The result of compiling Java source code; interpreted by the Java Virtual Machine.

Classpath The list of directories that Java will search to locate a required class.

Client A software application that runs on a PC or workstation and connects to a *server* that provides data or Web content.

Common gateway interface (CGI) A specification for dynamic transfer of information between a Web server and a client.

Component Object Model (COM) A Microsoft specification that allows programs to access objects that support this public interface.

Constructor A Java method that specifies the code to be executed when an object is created.

Hypertext markup language (HTML) A specification for creating documents for the World Wide Web.

Java A platform-independent, object-oriented programming language developed by Sun Microsystems; widely used for Web applications.

Java 2 Enterprise Edition (J2EE) A set of APIs that can be used to create multitier Web applications.

JavaBeans A specification developed by Sun Microsystems that specifies how Java objects interact; can be run on any platform that supports the Java Virtual Machine.

Java Database Connectivity (JDBC) API that allows Java programs to execute SQL statements.

JavaScript A scripting language design by Netscape to allow client-side dynamic content in Web pages.

JavaServer Pages (JSP) Web pages containing dynamic content supplied by Java servlets.

Java Virtual Machine (JVM) A platform-specific interpreter for running Java bytecode.

JSP action HTML-embedded code that is executed when a JSP page is requested by a client.

JSP custom tag JSP element tag that provides some user-specified action (see tag library).

JSP declaration Used to declare a scripting language variable or method; code is inserted into the generated servlet.

JSP directive Used to supply information that does not differ between requests for a page.

JSP expression A syntactically complete, valid expression inserted into the servlet code; must result in a String.

JSP scriptlet A code fragment inserted in the servlet.

Method A function or subprocedure in object-oriented programming, executed when the object receives a message.

Parameter An argument to a function or subprocedure; parameters in HTTP are content passed to or from a Web page.

PHP An open-source, server-side embedded scripting language; it is included with most Linux releases.

Scope The range of statements in which a variable or method is visible.

Server A software application that provides network-shared resources such as e-mail, printing, Web content, or data.

Servlet A small Java program run by the Web server to provide dynamic content.

Servlet engine A stand-alone program that supports the servlet API.

Structured query language (SQL) A standardized way to access information in relational databases.

Tag handler A JavaBean that provides methods to get and set properties of a custom action.

Tag library A collection of custom actions.

Tag library descriptor (TLD) An XML file that maps custom actions to tag handlers.

Tomcat A freely available servlet engine created by the Jakarta Project of the Apache Software Foundation.

Uniform resource identifier (URI) The generic term for an Internet address.

Uniform resource locator (URL) One type of URI.

CROSS REFERENCES

See *Active Server Pages; ActiveX Data Objects (ADO); Common Gateway Interface (CGI) Scripts; Dynamic HTML (HyperText Markup Language); HTML/XHTML (HyperText Markup Language/Extensible HyperText Markup Language); JavaBeans and Software Architecture; JavaScript; Perl; Web Content Management*.

REFERENCES

Active Server Pages (n.d.). Retrieved April 1, 2003, from http://msdn.microsoft.com/library/default.asp?URL=/library/psdk/iisref/aspguide.htm

Apache Jakarta Project (n.d.). Retrieved April 1, 2003, from http://jakarta.apache.org/tomcat/

Apache Jakarta Project Overview (n.d.). Retrieved April 1, 2003, from http://jakarta.apache.org/tomcat/tomcat-4.0-doc/config/

Goodwill, J. (n.d.) Java Web Applications. Retrieved April 1, 2003, from http://www.onjava.com/pub/a/onjava/2001/03/15/tomcat.html

Jakarta Project: DBTags Tag Library (n.d.). Retrieved April 1, 2003, from http://jakarta.apache.org/taglibs/doc/dbtags-doc/

Java 2 Platform, Enterprise Edition (n.d.). Retrieved April 1, 2003, from http://java.sun.com/dtd/

JavaServer Pages (n.d.). Retrieved April 1, 2003, from http://java.sun.com/products/jsp/

JavaServer Pages Tag Library (n.d.). Retrieved April 1, 2003, from http://java.sun.com/products/jsp/taglibraries.html

Java Servlet Technology (n.d.). Retrieved April 1, 2003, from http://java.sun.com/products/servlet/

JDBC Data Access API (n.d.). Retrieved April 1, 2003, from http://industry.java.sun.com/products/jdbc/drivers and http://java.sun.com/products/jdbc/

JDBC-ODBC Bridge Driver (n.d.). Retrieved April 1, 2003, from http://java.sun.com/j2se/1.3/docs/guide/jdbc/getstart/bridge.doc.html

PHP: Hypertext Preprocessor (n.d.). Retrieved April 1, 2003, from http://www.php.net/

Shachor, G. (n.d.) Working with mod_jk. Retrieved April 1, 2003, from http://jakarta.apache.org/tomcat/tomcat-3.3-doc/mod_jk-howto.html

Web Applications (n.d.). Retrieved April 1, 2003, from http://java.sun.com/webservices/docs/ea2/tutorial/doc/WebApp.html

Web Application Deployment Tool for Java Web Services Development Pack (n.d.). Retrieved April 1, 2003, from http://java.sun.com/webservices/docs/ea2/deploytool/ReleaseNotes.html

Web Clients and Components (n.d.). Retrieved April 1, 2003, from http://java.sun.com/j2ee/tutorial/1-3-fcs/doc/WCC.html

FURTHER READING

JavaServer Pages

Bergsten, H. (2002). *JavaServer Pages* (2nd ed.). Sebastopol, CA: O'Reilly.

Brogden, W. B. (2000). *Java developer's guide to servlets and JSP*. Alameda, CA: Sybex.

Burd, B. A. (2001). *JSP: JavaServer Pages developer's guide*. New York: Hungry Minds.

Callaway, D. R. R., & Coward, D. (2001). *Inside servlets: Server-side programming for the Java platform* (2nd ed.). Reading, MA: Addison-Wesley.

Fields, D. K., Kolb, M. A., & Bayern, S. (2001). *Web development with Java Server Pages*. Greenwich, CT: Manning.

Geary, D. M. (2001). *Advanced JavaServer Pages*. Englewood Cliffs, NJ: Prentice Hall PTR.

Goodwill, J., & Mehta, S. (2001). *Developing Java servlets*. Indianapolis, IN: Sams.

Hall, M. (2000). *Core servlets and JavaServer Pages*. Englewood Cliffs, NJ: Prentice Hall PTR.

Hall, M. (2001). *More servlets and JavaServer Pages*. Englewood Cliffs, NJ: Prentice Hall PTR.

Hougland, D., & Tavistock, A. (2000). *Core JSP*. Englewood Cliffs, NJ: Prentice Hall PTR.

Hougland, D., & Tavistock, A. (2001). *Essential JSP for Web professionals*. Englewood Cliffs, NJ: Prentice Hall PTR.

Hunter, J., Crawford, W., & Ferguson, P. (2001). *Java servlet programming* (2nd ed.). Sebastopol, CA: O'Reilly.

Kurniawan, B. (2002). *Java for the Web with servlet, JSP, and EJB: A developer's guide to scalable J2EE solutions*. Indianapolis, IN: New Riders.

Monson-Haefel, R. (2001). *Enterprise JavaBeans* (3rd ed.). Sebastopol, CA: O'Reilly.

Moss, K. (1999). *Java servlets*. New York: McGraw-Hill.

Smith, D. (2002). *Java 2 for the World Wide Web*. Berkeley, CA: Peachpit Press.

Williamson, A. R. (1999). *Java servlets: By example*. Greenwich, CT: Manning.

Custom JSP Tags

da Silva, W. L. S. (2001). *JSP and tag libraries for Web development*. Indianapolis, IN: New Riders.

Goodwill, J. (2002). *Mastering JSP custom tags and tag libraries*. New York: Wiley.

Heaton, J. (2002). *JSTL: JSP standard tag library*. Indianapolis, IN: Sams.

Schachor, G., Chace, A., & Rydin, M. (2001). *JSP tag libraries*. Greenwich, CT: Manning.

Weissinger, A. K., & Weissinger, K. (2002). *Developing JSP custom tag libraries*. Sebastopol, CA: O'Reilly.

JDBC

Harms, D. (2001). *JSP, servlets, and MySQL*. New York: Wiley.

Reese, G. (2000). *Database programming with JDBC and Java* (2nd ed.). Sebastopol, CA: O'Reilly.

Turner, J. (2002). *MySQL and JSP Web applications: Datadriven programming using Tomcat and MySQL* Indianapolis, IN: Sams.

White, S., Fisher, M., Cattell, R., Hamilton, G., & Hapner, M. (1999). *JDBC API tutorial and reference: Universal data access for the Java 2 platform (Java series)* (2nd ed.). Reading, MA: Addison-Wesley.

Williamson, A. R., & Moran, C. (1997). *Java database programming: Servlets and JDBC*. Englewood Cliffs, NJ: Prentice Hall PTR.

K

Knowledge Management

Ronald R. Tidd, *Central Washington University*

INTRODUCTION

Imagine that you recently returned from an adventure vacation in a equatorial rain forest. Prior to your trip, you checked the Web sites of the U.S. State Department to verify the area's security and the Centers for Disease Control (CDC) to identify the suggested vaccinations. During the trip, you consumed only food and water that had been properly prepared, avoided contact with animals that might transmit diseases or parasites, and generally practiced safe living without compromising your adventure. On your return, you "enjoyed" a relatively comfortable, but long and crowded, international flight.

Now you are experiencing physical symptoms that suggest a case of the flu. Because these symptoms have persisted and increased in severity over a couple of days, you go online and connect to your physician's Web portal. Your daily scheduling software connects with the physician's daily scheduling software, and they arrange an appointment for the following day. Before logging off, you complete a patient interview form that apprises your physician of your symptoms. Your responses include information about your recent international travel, including locations and flight schedules. You also authorize your physician to report your symptoms to the CDC, but with an appropriate degree of privacy that protects your identity. (The CDC will know your symptoms and travel itinerary but no personally identifiable information.) You authenticate your authorization with your digital signature.

While you try to sleep, the CDC computer system is collecting thousands of similar reports from around the World. It is using data mining techniques to discern patterns in the data, including a pattern related to your symptoms. The system notifies the appropriate CDC personnel and simultaneously e-mails an advisory to physicians who are treating affected patients. They are advised to monitor those patients closely and submit information about the success of treatments administered. The advisory is also sent to key medical centers around the World and the leading pharmaceutical companies.

Your medical condition continues to deteriorate during the night, but as your physician leaves home the next morning, he finds that his personal digital assistant (PDA) downloaded the CDC advisory and your interview form over a wireless wide area network during the night. When he arrives at his office, he discovers that his nurse received an update on the severity of the illness, the need to put you into quarantine, and the most successful treatment protocols. She has called you and the ambulance company closest to you to arrange for immediate transport to an isolation unit at the nearest hospital.

Simultaneously, your physician's computer system notified the hospital's computer system. It checked the medications inventory to verify the availability of the required medications and advised the necessary medical personal over their PDAs. Those medical personnel also continuously receive updates on their PDAs from the ambulance crew about your vital statistics while you are in transit to the hospital. Because the ambulance is equipped with a GPS location device, hospital personnel monitor the ambulance's expected time of arrival. Once your treatment starts, the hospital's data will be fed to the CDC system so that its success can be assessed and conveyed to other physicians.

Your physician's computer system also notified your medical insurance provider, which issued digital authorization for your emergency treatment to the ambulance company, hospital, and physician. The system has already started to generate the records necessary to ensure that all of the participants in this situation receive the appropriate reimbursements in a timely manner.

KNOWLEDGE MANAGEMENT

The medical community is at least several years away from implementing the processes and the technologies used in this hypothetical medical scenario, but the vignette illustrates how networked computers could be used to collect and process data and information, then deliver it to decision makers in a value chain. Sometimes the decision makers will use the prescriptions and procedures suggested by the information. Other times they will use the information as a catalyst for the generation of new knowledge based on their personal situations and experiences. The scenario also assumes that the participants in this value chain have overcome myriad organizational, cultural, and legal constraints. For example, concerns about patient privacy and proprietary research, plus the tendency of experts to enhance their professional reputations by hoarding rather than sharing knowledge, may prevent businesses from realizing the promise of knowledge management in e-commerce.

The extent to which an organization realizes that promise depends on the perspective of knowledge and knowledge management that is brought to bear on the design and implementation of its knowledge management system. That perspective also determines the focus of a knowledge management system (KMS) and its processes. Alavi and Leidner (2001, p. 111) identified six such perspectives:

1. **Knowledge vis-à-vis data and information.** Data are input that create the information, which becomes knowledge when an individual processes it (hierarchical model). A KMS focuses on user assimilation of information.
2. **State of mind.** Knowledge is a state of knowing and understanding. A KMS provides access to sources of knowledge.
3. **Object.** Knowledge is an object to be stored and manipulated. A KMS helps gather, store, and transfer knowledge.
4. **Process.** Knowledge is the process of applying expertise. A KMS provides links among sources of knowledge to increase the depth and breadth of knowledge flows.
5. **Access to information.** Knowledge is a condition of access to information. A KMS provides search and retrieval mechanisms for locating relevant information.
6. **Capability.** Knowledge is the potential to influence action. A KMS supports development of individual and organizational competencies.

The knowledge–data–information perspective guides the discussion in the remainder of this chapter. It should become obvious, however, that a "best practices" KMS draws from all of the perspectives.

What Is Knowledge?

As is frequently noted in discussions about knowledge management, philosophers have debated for centuries as to what constitutes knowledge and how it is distinguished from and related to data and information. There is some consensus about definitions that have pragmatic significance. One perspective uses a hierarchical model of data, information, and knowledge:

- Data are facts that have no intrinsic value. In the medical scenario, data includes the list of passengers on your international airline, the list of physicians on the CDC's system, the list of patients with symptoms reported to the CDC, and the medications used.
- Information is data that are processed and analyzed such that they can be used to help make decisions. In the medical scenario, information would include the correlation between your recently reported symptoms and your presence on a given international flight, as well as the analysis of treatments administered and their success.

In this hierarchical model, data are the input used to create information and information is one input used to create knowledge.

The definition of knowledge is more complicated (Polanyi, 1967) and depends to some extent on the other inputs that are combined with information to create the knowledge. When the inputs include an individual's context and experience but do not include an ability to articulate in language or a process for documentation, the knowledge is referred to as *tacit knowledge*. In other words, it is knowledge that is embedded in an expert's cognitive processes. In the medical scenario, the physician who first treated a patient reporting the symptoms associated with this international outbreak, analyzed patient data (symptoms and vital signs) to generate information, then processed that information within the medical context and added his or her professional beliefs and past experiences with similar symptoms.

When knowledge is articulated in language and documented, it becomes *explicit*. It is likely that the physicians also relied on explicit knowledge about the efficacy of certain medications in the treatment of certain symptoms and afflictions. Such knowledge became explicit when it was documented in medical journals and the publications distributed by pharmaceutical companies, including the *Merck Manual of Diagnosis and Therapy.* In this scenario, the CDC converted the tacit knowledge of the physicians into explicit knowledge through the processes of accumulating, analyzing, documenting, and disseminating the successes and failures of their prescribed treatments. Those processes are necessary to create explicit knowledge. Stewart (2001, p. 125) makes a useful distinction between how the two types of knowledge are developed: People are trained in the use of explicit knowledge; they are coached in the development of tacit knowledge.

Why Should Knowledge Be Managed?

Managers are responsible for managing organizational resources to fulfill the organization's mission effectively and efficiently. During the industrial age, this meant managing hard, tangible assets such as land, factories, equipment, and a significant amount of relatively unskilled labor. That labor "punched time clocks," performed the same task on a fixed production line day after day, and was treated as an expense rather than an investment.

Now, there is some consensus that civilization is entering an information age or a digital economy (Martin, 1999; Minkin, 1995; Stewart, 2001; Tapscott, 1996). One of the expected characteristics of the mature economies in this emerging epoch is that they will rely less on manufacturing industries and more on service industries. The evidence suggests that this is already happening (Sveiby, 1997). This supports the increased emphasis on knowledge workers (assets) such as the attorneys, accountants, and consultants that many service industries employ.

It would be a potentially disastrous mistake for managers in manufacturing enterprises not to become adept at managing knowledge resources, however. Computer technologies have created a more connected workforce in manufacturing companies and allowed decision making to occur at much lower levels in the hierarchy. Industrial age workers who used to tighten the same nut and bolt every day are now information age workers who now work at the same computer terminal every day. Their educational levels are increasing, and they are authorized to make decisions regarding the operation of production lines. They are increasingly recognized as a valuable source of knowledge with respect to the conversion of ideas on the drawing board to products on the store shelves and showrooms.

That employee knowledge and competence is the focus of a second expected characteristic of the mature economies: Mature economies will depend less on tangible assets and more on intangible assets that include an organization's employee competence and the internal and external structures that they create and support (Sveiby, 1997). Organizations that learn how to manage employee knowledge and competence can do the following:

- Reduce development time for workforce and innovation by delivering the appropriate information and knowledge to the workforce in a timely manner.
- Assimilate knowledge for improved customer support by analyzing customer interactions documented through help-desk inquiries, correspondence, and Web site activity.
- Organize increasing complexity in the organization and marketplace, in part by identifying an appropriate taxonomy for classifying key resources and processes.
- Retain knowledge within the organization regardless of workforce mobility by identifying and collecting the knowledge and information contained in the structured (e.g., databases and spreadsheets) and unstructured documentation (e.g., e-mails, Web pages, and correspondence) generated and collected by the workforce.

- Reduce development time for workforce and innovation by delivering the appropriate information and knowledge to the workforce in a timely manner.
- Assimilate knowledge for improved customer support by analyzing customer interactions documented through help-desk inquiries, correspondence, and Web site activity.
- Organize increasing complexity in the organization and marketplace, in part by identifying an appropriate taxonomy for classifying key resources and processes.

Thus, proper management of knowledge workers and their expertise helps organizations improve customer satisfaction and maintain a competitive edge.

What Is Knowledge Management?

There is no consensus about what constitutes knowledge management or even about whether knowledge can be managed. Data, information, and explicit knowledge are readily documented and, therefore, clearly manageable. Tacit knowledge is highly personalized and internalized. Thus, some claim it is unmanageable. Others claim that it is manageable to the extent that the expert who possesses it is manageable. Using the latter perspective for guidance, there are several possible definitions and descriptions of knowledge management from which to choose. Whichever is chosen, it is imperative that it recognize the strategic importance of knowledge management to organizational success. Knowledge management is about business, not technology.

Barclay and Murray (1997) provided one such definition. Their original version referred to businesses, but it can be generalized to any organization with minor modifications. With those modifications, knowledge management is defined as an organizational activity that

- treats knowledge as an explicit concern of the organization and
- links positive organizational results to explicit and tacit knowledge assets.

The former requires that organizations adopt strategies, policies, procedures, and practices that support the collection and sharing of knowledge throughout their operations—at all levels and in all functional areas. The latter requires that organizations identify the intellectual assets that can be used to enhance operations or sold as knowledge products (Dawson, 2000; Stewart, 2001).

THE KNOWLEDGE MANAGEMENT SYSTEM

There is some debate about the degree of dependence of knowledge management systems on information technologies (ITs; Alavi & Leidner, 2001, p. 114). Clearly, IT is not essential to the development of a viable KMS: Public accounting and legal firms maintained successful knowledge management programs well before desktop computers were integrated into their professional practices. IT does provide obvious enabling tools, however, tools that enhance the effectiveness and efficiency of a KMS. This is especially true when the KMS is integrated into an e-commerce system that relies on a solid IT infrastructure. Thus, the following discussion focuses on an IT-based knowledge management system. This does not imply that the IT department must manage the KMS.

When viewed in this manner, it is possible to design and implement a KMS using the same conceptual foundation that guides the design and implementation of any IT system or subsystem. That foundation starts with the specification of a goal for the KMS. That goal is the essential precursor for strategies that will combine process

Table 1 Conversion Modes for Knowledge Transformation (derived from Nonaka & Takeuchi, 1995)

FROM	TO	MODE
Tacit	Tacit	Socialization through face-to-face communications between people who share a common culture and can work together effectively
	Explicit	Externalization through conceptualization and elicitation that leads to articulation, typically in collaborative efforts
Explicit	Tacit	Internalization through reflection about multiple sources of information, including reports and journal articles and e-learning
	Explicit	Combination through exposure to knowledge through meetings, documents, e-mail, and training and education via e-learning

and an infrastructure to help fulfill the goal. The process starts with identifying useful knowledge and ends with delivering it to decision makers. The infrastructure comprises two components: The IT infrastructure focuses on computer technology and the human resources that use and manage that technology within an organization; the organizational infrastructure focuses on the organization's hierarchical structure and its impact on knowledge sharing efforts.

KMS Goals and Strategies

That system's ultimate goal is to help fulfill organizational goals by sharing, extending, and creating explicit and tacit knowledge. It must do so when and where the knowledge is needed and in a manner that is personalized for an individual decision maker. The strategies used to fulfill this goal must adhere to two important rules:

1. Business mission drives the KMS effort and computer technology enables that effort. The KMS must help solve business problems, not highlight technological capacities and abilities.
2. All stakeholders must be involved in the design of the KMS then trained in its use. Those who will use the system must buy into it and understand that the organization's management and leadership also have bought into the effort.

The latter rule suggests the importance of the human resource and organizational cultural components to the successful KM strategy.

The strategies for implementing a KM system must explicitly recognize the importance of properly training the user, because the investment in the human resources may generate a greater return on investment than the investment in the enabling technologies (Dawson, 2000, p. 75; Eisenhart, 2002). Because users work within an organizational culture, the implementation strategy also must consider how the culture influences their behavior with respect to the use of a KMS.

De Long and Fahey (2000) identified four "frameworks" in which this occurs, in their discussion of how culture

- Shapes assumptions about which knowledge is important,
- Mediates the relationships between knowledge that belongs to the individual, to his or her work group, or to his or her organization,
- Creates a context for social interaction, and
- Shapes creation and adoption of new knowledge.

Their examination of the linkage between culture and behavior toward knowledge resources provides evidence that "any discussion of knowledge in organization settings without explicit reference to its cultural contest is like to be misleading." De Long and Fahey also enumerated diagnostic actions that managers can take within each framework to help align the goals of the KM and the culture of the organization.

The strategies for the design and implementation of the KMS must also consider that the transformation of the different kinds of knowledge between people requires different conversion modes (Dawson, 2000; Marwick, 2001). Table 1 illustrates the four possible combinations for converting the tacit or explicit knowledge of one person to the tacit or explicit knowledge of another. It also shows that the KMS must provide the appropriate processes and technologies to support the conversion that an organization requires. A given transformation cannot occur in the absence of the appropriate system.

KMS Processes

Manville (1999) identifies two different philosophies that can guide the design of the KMS. The more established of the two reflects an "engineering mindset." It is based on the assumption that there is an omnipotent individual in the organization who is capable of defining the taxonomy used to catalog the organization's knowledge. It is a top-down approach.

The less established and potentially more powerful philosophy is based on the complex adaptive systems (emergent systems) paradigm. It relies on the assumption that only those who possess and create the knowledge can create the taxonomy. To prevent their diversity from degenerating into chaos, they will devise a system that allows only the most successful contributions to the schema to survive. It is a bottom-up approach.

Whether the KMS is guided by the first or the second philosophy, it still follows the same basic processes. It must identify, collect, store, and deliver knowledge

Table 2 KMS Processes

KMS PROCESS	KEY ACTIVITIES AND CONSIDERATIONS
Identify	Identify the appropriate communities of practice. Identify the mission-critical knowledge that needs to be managed for each community. Develop the taxonomy to be used to organize and catalog that knowledge.
Collect	Embed the collection process in the daily work process. Rely on the technologies that the contributors use on a daily basis. Develop a nurturing culture that values knowledge sharing. Recognize and reward those whose contributions add value to the organization's operations.
Store	Prevent unauthorized physical and virtual access to the KMS, including the hardware, software, and knowledge. Perform timely backups that are stored off site. Test the backup recovery process.
Deliver	Identify how those who will use the KMS work. Train them in the appropriate use of the KMS. Use some combination of push and pull strategies to deliver the information to them.

(Table 2). This model is most appropriate for the management of data, information, and explicit knowledge. With slight modification of the storage process, it can be altered to help manage the tacit knowledge embedded in the minds of knowledge workers.

The identification stage is the most complicated in the process. It also is where the differential impact of the two philosophies is most noticeable. The two main goals in this stage are to identify the knowledge that needs to be managed because it helps a unit of analysis achieve its goals and then to classify that knowledge according to an appropriate vocabulary or taxonomy. The knowledge that needs to be managed and classified may be found in documents with content that is structured (e.g., databases and forms) and documents with content that is unstructured (e.g., spreadsheets, Web pages, letters, e-mail, and reports), as well as embedded in the knowledge workers.

To fulfill those two goals, it is necessary to identify the appropriate organizational units for which knowledge is to be managed. For example, it could be by functional specialization. Thus, emergency room physicians and nurses, ambulance personnel, hospital inventory managers, and hospital accountants would each have their own separate taxonomy. Alternatively, the unit could be defined as a service (experience) community (Martin, 1999, pp. 185–220) or a community of practice in which networked computers transcend functional, geographic, and hierarchical boundaries. Then the physicians and nurses, ambulance personnel, hospital inventory managers, and hospital accountants might work together on a cross-functional emergency medicine project team, thus becoming a community of practice. (Geographically dispersed personnel can form a virtual community of practice.) Their focus shifts from their respective functional areas to the provision of cost-effective emergency medical services. Concomitantly, the knowledge each person needs may shift and the community of practice for which that knowledge must be managed also may shift. In fact, it may be constantly shifting as team members participate in multiple projects at any time and move to different projects over time.

Once the mission critical knowledge for the appropriate unit of analysis is identified, it must be classified with a taxonomy. A taxonomy provides a bridge between those who provide knowledge and information and those who need it. It does so by integrating

- structure with terms that are arranged in a hierarchy and cross-referenced and
- pointers to the relevant knowledge resource, whether a digital document or a human expert.

(A book's index and a library's card catalog are manifestations of taxonomies.) Thus, the taxonomy helps establish the credibility of the KMS and the trust that users will have in it. Its importance cannot be overstated.

Manville (1999) explained that creating a taxonomy is a daunting task: Users from different functional areas or with different academic, cultural, and experiential backgrounds use different terminology. They may also use the same terminology in different ways. In slightly different terms, the context in which terminology, information, and knowledge evolve varies. Also, terminology tends to be dynamic, changing as knowledge workers revise their terminology and as the KMS expands to include new sources. The consequent Tower of Babel requires a mediation mechanism that generates a taxonomy that adapts to needs that change over time or between user groups. This is especially important as the diversity of the users of a KMS increases.

Currently, the most feasible mechanisms include the designation of an individual or a committee to manage the taxonomy. As the amount of data, information, and knowledge that is managed by a KMS increases and diversifies, however, this approach becomes less feasible. Technologies such as XML (extensible markup language), fuzzy logic, and topic maps (Pepper, 2000) will supplement or supplant direct human intervention. They will enhance the value of a KMS by making it more accessible to a more diverse audience on a more timely basis. Pepper (2000) maintained that a topic map could become a knowledge asset itself, by virtue of its portability across organizational domains.

The second step in the process is to collect the knowledge from the individuals who have it and are motivated to share it. This is more likely to happen when they know that their organization rewards their efforts and they trust that it will not exploit them. In other words, there must be an appropriately designed incentive system and a nurturing culture.

For example, a consultant who uses his or her expertise to develop a leading-edge solution to a client's problem can "announce" his expertise via the consulting firm's communications channels. That might be through a listing in the firm's telephone directory of experts and expertise. When other consultants confront similar client problems, they contact the firm's expert to discuss the solution. The expert receives peer recognition via the discussion and firm "recognition" through the generation of billable hours that the client must pay for.

Alternatively, the expert would document the solution in a manner that is consistent with the community's vocabulary and standards, then submit it to the computerized KMS. When other consultants confront similar client problems, they contact that KMS to locate the solution. Clearly the documentation must identify the author to ensure peer recognition and establish the solution's credibility based on the expert's reputation. The KMS also must provide a means by which the expert receives firm recognition. This might include a method for assessing and rewarding how frequently the expert contributes to the KMS, how often those contributions are referred to by the community of practice, and how much value those contributions add to the service of other clients. Recognition and rewards must be commensurate with the value of the contributions.

Successful collection efforts also are more likely if the collection process is integrated into the normal workflow (rather than appended to it) using the technology that the expert uses on a daily basis. Regarding the consulting example, one collection approach requires the consultant to fill out a lengthy debriefing report at the end of every engagement. An alternative approach requires the consultant to use computer- based forms during the planning and performance of the consulting engagement, so as to capture the knowledge in an appropriate format and in a timely manner. Although debriefing can be a valuable means of collecting feedback and is preferred in many situations, the latter approach is less intrusive and more expeditious.

The processes for storing knowledge should not be as troublesome as those for classifying and collecting it. The primary goals are to keep the knowledge secure and protected, typically using the network technology discussed in a subsequent section. To this end, it is imperative that qualified personnel be hired to manage the KMS and that they recognize the importance of the resources that they manage. They should use a prevent-detect-correct-recover strategy to secure the hardware, software, and knowledge: The preventative measures should prevent unauthorized physical and virtual access to these resources; the detection processes should recognize and identify any unauthorized access, then prevent reoccurrences. In addition, the stored knowledge must be backed up on a regular and timely basis and stored in off-site locations.

The last process is to deliver relevant knowledge to the appropriate decision makers on a timely basis, such that they will use the KMS to fulfill the organization's mission more effectively and efficiently. As in the collection process, it is more likely that users will utilize the KMS if delivery relies on the technologies that they work with daily. To the extent the KMS uses unfamiliar technologies, such as search engines, the processes must provide training. The processes must also support more personal (face-to-face) communications when "delivering" tacit knowledge (Table 1).

Delivery will use some combination of push and pull strategies. The latter allows users to extract knowledge from the KMS using tools like the integrated search engines that search all of an organization's stored knowledge. The former sends the knowledge to those who need it, perhaps because they subscribed to the system or because the system tracked their search and work efforts to identify knowledge that is relevant to their efforts.

The KMS Technological Infrastructure

The technological infrastructure of the KMS has two main components: the technology component, which consists of hardware, software, and the data, information, and knowledge that the KMS processes; and the human resources component, which includes those that use and those that manage the KMS.

Computer networks are the "backbone" technology for a modern KMS. They facilitate knowledge sharing processes that can occur anytime, anyplace, with anyone who has access to the network. When used in a KMS, they must be configured to accommodate the technology used by the knowledge users and contributors—that is, the hardware (desktop and laptop personal computers, personal digital assistants, mobile phones, audio–video equipment, and perhaps automobiles and refrigerators) and software (spreadsheet, word-processing, database, e-mail, and scheduling applications).

While the taxonomy is probably the most important factor in the design of the KMS, a strong document management system is the most important factor for populating and using the KMS. The computer network must be configured to capture and store the documents prepared and used by employees. When the documents exist in printed from only, it is necessary to convert them to a digital format. Typically this is done with scanners and software with optical character recognition (OCR) abilities. Both the scanned documents and the documents already in a digital format (i.e., those prepared in spreadsheet, word-processing, database, and e-mail applications) need to be indexed in a manner that is consistent with the taxonomy.

There are two basic strategies for storing the digital documents and files. The first involves using the storage media (hard drives) on the computers used by those who create the documents. Those computers must be connected to the network and available when the

Table 3 Computer Applications for Knowledge Transformation (derived from Alavi & Leidner, 2001)

FROM	TO	MODE AND MAIN APPLICATIONS
Tacit	Tacit	Socialization: groupware, conferencing applications, personnel directories
	Explicit	Externalization: discussion lists, bulletin boards, groupware, conferencing applications
Explicit	Tacit	Internalization: search engines, e-learning
	Explicit	Combination: groupware, search engines, e-learning, Web portals

files are likely to be needed. Recent developments in peer-to-peer computing make this strategy possible. The second strategy involves using servers dedicated to storing and providing the contents of the KMS. The more sophisticated (and constantly evolving) implementations rely on high-speed servers in either storage area networks (SANS) or network-attached storage (NAS). The capacity and speed available with these configurations is essential given the vast amount of digital data and information that is being collected, such as that taken from Web sites.

The computer network must also be configured with applications that are appropriate for the conversion mode (socialization, externalization, internalization, combination) for the type of knowledge that an organization works with and needs to convert (Table 3).

The socialization mode has proven to be the most difficult mode for computer technologies to support. Trust is a prerequisite for converting one individual's tacit knowledge into another individual's tacit knowledge. In the absence of an existing trust relationship, it is not easy to establish or convey with the applications currently in use (Marwick, 2001), such as the following:

- **Groupware,** which establishes a virtual workspace in which the members of a community of practice can communicate and coordinate group efforts. Groupware supports communications that are synchronous (everyone is online and communicating at the same time, as in an online meeting) or asynchronous (everyone is online and communicating at a different time, as when posting to a discussion page or sending e-mail).
- **Conferencing software,** which provides a richer communications channel for synchronous communications through the use of audio and video in addition to text-based chat. Some applications support real-time document sharing and collaboration.
- **Personnel directories,** which identify the experts within the organization. A valuable carryover that predates the advent of computer networks.

Conferencing software will probably become the tool of choice, because it supports real-time communications and a view of the communicator's body language, which is critical to the socialization process.

The externalization mode requires a dialogue between individuals in which there is elicitation, articulation, and conceptualization, so it uses the same collaboration applications as used to support the socialization mode. Discussion lists and bulletin boards also facilitate the requisite dialog. Marwick (2001, p. 819) maintained that whichever application it used, the most effective results probably occur when the dialogue is more informal and free.

Computer technology is well suited to support the internalization mode, which requires an individual to reflect on multiple sources of information about a topic. The first task for the KMS is to locate and assess the relevance of documents and files with explicit knowledge. This requires the use of robust search engines and techniques, especially when the search involves the "infoglut" on the Internet, in addition to resources on an organization's intranet. The search engines and methods must be capable of locating relevant files whether they are textual, aural, or graphical in nature. The second task for the KMS is to help the individual digest the relevant explicit knowledge and convert it into tacit knowledge. This may require a structured "e-learning" experience that guides the individual through the process by an appropriate sequence of learning tasks (e.g., computer simulations) and reflection. In other situations, visualization techniques can help the individual understand the topic and its context and relationships. Visualization mechanisms include mind mapping, concept mapping, various kinds of "topographic" topic maps, and any one of the numerous charts and graphs commonly found in spreadsheet and statistical analysis software.

Computer technology is best suited to support the combination mode, which only involves explicit knowledge that is, by definition, documented. Once again, the first task of the KMS is to locate and evaluate that documentation and files for relevance. The second task is to make it readily available to the user. Once again, good search engines and search techniques for various types of media files are essential. Web (knowledge) portals provide an increasingly familiar and powerful means of consolidating the sources of knowledge from multiple locations on an intranet or the Internet.

As discussed earlier, computer technology is an enabling tool used to collect, store, and deliver the knowledge created by humans. In the long run, it is likely that the technology will add a second role in knowledge management. With the advent of more sophisticated technologies (e.g., intelligent agents and grid computing), the technology could become the creator of new knowledge, the discoverer of deep knowledge. Current data mining techniques provide a glimpse of that future.

The main considerations and concerns about the human resource component of the technological infrastructure were discussed under the policies section, summarized as follows:

- KMS users must be trained in its use and motivated to use it in a manner that enhances organizational performance.
- KMS administrators must be trained in its administration and aware of its strategic significance.

An additional consideration is the need to provide a third set of skills and knowledge to the KMS project—a knowledge expert. Whereas an organization's topic experts understand the data, information, and knowledge that the organization needs and its KMS administrators understand the computer networks that the organization uses, the knowledge expert understands how knowledge is created and used by people. In a sense, the knowledge expert is the bridge between the KMS users and the KMS administrators. This triumvirate of skills and knowledge is necessary to a successful KMS implementation. Rarely is it found in a single individual.

KMS Organizational Infrastructure

For this chapter, the organizational infrastructure refers to the organizational structure that houses those who use the technology (e.g., business units) and those who manage the technology (e.g., the IT department). It is not likely that a KMS will require or cause a significant change in the formal organizational structure of the business units or the IT department. Computer networks, however, transcend the formal boundaries that define hierarchies, chains of command, and functional departments. Consequently, implementation of a KMS probably requires a change in the organizational culture. That culture must minimize the barriers to collaboration and support knowledge sharing rather than knowledge hoarding and facilitate a frictionless flow of knowledge among fluid communities of practice.

One requirement for implementing the cultural change (that may have a corollary impact on organizational structure) is the appointment of respected experts within the organization as knowledge managers and knowledge arbiters. Their role is to facilitate wise and informed use of the KMS. Managers may play a dual role as coaches who need to motivate participation in the system and gatekeepers who determine which contributions deserve to be included in or need to be removed from the KMS. Arbiters restrict their efforts to the latter task and to ruling on changes to the system, including revisions to the vocabulary and taxonomy. Within any organization, these roles will probably be filled by multiple individuals and perhaps by committees (Manville, 1999).

KMS Metrics

As with all IT projects, it is advisable to assess the success of a KMS effort and just as difficult to do so. Whether using return on investment (ROI) or total cost of ownership (TCO) as the assessment criterion, there are a myriad of costs and benefits that are difficult to measure. For example, the impact on employee morale is impossible to determine and will go in both directions—the implementation of a KMS will distress some employees and elate others.

Regardless of the known difficulties and inaccuracies, ROI and TCO should be calculated and analyzed, as should other established financial metrics such as revenues per employee and market share. There also are nonfinancial ("knowledge") metrics that deserve consideration, such as the following:

- Time to market for new products and services,
- Employee satisfaction as measured by retention,
- Customer satisfaction as measured by retention and repeat business, and
- Time to resolve customer problems.

Sveiby (1997, pp. 185–202) devoted an entire chapter to similarly nontraditional metrics for intangible assets. These emerging metrics may not be valid in the minds of some managers at this time, but they are quite appropriate in a digital economy. They will gain acceptance as more organizations recognize that the intangible assets such as employee knowledge are the most significant assets that they have.

Unlike other IT projects, it also is necessary to assess individual performance within the KMS implementation. This is particularly true for those whose contributions are a factor in the determination of incentives, rewards, and recognition. The assessment must focus on some combination of the quality and quantity of the contributions. The KMS technology can be useful in the assessment process because it can capture quantity data such as the number of contributions and the number of times that a contribution was accessed. The technology cannot replace human judgment, however, when assessing the criteria that measures the real value of contributions—their quality.

It is advisable to use multiple metrics to triangulate on a measure that is as elusive as quality. To start, knowledge managers and arbiters can serve in an assessment capacity given that they should be familiar with the contributions. In addition, organizations could mimic online publishers that elicit reader feedback about the usefulness of a document. Extending the philosophy espoused by Sveiby (1997), it might be possible to develop new measures based on the impact that a contributor's efforts had on the success of the organization and its clients and customers and the creativity and complexity of the contributions. Clearly, there is a definite need for new assessment metrics at both the individual and the project levels.

THE IMPACT OF E-COMMERCE

When an organization engages in e-commerce or b-commerce, it implements a business strategy that allows it to exchange economic value and information with the participants in its supply chain across electronic networks. When the organization implements a knowledge strategy that extends its computer network to include all of those supply chain partners, it creates a network that currently goes by several different names:

Business web or B-web (Tapscott, Ticoll, & Lowry, 2000)
Collaborative commerce or c-commerce (McCoy, 2002)
Value network (Keen & McDonald, 2000).

Whatever it is called, McCoy (2002) noted that this Internet-based network integrates the organization's relationship management systems such as its customer relationship management (CRM) and supply chain management (SCM) systems, with its enterprise resource planning (ERP) system. (Each of these topics is discussed in greater depth in other chapters of this encyclopedia.)

Metcalfe's Law suggests the potential power of this extended configuration. It states that computer power increases by the square of the number of nodes on a network. Consequently, an organization may experience an exponential increase in its exposure to data, information, and, it is hoped, knowledge by virtue of establishing a knowledge network with its business partners. When properly configured, the shared data-information-knowledge would be available to all authorized participants in the network virtually immediately, regardless of their location. Thus, as the medical example at the beginning of this chapter illustrated, a patient (customer) can use a Web browser to verify the status of an appointment (order). Simultaneously, the CDC (supplier) can detect geographically specific patterns in patient responses (customer behavior) and direct the appropriate medical information and supplies (resources) to the locations where they add value to the process.

The theoretical exponential increase in power does not come cheaply. There are goals and strategies to specify, processes to define, and technologies to be designed and implemented. All of this must be accomplished using a computer network, just like before. But now they must be accomplished across the hierarchical and cultural boundaries of multiple organizations. The requisite coordination and cooperation is guaranteed to extract a heavy price.

E-commerce KMS Goals and Strategies

The guiding concept for the extended e-commerce KMS is that it is only as strong as the weakest participant. With the extension of the network's reach, there must be a concomitant extension of the KMS goals and strategies. It is necessary to find a common thread to bind the participants in the extended KMS, however, or they will degenerate into the chaos of self-interested competition with each other.

The consensus is that the end customer provides that common thread. As the final arbiter of the supply chain's value, the supply chain participants should share information and cooperate to enhance the value delivered to that customer. Many commentators share the sentiments expressed by Keen and McDonald (2000, p. 91): "participating organizations maintain their associations so as to leverage their capabilities and opportunities in a way that attracts customers, keeps customers, and creates associations of mutual benefit."

Although the customer serves as a common goal, the participants still need to develop an appropriate strategy. Technologies with cross-platform compatibilities may simplify the coordination process, but that process will still require a significant effort. As before, the most appropriate "meta-strategy" is to develop a pilot project with a high-profile participant where there is both a high probability of success and high political and economic payoffs. Also as before, it must be based on the types of knowledge that the participants are trying to convert.

E-commerce KMS Processes

Once again, the interorganizational KMS requires few changes to the KMS processes vis-à-vis the intraorganizational KMS (Table 2), but it injects several more coordination challenges. The most obvious is related to the taxonomy. The language used by knowledge workers who share a similar professional background may be common, but still influenced and differentiated by their respective organizational environments. The severity of those influences increases as the communities of practice become cross-functional and international.

An even bigger challenge exists with respect to the nature of the alliances in the extended KMS. Those alliances (virtual enterprises) may be virtual and ephemeral. An alliance formed for one project may involve previously unknown partners and then be disbanded with the project's completion. Trust is going to be difficult to develop. This situation is likely to cause problems for interorganizational knowledge sharing, which, although often difficult even within an organization because of a lack of trust, becomes even more difficult when a knowledge transfer is to an alliance participant that could be a competitor tomorrow. Network security poses another problem. The security of a multiorganizational network is only as strong as the security used by the weakest participant in that network. Organizations will be reluctant to expose their data-information-knowledge to others if there is a concern that it will be compromised. This would be a more significant problem for professionals who are legally obligated to protect confidential client information (e.g., doctors and attorneys).

The absence of trust will require the emergence of third-party services or other mechanisms to guarantee the fidelity of the participants and provide indemnity. This will be problematic in a global economy in which there are significant cultural differences in the definition of copyrights, trademarks, and patents.

The E-commerce KMS Technological Infrastructure

A computer network is the backbone of the extended KMS, just as it was for the organizational KMS. Basically, it uses the same technological infrastructure: digital devices for inputting and accessing knowledge, servers for storing knowledge, search engines and portals for organizing knowledge. It also has two special concerns: security and scalability.

The Internet is *the* computer network and channel for e-commerce and extended KMS. In its current configuration, it also is inherently insecure. As mentioned, a network's security is only as strong as its weakest link. For example, a denial of service attack on one participant with weak security could cripple the entire supply chain. Thus, it is imperative that all members of an alliance implement security that is appropriate for the data-information-knowledge that is accessible from or affected by their node. It is likely that third-party verification and audits of that security will be necessary to instill the requisite trust.

In the context of this topic, scalability refers to the ability of the KMS to change as the number or sophistication of its users change. In the dynamic and fluid world of virtual enterprises, members frequently may be added and dropped. This situation could make the coordination process prohibitively expensive unless the participants adopt standardized technologies. The goal is to ensure cross-platform compatibility, such that communication is possible regardless of the technology used by any of the participants. The use of proprietary applications and protocols that only work when both ends of the communications channel have identical systems must be avoided.

XML is one such standard. It should have a tremendous impact on c-commerce because it is platform neutral. In addition, and perhaps more important, many communities of practice are emerging to device XML taxonomies that are appropriate for their needs (e.g., mathematics, chemistry, and financial reporting). Those taxonomies could provide guidance on the development of taxonomies for a KMS.

E-commerce KMS Organizational Infrastructure

The organizational infrastructures of both the business units and the IT departments is still the concern, except there are now multiple organizations. It also still is unlikely that these infrastructures need to be reorganized for any of the participating organizations with the implementation of an extended e-commerce KMS. The biggest concern is again organizational culture and the extent to which it supports knowledge sharing across organizational borders. As before, management must establish a culture that works to help the supply chain fulfill its mission.

Knowledge managers and arbiters are more necessary to the successful KMS in the e-commerce environment. First, the effort requires an overall manager. The potential difficulty is in deciding which organization he or she should come from. If one organization in the supply chain is the proverbial 800-pound gorilla, the choice may be obvious. Alternatively, there should be at least one keystone organization that provides a permanent nucleus to the ephemeral supply chain. Again, the choice may be obvious. In the absence of a dominant or permanent force, however, it will be problematic. The biggest problem is the apparent transfer of power and control over components of the supply chain by all of the participants to a manager employed by only one of the participants.

Similar concerns affect the selection of managers and arbiters for the different communities of practice. Clearly, each manager should be a content expert, but political and cultural issues may intervene in the selection. Also of concern is the impact that the knowledge manager from one organization might have on the rewards earned by a member of a community of practice from another organization. In effect, the knowledge manager is making human resource decisions for another supply chain participant.

E-commerce KMS Metrics

The e-commerce KMS also creates challenges for the assessment process. One significant challenge occurs when a cost–benefit analysis for the entire supply chain proves that the extended KMS makes economic sense. Cost–benefit analyses for individual organizations may show that it does not make economic sense for them, however. If bargaining and negotiation about how to share the profits of cooperation are not successful, then new partners must be found. A second challenge involves the measurement of the contribution that each participating organization in the supply chain makes to the fulfillment of the supply chain's goal. (The challenges of measuring an individual's contribution to the organization have evolved into measuring an organization's contribution to the supply chain.) To some extent, this is dependent on the synergies that the supply chain can exploit, which involves interorganizational performance measurement. Traditional assessment (and accounting) practices are not suited for such measurement.

Ultimately, the goal of the e-commerce KMS is the same as the goal of the organizational KMS: Optimize customer satisfaction by accessing the appropriate knowledge in a timely manner. Consequently, at least some of the metrics for assessing success are the same. Product cycle times and the satisfaction of customers, employees, and business partners must and can be measured using the previously mentioned metrics.

CONCLUSION

Networked computers are promoting the evolution of a digital economy that is characterized by the following:

- Workers who are mobile and increasingly dependent on knowledge rather than physicality,
- Business partners that are connected to each other along the entire value chain, and
- A competitive environment that rewards speed and innovation.

To be successful managers in this economy, managers must properly manage relationships with employees, business partners, and customers. Managers also must effectively and efficiently manage the vast quantity of data, information, and knowledge generated and required by these participants in the value chain.

A knowledge management system provides an appropriate means of fulfilling the latter. Certainly there are significant challenges and costs involved in implementing a KMS, especially with respect to overcoming cultural barriers and human tendencies to hoard knowledge. The main benefit is substantial, however: An organization's knowledge becomes a strategically significant resource that helps it streamline operations, speed up decision making, and devote more time to proactive strategic activities. The net benefit can be magnified if the KMS is integrated into the IT infrastructure of an e-commerce endeavor, such that the participants in the value chain are able to share and leverage their collective knowledge resources.

In general, the business community is several years away from implementing the vision of knowledge management held by many knowledge management experts, but the promise is there and the technological capacity for fulfilling the promise is evolving. It remains to be seen whether the cultural capacity to embrace knowledge management will also evolve.

GLOSSARY

Business Web (B-Web) A distinct system of suppliers, distributors, commerce services providers, infrastructure providers, and customers that uses the Internet for its primary business communications and transactions (Tapscott, Ticoll, & Lowy, 2000, p. 17).

B-commerce The exchange of economic value that is mediated and enabled by electronic networks and the exchange of information not directly related to the exchange of economic value.

C-commerce Collaborative commerce involving the flow of product, data, and metadata along the entire value chain from supplier to customer and back.

Communities of practice Groups of people who share similar goals and employ common practices, tools, and vocabulary language to fulfill those goals.

Data Facts that have no intrinsic value.

E-commerce The exchange of economic value that is mediated and enabled by electronic networks.

Information Data that is processed and analyzed such that it can be used to help make decisions.

Knowledge, explicit Knowledge that can be articulated in language and documented.

Knowledge, tacit Knowledge that cannot be articulated in language or documented because it is based on the personal experiences and context of the individual who possesses it.

Knowledge management An organizational activity that treats knowledge as an explicit concern of the organization and links positive organizational results to knowledge assets (Barclay & Murray, 1997).

Structured documentation Data and information that are organized or cataloged according to a prescribed schema and stored in applications such as databases.

Unstructured documentation Data and information that are not organized or cataloged according to any schema, such as that in e-mails, letters, and reports.

Value Networks A networked supply chain in which the participating organizations maintain their associations so as to leverage their capabilities and opportunities in a way that attracts customers, keeps customers, and creates associations of mutual benefit (Keen & McDonald, 2000, p. 91).

CROSS REFERENCES

See *Customer Relationship Management on the Web; Data Mining in E-Commerce; Electronic Commerce and Electronic Business; Enterprise Resource Planning; Extensible Markup Language; Fuzzy Logic; Online Communities; Peer-to-Peer Systems; Privacy Law; Supply Chain Management; Virtual Enterprises; Web Search Fundamentals; Web Search Technology.*

REFERENCES

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly 25,* 107–136.

Barclay, R., & Murray, P. (1997). What is knowledge management? Retrieved October 24, 2002, from http://www.media-access.com/whatis.html

Dawson, R. (2000). *Developing knowledge-based client relations: The future of professional services.* Boston: Butterworth-Heinemann.

DeLong, D. W., & Fahey, L. (2000). Diagnosing cultural barriers to knowledge management. *Academy of Management Executive, 14,* 113–127.

Eisenhart, M. (2002). *The human side.* Retrieved October 24, 2002, from http://www.line56.com/articles/default.asp?ArticleID=3545

Keen, P., & McDonald, M. (2000). *The eprocess edge.* Berkeley, CA: Osborn/McGraw-Hill.

Manville, B. (1999). Complex adaptive knowledge management. In Clippinger J. H., III, (Ed.), *The biology of business* (pp. 89–111). San Francisco: Jossey-Bass.

Martin, C. (1999). *Net future.* New York: McGraw-Hill.

Marwick, A. (2001). Knowledge management technology. *IBM Systems Journal, 40*(4), 814–830.

McCoy, F. (2001). Preparing the enterprise chain for c-commerce. Retrieved October 24, 2002, from http://www.techrepublic.com/article_guest.jhtml?id=r00520020201fmc01.htm&fromtm=e105-2

Minkin, B. H. (1995). *Future in sight.* New York: Macmillan.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation.* Oxford: Oxford University Press.

Pepper, S. (2000). The TAO of topic maps. Retrieved October 24, 2002, from http://www.gca.org/papers/xmleurope2000/pdf/s11-01.pdf

Polanyi, M. (1967). *The tacit dimension.* London: Routledge and Kegan Paul.

Stewart, T. A. (2001). *The wealth of knowledge: Intellectual capital and the twenty-first century organization.* New York: Doubleday.

Sveiby, K. E. (1997). *The new organizational wealth.* San Francisco: Berrett-Koehler.

Tapscott, D. (1996). *The digital economy.* New York: McGraw-Hill.

Tapscott, D., Ticoll, D., & Lowy, A. (2000). *Digital capital: Harnessing the power of business webs.* Boston: Harvard Business School Press.

L

Law Enforcement

Robert Vaughn, *University of Memphis*
Judith C. Simon, *University of Memphis*

INTRODUCTION

In the mid-1960s the Internet was first formed using a computer network to link together a small number of universities and military research laboratories in various locations across the United States (Kalakota & Whinston, 1996). The Internet was initially a military communication network established for use by U.S. military personnel, some educational institutions, and defense contractors. The "official" start of the Internet is thought by many to have occurred in 1983 as a result of improved protocols, the establishment of a separate network for the military, and the formation of additional networks (Ruthfield, 1995). Since that time, Internet usage has soared beyond expectations and continues to do so now, along with a related rapid increase in crimes associated with the Internet.

As a new technology, the Internet has exhibited a much greater growth rate than any other recent technology, with the number of host computers connected to the Internet increasing from approximately 1 million in 1993 to 20 million in 1997 (Gallagher, 1999). NUA.com (2002) estimated that worldwide Internet use has grown from 26 million users in 1995 to 580.78 million users as of May 2002, with Internet-generated revenues increasing from US$8 million in 1994 to US$1,234 billion in 2001. NUA.com is an Irish Web site, where "nua" means "new." The NUA Web site claims to be "the world's leading resource for Internet trends and statistics" and estimates that over "200,000 people in more than 140 countries read" their news and analysis off the Internet every week. InterGov International (2002) reported 574 million Internet users for 2001 with a projected 741 million users in 2002. Gallagher (1999) reported that the level of economic development reflects the amount of Internet use within a particular country.

The dramatic increase since the early 1990s in the number of persons using the Internet, both for personal and for business reasons, has resulted in changes in the manner that business is conducted worldwide. This includes the business of law enforcement and the negative consequences resulting from the escalation of criminal activities over the Internet and the resulting jurisdictional issues, conflicts, and confusion. As the lawful use of the Internet has increased, so too has the unlawful use of the Internet, resulting in an increased assortment of new crimes and changes in the modus operandi of traditional crimes, such as money laundering, fraud, drug trafficking, and even murder. The communication capabilities, the concealment facet, and the global nature of the Internet can be used advantageously for criminal purposes. The association with increased criminal activity was characteristic of the telephone, the mobile telephone, and other technologies, and now of the Internet. The Internet has enabled new avenues for the commission of both personal and property crimes on a global scale that have left law enforcement personnel worldwide searching for legal solutions with which to address the emerging Internet crimes.

Although the Internet has enhanced law enforcement information sharing and law enforcement related to electronic commerce ("e-commerce"), it has also increased issues regarding legality, privacy, and security, as well as providing a new medium of communication that can be exploited by those with criminal intentions. In this increasingly technological world, law enforcement and security personnel face new and ever-changing challenges to effectively respond to increased criminal exploitation of the Internet, technological advances, and increased globalization.

The sections below discuss ways in which law enforcement has taken advantage of the capabilities of the Internet, as well as new criminal activities that law enforcement must handle due to the Internet.

LAW ENFORCEMENT AGENCIES' WEB SITES

Information on the Internet regarding any law enforcement agencies that have a public Web site is accessible globally. For reasons of ongoing investigations, sensitivity of the type of investigation, victim protection, and other legalities, not all law enforcement sites or their subsites on the Internet are accessible to the public at large.

Currently, law enforcement Web sites are numerous and easily found using Web browsers to search for any city, town, county, borough, metropolitan, federal, state, country, or international law enforcement agency. Some Web sites have compiled Web addresses and links to law enforcement agencies worldwide, making it even easier to find multiple law enforcement agencies. Businesses and individuals now have ready access to information regarding law enforcement agencies, their jurisdictions, addresses, phone numbers, and e-mail information, which allows for increased information exchange between the public and law enforcement. The Bureau of Justice Statistics site references law enforcement information and reported that in 1996 there were a total of 738,028 full-time sworn law enforcement officers and over 36,800 different law enforcement agencies, including local police, state police, special police, Texas constables, and Federal personnel, within the United States (http://www.ojp.usdoj.gov/bjs/lawenf.htm). Exact figures for law enforcement organizations worldwide are not known since not all countries make that information available, but the Internet has made access much easier to the large number of law enforcement organizations in existence.

U.S. and Canadian Agencies' Web Sites

The United States and Canada appear to have the most Web sites for law enforcement agencies, although this is not empirically known. This is intuitive primarily because of the 181.23 million Internet users in the United States and Canada, compared with Europe's 171.35 million users, Asia/Pacific's 157.49 million users, Latin America's 25.33 million users, the Middle East's 4.65 million users, and Africa's 4.15 million users (NUA, 2002). The continued growth of the Internet is such that a survey of specific types of sites would be but a snapshot of the then-current status of the Internet.

Directories of law enforcement agencies and related organizations in many countries can be accessed using browser searches (http://search.officer.com/agency search/, http://www.info-s.com/police-d.html, http://info-s.com/laworg.html). FirstGov (http://www.firstgov.gov) is a good locator Web site for both federal and state government law enforcement sites, which in turn have other links to multiple law enforcement sites. State trooper official sites are listed on the Internet (http://www.statetroopersdirectory.com/). U.S. agencies such as the Federal Bureau of Investigations, the Central Intelligence Agency, the Drug Enforcement Agency, and the Bureau of Alcohol, Tobacco, Firearms, and Explosives all have Web sites accessible by the public that contain a wide variety of information pertaining to law enforcement, including law enforcement on the Internet.

Agency Web Sites in Other Countries

Law enforcement agency Web sites are dependent upon the technological infrastructure available within any particular country. Underdeveloped countries may not have any law enforcement agencies with accessible Web sites, while well-developed countries would be expected to have numerous law enforcement agency Web sites. As infrastructure and other needed technologies become available within underdeveloped countries, the number of Internet law enforcement listings is expected to increase.

Law enforcement agencies' Web sites worldwide can be accessed to obtain information on their agencies. Examples include countries such as Russia (http://www.mvdinform.ru/index.php), Pakistan (http://www.sindh police.gov.pk), Hong Kong (http://info.gov.hk/police/index.htm), and Iceland (http://www.logreglan.is/displayer.asp?catid=217). The U.S. Department of Justice Bureau of Justice Statistics developed a factbook, *The World Factbook of Criminal Justice Systems*, which gives the descriptions of criminal justice systems in 42 countries as of 1993 (http://www.ojp.usdoj.gov/bjs/abstract/wfcj.htm).

International Agencies' Web Sites

International agencies are not as numerous because they represent larger segments of the population in various regions of the world. Also, they are dependent upon the technology available in each region, which in turn determines the capability to establish and maintain Web sites. Regardless of this limitation, there are several international police agencies that can be accessed via the Internet (http://www.policeinternational.com, http://www.interpol.int) including a Universal Police site (http://www.spinet.gov.sg/pollink/index.html) that links to law enforcement agencies in 20 countries and Interpol. There are also Internet listings of international government, military, and intelligence agencies that can be accessed (http://www.sagal.com/ajax, http://www.gksoft.com/govt/en/world.html, http://www.police.uk, http://www.polfed.org/main_frame.htm).

InterGov International (http://www.intergov.org), formed in 1986, operates under the supervision and guidance of staff and management brought together from countries around the world with the goals of protecting, informing, and serving persons within the Internet community. Associated with InterGov International is the International Web Police (http://www.web-police.org), which claims recognition internationally as the Internet Police Authority since 1986. Its sole mission is to protect users of the Internet from the vast amount of criminal activity present on the Internet. Both InterGov International and International Web Police are nonprofit organizations providing services to Internet users globally.

Web Sites of Agency Divisions and Bureaus

Information concerning virtually every conceivable law enforcement division or bureau can be found on the Internet, including airport, auto theft, aviation, bomb squad, burglary, canines (http://www.policek9.com), child abuse, community policing, corrections, criminal intelligence, crime prevention, crime scene, crime stoppers (http://www.c-s-i.org/), crisis intervention, domestic violence (http://info-s.com/domestic.html), firearms instruction (http://www.ialefi.com/), forensics, hostage negotiation, internal affairs, juvenile justice, narcotics, organized crime, patrol, robbery, security squad, sex crimes, transit, and vice.

WEB SITES FOR SPECIFIC LAW ENFORCEMENT INTERESTS AND CONCERNS

Numerous Web sites are available for various law enforcement needs and interests. Several of these areas of interest are mentioned in this section, providing examples of the range of topics available through the Internet.

Government Resources

Access to a wide variety of government resources is available through the Internet. The National Criminal Justice Reference (http://www.ncjrs.org/lewww.html) allows access to law enforcement resources related to corrections, courts, drugs, crimes, juvenile justice, statistics, victims of crime, and international interests. The U.S. Department of Justice Bureau of Justice Statistics Web site (http://www.usdoj.gov/bjs) reports statistics concerning crime, crime victims, criminal offenders, and special topics such as homicide trends, firearms crimes, drug crimes, and international justice statistics.

Grants and Funding

Grants and assorted funding agreements are available to assist the public in their efforts to fight crime and protect the public welfare. The U.S. Department of Justice is involved in community-oriented policing services (COPS). COPS grants have provided communities with funds for hiring and redeploying police officers into communities. Other COPS initiatives include targeting specific crimes, such as school-related crimes, methamphetamine labs, and domestic violence, as well as cultivating partnerships and community government. COPS grant and funding information can be accessed at the U.S. Department of Justice Web site (http://www.cops.usdoj.gov).

Grants are currently offered by the U.S. Department of Justice Office of Justice Programs (http://www.opj.usdoj.gov/fundopps), including grant programs for law enforcement assistance, technology, anti-terrorism, juvenile justice, community-based initiatives, violence against women, traffic safety, research, and statistics. Through these grants, funding is available for law enforcement efforts addressing specific concerns such as safe schools, sex offender management, crime victims, state domestic preparedness, sexual assaults, domestic violence, and "weed and seed." Weed and seed grants fund law enforcement efforts to prevent, control, and reduce violent crimes, drug trafficking, drug abuse, and gang-related activities in specific targeted high-crime neighborhoods throughout the United States. The targeted neighborhoods range in size from several city blocks to as large as 15 square miles.

From 1981 through 2000, the National Center for Injury Prevention and Control reported that motor vehicle crashes were the leading cause of death in the United States for those between the ages of 4 to 35 (NCIPC, n.d.). Peoria, Illinois, for a three-year period 1994 through 1996, experienced a significant reduction in both violent and property crimes, and traffic crashes, which was the direct result of establishing traffic enforcement as a first priority (http://www.nhtsa.dot.gov). "The Peoria Experience" demonstrated the importance of traffic enforcement, not only to reduce traffic injuries and fatalities, but also to significantly reduce the crime rate in the targeted area. While Peoria relied primarily upon internal sources of revenue, the U.S. Department of Justice offers funding assistance to law enforcement agencies for traffic enforcement programs in an effort to reduce traffic fatalities and crime (http://www.nhtsa.dot.gov/nhtsa/whatsup/tea21/index.html).

Grants and funding information regarding juvenile-related issues are available through the Office of Juvenile Justice and Delinquency Prevention (http://ojjdp.ncjrs.org/grants/grants.html). This office offers funding for missing and exploited children programs, juvenile accountability incentives, programs combating underage drinking, juvenile mentoring programs, and drug-free communities support programs.

Forensics

The American Academy of Forensic Sciences (http://www.aafs.org), with over 5,000 members, is a professional society dedicated to the use of science to further law enforcement efforts. Members come from a diverse variety of professions such as physicians, dentists, physical anthropologists, toxicologists, attorneys, psychiatrists, document examiners, engineers, criminology personnel, and educators. These members represent over 50 countries, including the United States and Canada.

The Kentucky State Medical Examiners Office and the Justice and Safety Center in the College of Justice and Safety at Eastern Kentucky University developed a Web site (http://www.unidentifiedremains.net/) designed to identify human remains within the state. The Michigan State Police have a similar site (http://www.msp.state.mi.us/persons/remains.htm) that shows reconstructions produced by putting clay onto the actual skulls of unidentified victims in order to attain renderings that would represent the individuals as they would have appeared before their deaths. In Atlanta, Georgia, the Fulton County Medical Examiner's Center Web site (http://www.fcmeo.org/UIDPage.htm) posts descriptions, clay reconstructions, sketches, and photographs of unidentified deceased persons in an attempt to gain identification of those persons by someone visiting the site in search of missing relatives.

The DOE Network (http://www.doenetwork.org) is a volunteer organization dedicated to the identification of cold case victims and missing persons in North America, Europe, and Australia. The name "DOE" comes from law enforcement usage of "John Doe" or "Jane Doe" in reference to unidentified persons, which in the case of the DOE Network relates to unidentified human remains. The DOE Network Web site uses only reconstructed images and sketches of both unidentified adult and juvenile human remains found throughout the United States, Europe, and Australia, along with known law enforcement case information available for each unidentified victim presented. The DOE Network focuses on cold cases that have received little public attention in recent years, as well as a few well-known cases where little evidence is available. The criteria for unidentified cases are that the victim died prior to or during 1999 in Europe, Canada, or the United States; the case is on file with a law enforcement agency; and there is a picture or reconstructed image of the victim available. Unexplained disappearance cases prior to 1993 are also featured on the site.

The Missing Persons Cold Case Network website (http://www.angelfire.com/mi3/mpccn/index.html) is a sister site of the DOE Network and provides information for more than 3,000 missing person cases that have been active for six or more months. The primary requirements for listing missing persons on the site include cases must originate from within the United States, cases must be on file with and submitted by a law enforcement agency, and cases must have been active for at least six months prior to the request for a missing person listing.

Crime Scenes

It has always been crucial to any investigation to protect the scene of the crime. The Internet has enabled increased information sharing concerning crime scene preservation and training. It has become increasingly important to protect personnel investigating crime scenes because of the risks associated with body fluid contaminations often found at crime scenes. In light of this increased risk of exposure to body fluids, pathogens, and communicable diseases, a crime scene safety handbook was produced by the FBI and is accessible on the Internet (http://www.fbi.gov/programs/lab/handbook/safety.htm). Other sites, including numerous medical sites, present information related to contamination safety regarding body fluids and other biohazards.

Biological and Chemical Acts

Terrorism acts and the threat of terrorist acts are an increasing global concern for law enforcement and the general public. This has led to an increased interest regarding the possibility of biological and chemical acts of terrorism. The Henry L. Stimson Center was founded in 1989 as a nonprofit institution devoted to international security and peace (http://www.stimson.org/cbw/?SN=CB2001121259). Johns Hopkins University hosts a Center for Civilian Bio-Defense Strategies Web site (http://www.hopkins-biodefense.org).

The U.S. Department of Health and Human Sciences Center for Disease Control and Prevention (CDC) Web site provides a list of biological diseases and agents according to category and level of threat of the organism and a list of chemical agents that pose threats to society. Other relevant information is also available on the home page of the CDC (http://www.cdc.gov) and through MEDLINEplus (http://www.nlm.nih.gov/medlineplus), which posts information and links to numerous other related government sites. Information used by MEDLINEplus is obtained from the world's largest medical library, the National Library of Medicine (http://www.nlm.nih.gov).

Other Forms of Terrorism

Since September 11, 2001, the United States and other countries throughout the world have increased their focus on the issues of terrorism. The U.S. Department of State (http://usinfo.state.gov/topical/pol/terror/) has made information available regarding terrorism. For example, in 2001 there were over 3,700 deaths associated with 348 terrorist attacks. In 2000, there were 409 deaths that resulted from 426 terrorist acts. The increase in the number of deaths in 2001 resulting from terrorist acts can be attributed to the terrorists' acts at the New York City World Trade Center.

The United Nations hosts a Web site (http://www.un.org) regarding international terrorism and global resolutions and declarations against acts of terrorism worldwide. The Federal Emergency Management Agency, through a virtual library and electronic reading room (http://www.fema.gov), presents a fact sheet regarding terrorism acts before, during, and after they occur that is relevant to public safety and the safety of individuals targeted by such terrorism. The Federation of American Scientists (http://www.fas.org/index.html) also hosts a site that has information concerning terrorism, especially as it relates to nuclear terrorism and dirty bombs.

In 1994, the University of St. Andrews in the United Kingdom established a Center for the Study of Terrorism and Political Violence (CSTPV) (http://www.st-and.ac.uk/academic/intrel/research/cstpv/). The purposes of the center were to investigate political violence, to develop a body of theories encompassing the various dissimilar elements of political violence, and to examine the impact of violence and resulting responses to violence at the various levels of society and government worldwide.

A University of Pittsburgh Web site (http://jurist.law.pitt.edu/terrorism.htm) provides information on terrorism laws and policies. International terrorism information is also attainable at the International Policy Institute for Counter-Terrorism Web site (http://www.ict.org.il). In the United States, the Central Intelligence Agency (http://www.cia.gov/terrorism) and the Federal Bureau of Investigations (http://www.fbi.gov/terrorinfo/terrorism) both present information pertaining to terrorism and will respond to requests for information from anyone regarding terrorism and suspected terrorists. Domestic security information can be found at various U.S. sites such as the State of Florida Web site (http://www.myflorida.com/myflorida/domestic_security/index.html) and a New Jersey state Web site (http://www.state.nj.us/lps/dsptf/dsptfhome.html). Other state Web sites also contain domestic security-related information.

The U.S Department of State Web site (http://www.state.gov) maintains up-to-date information for international travelers regarding warnings and available assistance, in addition to information on terrorism and counterterrorism.

Drug Abuse

Drug abuse affects persons worldwide and is often associated with increased crime rates as a result of drug abusers striving to obtain more drugs as their addiction increases. Drugs are sold openly over the Internet, primarily from countries where drug laws are liberal and drugs can be readily shipped internationally without consequences, except to the recipient.

The Drug Enforcement Agency (DEA) Web site (http://www.usdoj.gov/dea) has a large volume of drug-related information. The DEA Web site includes lists of a wide variety of drugs, with detailed information concerning drug effects.

The National Institute on Drug Abuse (NIDA) Web site (http://165.112.78.61/DrugAbuse.html) has information about common drugs of abuse including LSD, alcohol, cocaine, club drugs, inhalants, heroin, marijuana, MDMA, methamphetamine, nicotine, PCP, prescription drugs, and steroids. The NIDA site also provides information concerning drug abuse, prevention and testing for drug abuse, and statistics related to drug abuse trends and costs.

The National Criminal Justice Reference Service Web site (http://virlib.ncjrs.org) hosts a variety of articles and information related to drug abuse.

The California Department of Justice Web site (http://www.stopdrugs.org/identification.html) has information that aids in the identification of unknown drugs and pills and also describes an assortment of drug types and the effects of those drugs. This Web site also presents pictures of drugs and associated drug paraphernalia. The Orange County, CA, Sheriff's Office (http://www.ocso.comocso.com/sid/narcotics/identification/drugs.html) prepared an educational photo album of drugs that is beneficial to parents and others attempting to identify common street drugs that may be abused by children, friends, or family.

Locum International Group (http://www.locum.co.il/druglist/index.php) is an organization focused on generic drug development needs. Its Web site contains pictures and descriptions of numerous generic drugs along with the manufacturers' information. A related Web site (http://www.iagim.org) belonging to IAGIM is an international nonprofit pharmaceutical drug development association that has also made information available related to drugs internationally.

Investigative Resources

A major benefit of the Internet to law enforcement organizations worldwide is the vast amount of investigative resources available over the Internet that can provide assistance to investigators. The *CIA Factbook* Web site (http://www.odci.gov/cia/publications/factbook/index.html) is one example of a good source of facts and information regarding virtually any country in the world.

Other Internet resources available for investigative or informational purposes include adoption records and searches, background checks, credit information, college and university phonebooks, firearms and firearms identification, fugitives, genealogy, handwriting examinations, mapping, military personnel, organized crime and corruption, political, public records, Regional Information Sharing Systems (RISS) (http://www.iir.com/riss), serial killers, telephone and e-mail information, telephone reverse lookup, vital records, unclaimed property, and urban legends. Other Internet-available information includes prison (http://www.soci.niu.edu/~critcrim/prisons/prisons.html), corrections (http://www.officer.com/correct.htm), and sentencing (http://www.ojp.usdoj.gov/bjs/sent.htm).

Training and Education

While some law enforcement agencies provide their own facilities for training police personnel, there are numerous other training facilities available for such training including both government and privately owned facilities and programs. Many universities also offer training for police personnel. For instance, the Police Training Institute at the University of Illinois was established in 1955 and has since trained over 58,000 law enforcement personnel (http://www.pti.uiuc.edu/). The FBI Academy offers advanced training for police personnel in a wide variety of areas (http://www.fbi.gov/hq/td/academy/academy.htm). Many schools are increasing the online training and educational courses accessible through Web sites. In the field of law enforcement, this trend allows officers to take courses while on or off duty without the necessity of the officers having to physically attend classes.

Criminal justice education courses are offered by most colleges and universities. Many also offer advanced criminal justice degrees at the masters and doctoral levels. Florida State University, through its School of Criminology and Criminal Justice, has offered courses in criminology since 1952 with the mission of providing national and international leadership on issues related to crime and the response of society to crime (http://www.criminology.fsu.edu/). The Northern Michigan University Department of Criminology (http://www.nmu.edu/cj/default.htm) offers undergraduate and graduate degrees in criminology as does the University of Memphis (http://www.memphis.edu). Rated #1 in nine national surveys was the University of Albany Hindelang Criminal Justice Research Center program that leads to careers in research, higher education, and management positions in criminal justice (http://www. albany.edu/hindelang/). These colleges and universities are representative of the many colleges and universities that offer degree programs and training in criminology and law enforcement.

Professional Associations and Memorials

There are numerous Web sites with information regarding law enforcement professional associations and memorials. For example, the National Law Enforcement Officers Memorial Fund (http://www.nleomf.com) has a Web site.

The National Association of Chiefs of Police (http://www.aphf.org/nacop.html) is a tax-exempt, nonprofit educational association of law enforcement command officers with 60,000 members within the United States. Membership is open to any command law enforcement officer and directors of U.S. security agencies.

The International Association of Chiefs of Police (http://www.theiacp.org/) has over 19,000 members in more than 100 countries and states that it is the world's largest and oldest nonprofit police executive organization.

The International Association of Women Police (http://www.iawp.org) is an organization of law enforcement women that strives to ensure equity for women serving in law enforcement agencies.

The National Center for Women and Policing (http://www.womenandpolicing.org) advocates increasing the number of women actively involved in law enforcement across all ranks as an approach for enhancing responses to crimes and violence against women, for decreasing incidents of police brutality, and for stronger community policing initiatives.

The Police Executive Research Forum (PERF) (http://www.policeforum.org) is an organization of leading police executives representing the largest law enforcement agencies at the city, county, and state levels that is committed to improved policing and enhanced professionalism through research and public policy debate involvement.

The Community Policing Consortium comprises the International Association of Chiefs of Police (ICAP), the National Organization of Black Law Enforcement Executives (NOBLE; http://www.noblenatl.org/), the National Sheriff's Association (NSA), the Police Executive Research Forum (PERF), and the Police Foundation (http://www.communitypolicing.org). The purpose of the consortium is to further the development of community policing research and to provide training and assistance. The consortium is supported by the U.S. Department of Justice Office of Community Policing Services (COPS).

The Fraternal Order of Police (http://www.grandlodgefop.org), founded in 1915, has over 2,100 lodges and almost 300,000 members and is the world's largest organization of sworn law enforcement personnel. The Fraternal Order of Police is dedicated to improving the working conditions and the safety of law enforcement officers through "education, legislation, information, community involvement, and employee representation."

Death Penalty

Many Web sites (e.g., http://www.info-s.com/deathpenalty.html) that relate different viewpoints regarding the issue of the death penalty exist. For example, a major concern on either side of the issue is the execution of a death sentence upon a person who is actually innocent of the accused crime.

Some sites provide data related to the death penalty. According to a University of Alaska Justice Center Web site, "by February 1999, 67 countries had abolished the death penalty entirely and 14 had abolished it for all but exceptional crimes such as crimes under military law or crimes committed in exceptional circumstances such as wartime" (University of Alaska at Anchorage Justice Center, 2000). Amnesty International lists countries where death penalty executions were conducted (http://web.amnesty.org/pages/deathpenalty_index_eng?OpenDocument).

COMMUNICATION ACTIVITIES

The Web sites described in the previous section provide access to online information, but law enforcement agencies have additional communication activities that use Internet technology.

Communication among law enforcement agencies, other public service agencies, and the public is important to the public welfare. Also, the public has the right to be provided with information relative to public safety. Worldwide communication with law agencies can now be conducted over the Internet at a lower cost than international telephone rates. E-mail communications also reduce the cost of communicating on a global scale for law enforcement organizations and personnel.

The Internet is a communication medium through which information in the form of text, audio, voice over IP (the transmission of voice communications using the Internet protocol for data transmission), graphics, videos, and live video transmissions can be transferred. Mug shots, drivers' license photographs, and crime scene photographs and sketches can be e-mailed anywhere in the world on an as-needed basis at any time, as requested by other law enforcement or governmental agencies. The Internet is also a good communication medium for teleconferencing purposes that enables conferencing among people located remotely from one another with both audio and visual connectivity in real time, which gives the illusion of face-to-face conferencing.

Prior to the explosion of Internet use, most law enforcement-related information was primarily transmitted via telephone, Teletype, fax, postal mail, and secure direct connections to centralized mainframes. Therefore, most agencies were rather isolated, especially regarding international crime-related information. Today, the amount and variety of law enforcement information available over the Internet are much more extensive and accessible.

Information Sharing

The capability to share information is one of the most significant aspects of the Internet that has affected law enforcement, as evidenced by the volume of information regarding law enforcement now available. Previously, this information could only be shared on an almost strictly individual or agency-to-agency basis, with minimal information sharing with the general public.

Some sites, such as Police-L (http://www.police-l.org), are restricted to use by sworn law enforcement officers and serve as a forum for officers to exchange information regarding practices and procedures or issues of interest or concern, as well as to communicate freely in a police-only setting. Others are public service sites, such as the site at http://www.copnet.org/pubserv.html that serves to link the community with police as a united front to crime. Many law enforcement personnel are building their own personal Web sites for exchanging information, with some of the sites publicly accessible and others requiring user names and passwords and primarily for use by other law enforcement persons.

The most important information-sharing capability is that of law enforcement access to information contained within databases for law enforcement and criminal justice purposes. In the United States, law enforcement and criminal justice agencies, in addition to some authorized noncriminal justice agencies, can access information contained in multiple databases through the National Law Enforcement Telecommunications System (NLETS) and the National Crime Information Center (NCIC).

NLETS links local, state, and federal agencies and provides nationwide interstate communications for the purpose of exchanging criminal justice information. This computer controlled message switching system is owned, managed, and operated by the participating states (NLETS, n.d.). The NLETS allows access to drivers license information, vehicle registration information, boat/snowmobile registration information, vehicle identification number (VIN) data from the National Crime Bureau (NICB), Federal Aviation Administration (FAA) aircraft data, Canadian information, state and local criminal history information, state road and weather condition reports, hazardous materials information, National Center for Missing and Exploited Children information, International Criminal Police Organization (INTERPOL) information, El Paso Intelligence Center narcotics, alien, and weapons smuggling information, sex offender information, and Alcohol, Tobacco, and Firearms (ATF) Tracing System information. NLETS also allows for point-to-point administrative messages, statewide broadcasts, regional broadcasts, all-points broadcasts, and hit confirmation.

The Federal Bureau of Investigations maintains person and property data in the NCIC with allowed access made available to Federal Criminal Justice agencies, the 50 states, the District of Colombia, Puerto Rico, U.S. possessions, U.S. territories, and Canada. The NCIC provides stolen vehicle, license plate, boat, article, and securities information along with stolen/recovered gun information. It also contains information on wanted, missing, and unidentified persons and information on violent gangs and terrorists, violent felons, foreign fugitives, deported felons, and U.S. Secret Service Protective file information.

Chat Rooms and Message Boards

Chat rooms and message boards are available to assist in information exchange by law enforcement personnel and the general public. Chat rooms and message boards allow persons to ask questions, discuss issues, and post messages. An added feature of chat rooms is that they allow for real-time conversation and information exchange. CopsOnLine (http://www.copsonline.com) and Police World (http://www.policeworld.net) both have chat rooms and message boards available for anyone to post questions and answers or to chat with others about areas of interest.

E-mail

The use of the Internet for law enforcement e-mail allows for quick and rich communication. Color crime scene and suspect photos can be transmitted easily between law enforcement agencies. For suspects with no prior arrests, driver's license photographs can be obtained for photo lineups. For suspected serial crimes, whether committed nationally or internationally, crime scene photographs can be transmitted for crime scene comparisons, enabling law enforcement personnel to determine whether crimes are related.

Communication Security Issues

A large volume of information is readily available to law enforcement agencies and the general public, except where restricted due to reasons such as (1) the information is sensitive, (2) the investigation is ongoing, or (3) law restricts the information. Virtual private networks, encryption, and secure passwords protect some information-sharing connections so that the general public or persons with criminal intent cannot gain access to any sensitive or critical crime information.

Security of communications for law enforcement agencies is an important issue due to the nature of most interagency and intra-agency communication restrictions placed upon the transmission of sensitive law enforcement information. Communication of information stored in federal databases is restricted to communication across land telephone lines (physically wired telephone lines) or other hard-wired connections (point-to-point physically wired data communication networks). Although these restrictions serve to protect the integrity and security of communicated information, it restricts the type of information that can be transmitted via wireless Internet connections to officers using laptops in patrol units that need to access criminal history information in addition to a wide variety of other restricted information available via NLETS, "a sophisticated message switching network linking local, state, and federal agencies together to provide the capability to exchange criminal justice and public safety related information interstate" (NLETS, n.d.). The transmission of such information on the Internet over hard-wired lines is restricted because of the risk that it might be intercepted by unauthorized persons.

Virtual private networks (VPNs) may offer a solution to secure communications over the Internet. A VPN ensures point-to-point communication of messages and protects against interception by unauthorized persons through the use of a combination of firewalls, encryption, and authentication. The use of encryption alone, depending on the strength of the algorithm used, would provide security for law enforcement communications across the Internet, while also addressing issues of confidentiality and privacy whether the communication occurred over hard-wired or wireless communication channels because the communication could not be decrypted even if intercepted by unauthorized persons. In like manner, those with criminal intent can also use encryption and VPNs to secure their Internet communications. This would allow criminals to exchange communications with impunity over the Internet without concern for law enforcement interception and decryption.

Digital signatures have yet to be established for the communication of law enforcement-related information and for authenticity verification of senders and receivers to resolve issues of privacy and security. Oftentimes the requesting agency is first required to fax their request using a departmental coversheet before the requested agency will send any information over the Internet. Even then, the information may be restricted, because certain information is commonly restricted to transmission only over hard-wired communication media. With the advent of wireless networks, transmission of such restricted information is raising legal issues. Regardless of these information restrictions, the Internet has impacted law enforcement by enabling significant changes in the manner that law enforcement agencies share information both publicly and privately.

The primary advantages of the Internet for law enforcement agencies have been that of communication, information sharing, and access to information contained in shared databases. This focus is gradually shifting to include the e-commerce capabilities of the Internet.

LAW ENFORCEMENT USE OF E-COMMERCE

Kalakota and Whinston (1996) indicated that the term "e-commerce" was related to the buying and selling of products, services, and information over computer networks today and in the future by means of the Internet. Within a few years of that statement in 1996, the Internet has become the major technology for e-commerce.

Law enforcement has also expanded e-commerce on the Internet, as evidenced by the myriad of commercial Web sites selling law enforcement-related products, services, training, and information. Virtually any police-related product can be obtained at any of the numerous commercial police supply companies that have established Web sites.

Among the many services being offered over the Internet for law enforcement, one is significant in that it is enabling law enforcement to establish Web sites for the purpose of collecting traffic fines and fees (http://www.ezgov.com/newsfullpage.jsp?ArticleID=3684&category=pressreleases, http://www.cobbstatecourtclerk.com/). The Courts of New Zealand opted to have their own Web site developed for collecting fines online (http://www.fines.govt.nz). In addition, there are several commercial companies that offer services to governmental agencies that establish, operate, and maintain fee and fine-collecting Web sites.

CRIME AND THE INTERNET

The Internet provides both a means for law enforcement to obtain and report data regarding criminal activities and a way for criminals to conduct their activities. Both are discussed in this section.

With the rapid growth of the Internet has come an increased risk of exposure to illegal activities and criminal elements seeking to use new technologies to exploit vulnerabilities in those technologies (Macodrum, Hasheri, & O'Hearn, 2001). This exploitation of the Internet has not only resulted in new ways to commit "traditional" crimes but also in entirely new crimes made possible because of the Internet and new technologies.

Law enforcement has the responsibility to enforce laws that serve to protect persons, and this includes protection from those seeking to take advantage of vulnerabilities for unlawful gain at the expense of other persons or businesses.

International Crime

The U.S. International Crime Control Strategy as of May 1997 was to obtain an environment whereby the global economy and global trading would increase, where respect for human rights and democratic norms would be more widely accepted, and where drug trafficking, terrorism, and international crime would not be allowed to undermine global economic stability and destroy peaceful worldwide relations (NSC, 1997).

International crime has increased dramatically in the wake of globalization resulting from the increased global usage of the Internet by both individuals and businesses. International criminals conduct a wide range of crimes using the technology of the Internet. These crimes include fraud; extortion; counterfeiting; espionage including economic, military, and governmental; drug trafficking; terrorism; smuggling; money laundering; bribery; intellectual property thefts; slavery; and child pornography. Borders have no meaning to international criminals in conducting their activities but serve to hide and protect them from retribution for their crimes. The Internet has served as a communication medium through which criminals can form alliances and expand their realm of control on a global scale (NSC, 1997). The United Nations Office on Drugs and Crime Web site (http://www.unodc.org/unodc/crime_cicp_sitemap.html) provides information regarding international terrorism, corruption, organized crime, and trafficking in human beings. One of the international crime trends is the expansion of computer and high-tech crimes and the resultant merging of powerful organized crime groups in various parts of the world such as Russia, Asia, Europe, Latin America, and Africa (NSC, 1997).

Web Sites for Crime Statistics

Crime statistics for the United States are available through the Federal Bureau of Investigations Uniform Crime Reports Web site (http://www.fbi.gov/ucr/ucr.htm). Information regarding the National Incident-Based Reporting System (NIBRS) is available on the same Web site, in addition to hate crime statistics and information pertaining to law enforcement officers killed and assaulted.

The U.S. Department of Justice also maintains available crime statistics on its Web site (http://www.usdoj.gov). The Fed Stats Web site (http://www.fedstats.gov/agencies) provides links to an assortment of official statistical information for more than 70 Federal agencies that is publicly available from the Federal Government. The University of Maryland Libraries Web site (http://www.lib.umd.edu/MCK/STATS.html) has links to a large number of statistics sources including the U.S. Government, other countries, the United Nations, intergovernmental, nongovernmental, academia, search engines, and metasite sources, some of which contain crime statistics.

Internet Crime

Internet-related crimes continue to rise as Internet usage expands and as criminals, or those with criminal intentions, increase the scope and means of criminal activities conducted over the Internet.

Internet Crime Statistics

Crime statistics for Internet-related crimes in 2001 were reported by InterGov International as follows: Fraud (scams) 26%, child pornography 17%, stalking 11%, e-mail abuse 9%, harassment and threats 9%, hacking or viruses 9%, child related 6%, copyright violations 4%, terrorism 3%, chat room abuse 2%, and other 4% (InterGov, 2002). Some of these crimes, as well as other concerns, are discussed in the following paragraphs.

Fraud and Scams

The Internet Fraud Complaint Center (IFCC) represents a partnership between the National White Collar Crime Center and the Federal Bureau of Investigation and has published its first-ever IFCC 2001 Internet Fraud Report (IFCC, 2001). The Internet Fraud Complaint Center Web site received 49,711 complaints during 2001 related to Internet crimes such as complaints of fraud, computer intrusions, spamming, unsolicited e-mail, and child pornography. Of the complaints received, 42.8% were related to Internet auction fraud, 20.3% were related to payment account and nondeliverable merchandise, and 15.5% were related to the Nigerian Letter fraud (http://www1.ifccfbi.gov/strategy/statistics.asp). This last one is a fraud scheme that combines the threat of impersonation fraud with a modified advanced payment scheme where a supposedly government official is attempting to transfer illegally out of Nigeria with an alleged millions of dollars, but is in need of monetary assistance until the alleged funds become available after escaping from Nigeria, at which point the alleged millions will be shared (FBI, n.d.; NCIS, n.d.).

Cyberstalking

Prior to the Internet, cyberstalking did not exist. The first complaints of Internet stalking made to law enforcement were met with blank stares, laughs, or comments suggesting that the complainant turn the computer off (Hitchcock, 2000). Hitchcock (2000) went on to define cyberstalking as an online confrontation that spirals out of control, resulting in victims fearing for their lives.

Exploitation of Children

Baker (1999) reported that the FBI is catching sexual predators for soliciting minors over the Internet and that the FBI Innocent Images program has been successful. The Internet allows anonymous communications and easy access to children surfing the Web and talking in chat rooms anywhere worldwide. There is no way of determining the number of children who have been lured over the Internet to meet a predator, nor is it known what has happened to those children who are now missing. Dube (1999) reported that the National Center for Missing and Exploited Children was to create an International Center for Missing and Exploited Children with offices in London and Washington, DC, in order to establish a global network whose purpose would be to present a more uniform response to the worldwide problem of missing children. The present mission of the International Center for Missing and Exploited Children is stated as "to provide a coordinated, international response to the problem of missing and exploited children" by establishing a global network of participating countries that includes the United States, the United Kingdom, South Africa, the Netherlands, Italy, Canada, Argentina, Brazil, Chile, Belgium, and Australia (ICMEC, n.d.).

Child Pornography

The U.S. Customs CyberSmuggling Center's Child Exploitation Unit (http://www.cbp.gov/xp/cgov/enforcement/investigative_priorities/c3/child_exploitation.xml) is seeking to address the rapidly growing presence of child pornography on the Internet. Duffield (2000) stated that the Internet is being used to peddle and distribute pornography, including child pornography, worldwide, which has created the necessity of an international approach to child pornography. In 1997 the Swedish National Crime Intelligence Division was given the task of establishing centralized image storage for known child pornography images that had been seized, whether from over the Internet or in actual physical raids. This single library of child pornography images has since grown to include over 150,000 images from several countries that include Norway, Germany, France, England, Finland, Denmark, Switzerland, Portugal, Belgium, Ireland, and The Netherlands (Duffield, 2000). In the United States, the FBI Online Child Pornography Innocent Images National Initiative, which is a part of the Crimes Against Children (CAC), is a proactive, intelligence-driven investigative initiative with the goal of combating the proliferation of child sexual exploitation and child pornography on the Internet (http://www.fbi.gov/hq/cid/cac/innocent.htm).

Hacking and Viruses

Bequai (2000) stated that viruses and hacking are routine aspects of the Internet and that new solutions would be needed to deal with this area of Internet fraud. Hacking is done for a variety of reasons, such as for political purposes, to gain publicity, to point out security weaknesses, for grudges, and for profit. Hacking can also be performed through the placement of viruses across the Internet in order to disrupt information exchange across computers accessible to the Internet, whether it is for the purpose of disrupting e-commerce or simply for the satisfaction of disrupting communications or destroying data.

Theft

There are many forms of theft, including monetary, intellectual, goods, services, identity, and confidential information. Espionage once mainly targeted governments to gain secrets that could be sold to other governments. With the Internet, corporate espionage has dramatically increased. Organizations and businesses need to be connected to the Internet to remain competitive or to access the wealth of information available. Credit card fraud is still a concern primarily because of hackers breaking into commercial sites and stealing credit card numbers that can then be sold or held for ransom. Electronic theft (e-theft) involves hackers using Internet banking connections to hack into banks for the purpose of making money transfers from legitimate accounts to bogus accounts, which could then be accessed by the perpetrators (Philippsohn, 2001b). Philippsohn (2001b) also reported that the number one Internet fraud within the United States is that of online auction fraud. Other thefts involve using the Internet in advance fee frauds, bogus investments, pyramid frauds, bogus Internet banking Web sites, cyberterrorism, virus propagation, and cyberpiracy (Philippsohn, 2001b). Money laundering over the Internet has increased dramatically as a result of the Internet's characteristics of ease of access, anonymity, transaction speed, global reach, and nonexistent jurisdictional boundaries (Philippsohn, 2001a).

INTERNET LEGAL ISSUES

The primary legal issues related to the Internet are the location or venue where Internet crimes occur and the jurisdictions where they are to be prosecuted. Jurisdictional issues have yet to be resolved and are often further complicated by the difficulty in establishing the venue of some Internet crimes. Because of this, legal opinions are widespread, with some proposing that international laws are needed while others propose using existing laws already in place through an extension of their jurisdictional power and reach.

Geist (2001) stated that local law compliance is rarely enough to ensure that a business on the Internet is limited in exposure to legal risk because of the uniqueness of the Internet. Geist was referring to legal companies doing business over the Internet, and courts worldwide are having difficulties deciding jurisdictional issues involving legal entities. The jurisdictional issues related to criminal activities over the Internet are compounded in deciding what country should criminally try a case where the offender is on one side of the world and the victim is on the other side. In the Nigerian fraud scheme, some victims were actually lured to Nigeria, which puts them at risk because they have been involved in an attempt to defraud the Nigerian government (Internet ScamBusters, 1996, (http://www.scambusters.org/NigerianFee.html; U.S. Postal Inspection Services, n.d., http://www.usps.com/websites/depart/inspect/pressrel.htm). In 1994 the Department of Justice announced that a Federal Grand Jury indicted three Nigerian citizens on nine counts for wire fraud and that extradition proceedings would begin (U. S. Department of Justice, n.d., http://www.usdoj.gov/opa/pr/Pre_96/September94/519.txt.html).

Security and privacy of information and data over the Internet are other critical concerns of Internet users, not only from the standpoint of being exploited by those with criminal intent, but also from the standpoint of governmental intrusions into private areas protected by law. The Department of Justice Computer Crime and Intellectual Property Section Web site has posted legal issues as related to electronic commerce (http://www.cybercrime.gov/ecommerce.html). Another Department of Justice Computer Crime and Intellectual Property Section Web site iterates the new authorities that relate to computer crime and electronic evidence enacted in the USA Patriot Act of 2001 (http://www.usdoj.gov/criminal/cybercrime/PatriotAct.htm).

The legal issues of encryption and the right to privacy pose a unique dilemma that has placed governments in the precarious position of having to balance the rights of individuals to privacy with the right and obligation of governments to protect its citizenry and industries. One proposed solution to the security and privacy issues on the Internet is the FBI Carnivore program, a packet-filtering program (FBI, n.d., http://www.fbi.gov/hq/lab/carnivore/carnivore.htm) designed to examine packets of information on the Internet. This information is broken into packets, and then the individual packets are transmitted. With proper court authorization, the FBI can use Carnivore to view suspicious message transmissions over the Internet. The problem encountered with Carnivore is that the information packets composing one message are not transmitted together in sequence. When transmitted, they become intermingled with unrelated information packets not authorized for viewing by Carnivore. This intermingling of unrelated packets with the suspect packets makes it impossible to view only the authorized packets of information. The resulting issue becomes one of government trust and trust of law enforcement to only view, save, or in any way make retrievable only the packets of information authorized to be viewed. On July 24, 2000, Donald M. Kerr, Assistant Director, Laboratory Division, FBI, presented a Congressional Statement for the Record before the U.S. House of Representatives, the Committee on the Judiciary, Subcommittee on the Constitution, the Internet and Data Interception Capabilities Developed by the FBI. Donald M. Kerr in his statement discussed issues related to the control and use of Carnivore (Kerr, 2000, http://www.fbi.gov/congress/congress00/kerr072400.htm).

CONCLUSION

The ramifications of the Internet upon law enforcement, locally and globally, are widespread, numerous, and even controversial. Organized crime has at its disposal monetary support received from illegal transactions, while government must rely upon monies received from its citizenry.

The greatest challenge faced by law enforcement in this technological age of the Internet is that of becoming technologically advanced beyond the level of criminal expertise and knowledge.

GLOSSARY

Authentication The use of a password, encryption, or a combination of both, to confirm the identity and right of an individual to access protected networks or Web sites.

Community policing A philosophy based upon the recognition that dedicated people working together through community partnerships can make their communities safer and better and as related to law enforcement, accomplished by putting more police officers into communities to address not only crimes but also quality of life issues.

Cyberstalking The use of e-mail, instant messages, bulletin boards, chat rooms, or Web-based forums via the Internet by an individual to pursue and repeatedly attempt to contact another individual, the victim, which results in the victim feeling threatened and fearful of the contact.

Digital signature A protocol that has the same effect as that of a real signature in that it is a mark that only the sender can make, it is readily identified by others as belonging to the sender, it confirms agreement, and it possesses the additional properties of being unforgeable, authentic, unalterable, and nonreusable.

Encryption The use of coding to transform data or information into an unintelligible form that is confidential, unmodifiable, and not reproducible by unauthorized persons.

Firewall A computer or computer software that prevents unauthorized access to private data or information located on a organization's internal network by individuals located externally on the Internet.

Jurisdiction The right, power, or authority to interpret and apply law within the limits or region where the authority may be lawfully exercised.

Modus operandi As related to crime it is a distinct method used or operating pattern that indicates the work of a single criminal or a criminal group in two or more crimes.

NIBRS The FBI's National Incident-Based Reporting System using new guidelines beyond those of the UCR to better meet the law enforcement needs in the 21st century by reporting arrests and offenses using an incident-based system; collecting data at two levels, one of which would include more detail, and using a quality assurance program to improve the quality of the collected data.

Packet-filtering program A computer program that examines sender, receiver, and sequencing information contained in the header of a packet.

Protocol A set of rules that manage how software and hardware components communicate with one another or cross interfaces.

Spamming Sending unsolicited e-mail to a large number of addresses.

Stalking The repeated and unwanted behaviors of an individual attempting to contact another individual, the victim, which results in the victim feeling threatened and fearful of the contact.

UCR FBI Uniform Crime Report instituted in the 1930s whose purpose is to generate a reliable set of criminal statistics that can be used in the administration, operation, and management of law enforcement.

Urban legend An often lurid anecdote or tale based on hearsay and circulated as a true and factual story.

Venue The country or place where alleged events take place that results in legal actions.

Virtual private network An encrypted Internet or network "tunnel" to ensure point-to-point communication of messages and protects against interception by unauthorized persons through the use of a combination of firewalls, encryption, and authentication.

Virus A computer program typically secluded within an innocuous program that produces copies of itself and then inserts the copies into other programs usually for the purpose of performing a malicious action such as denial of services or destruction of data.

Voice over IP The transmission of voice communications using the Internet protocol for data transmission where the voice signal is converted into digital format, divided into packets, and transmitted over the Internet.

Weed and seed Law enforcement grant-supported efforts to prevent, control, and reduce violent crimes, drug trafficking, drug abuse, and gang-related activities in specific targeted high-crime neighborhoods ranging in size from several city blocks to as large as 15 square miles throughout the United States.

CROSS REFERENCES

See *Authentication; Cybercrime and Cyberfraud; Cyberstalking; Cyberterrorism; Digital Signatures and Electronic Signatures; Encryption; Firewalls.*

REFERENCES

Baker, D. (1999). When cyber stalkers walk. *ABA Journal, 85,* 50–54.

Bequai, A. (2000). America's Internet commerce and the threat of fraud. *Computers & Security, 19,* 688–691.

Computer Crime and Intellectual Property Section of the Criminal Division of the U.S. Department of Justice (n.d.). *Electronic commerce: Legal issues.* Retrieved January 1, 2003, from http://www.cybercrime.gov/ecommerce.html

Computer Crime and Intellectual Property Section of the Criminal Division of the U.S. Department of Justice (n.d.). *Field guidance on new authorities that relate to computer crime and electronic evidence enacted in the USA Patriot Act of 2001.* Retrieved January 1, 2003, from http://www.usdoj.gov/criminal/cybercrime/PatriotAct.htm

Dube, N. A. (1999). Globalsearch: The international search for missing and exploited children. *International Review of Law, Computers and Technology, 13,* 69–74.

Duffield, C. (2000). Turning the technology tables of serious crime. *Document World, 5,* 40–42.

Federal Bureau Investigations (FBI) (n.d.). *Carnivore.* Retrieved January 1, 2003 from http://www.fbi.gov/hq/lab/carnivore/carnivore.htm

Federal Bureau Investigations (FBI) (n.d.). *Common fraud scams.* Retrieved January 1, 2003, from http://www.fbi.gov/majcases/fraud/fraudschemes.htm

Gallagher, P. (1999). E-commerce trends. *International Trade Forum, 2,* 16–18.

Geist, M. A. (2001). Is there a there there? Toward a greater certainty for Internet jurisdiction. *Berkeley Technology Law Journal, 16,* 1345–1406.

Hitchcock, J. A. (2000). Cyberstalking. *Link-up, 17,* 22–23.

InterGov International (2002). *Latest Web statistics.* Retrieved January 1, 2003, from http://www.intergov.org/public_information/general_information/latest_web_stats.html

International Center for Missing and Exploited Children (ICMEC) (n.d.). Retrieved January 1, 2003, from http://icmec.missingkids.com/

Internet Fraud Complaint Center (IFCC) (2001). *IFCC 2001 Internet fraud report*. Retrieved January 1, 2003, from http://www1.ifccfbi.gov/strategy/IFCC_2001Annual_Report.pdf

Internet ScamBusters (1996). SCAM: The Nigerian advance fee scheme. Reprinted from *Internet ScamBusters, 11*. Retrieved January 1, 2003 from http://www.scambusters.org/NigerianFee.html.

Kalakota, R., & Whinston, A. B. (1996). *Frontiers of electronic commerce*. New York: Addison-Wesley.

Kerr, D. M. (2000). Internet and data interception capabilities developed by the FBI. Congressional Statement Federal Bureau of Investigation. Retrieved January 1, 2003 from http://www.fbi.gov/congress/congress00/kerr072400.htm

Macodrum, D., Hasheri, H., & O'Hearn, T. J. (2001). Spies in suits: New crimes of the information age from the United States and Canadian perspectives. *Information and Communications Technology Law, 10*, 139–166.

National Center for Injury Prevention and Control (NCIPC) (n.d.). *Leading causes of death reports*. Retrieved January 1, 2003, from http://webapp.cdc.gov/sasweb/ncipc/leadcaus.html

National Criminal Intelligence Service (NCIS) (n.d.). *West African organized crime section*. Retrieved January 1, 2003, from http://www.ncis.gov.uk/waocu.asp

National Law Enforcement Telecommunications System (NLETS) (n.d.). Retrieved January 1, 2003, from http://www.nlets.org/index.htm

National Security Council (NSC) (1997). *International crime control strategy*. Retrieved January 1, 2003, from http://clinton4.nara.gov/WH/EOP/NSC/html/documents/iccs-frm.html

NUA, Scope Communications Group (2002). *How many online*. Retrieved January 1, 2003, from http://www.nua.ie/surveys/how_many_online/world.html

Philippsohn, S. (2001a). Money laundering on the Internet. *Computers and Security, 20*, 485–490.

Philippsohn, S. (2001b). Trends in cybercrime—An overview of current financial crimes on the Internet. *Computers and Security, 20*, 53–69.

Ruthfield, S. (1995). The Internet's history and development: From wartime tool to the fish-cam. *ACM Crossroads, 2*(1). Retrieved January 1, 2003, from http://www.acm.org/crossroads/xrds2-1/inet-history.html

University of Alaska at Anchorage Justice Center (2000). *Focus on the death penalty: International context*. Retrieved January 1, 2003, from http://www.uaa.alaska.edu/just/death/intl.html

U.S. Department of Justice (n.d.). *Grand jury indicts Nigerians in wire fraud scheme*. Retrieved January 1, 2003 from http://www.usdoj.gov/opa/pr/Pre_96/September94/519.txt.html

U.S. Postal Inspection Services (n.d.). *Postal inspectors crackdown on Nigerian scam*. Retrieved January 1, 2003 from http://www.usps.com/websites/depart/inspect/pressrel.htm

FURTHER READING

All sites accessed and available as of January 2003.

Introduction

NUA (Scope Communications Group) Internet Surveys: http://www.nua.ie/surveys/

Law Enforcement Agencies' Web Sites

Bureau of Alcohol, Tobacco, Firearms, and Explosives: http://www.atf.treas.gov/

Bureau of Justice Statistics law enforcement information: http://www.ojp.usdoj.gov/bjs/lawenf.htm

Canine Police: http://www.policek9.com

Crime Stoppers: http://www.c-s-i.org/

Domestic violence: http://info-s.com/domestic.html

FirstGov, a locator for federal and state government law enforcement sites: http://www.firstgov.gov/

Hong Kong law enforcement: http://www.info.gov.hk/police/index.htm

Icelandic Police: http://www.logreglan.is/displayer.asp?catid =217

InterGov International: http://www.intergov.org/

International Association of Law Enforcement Firearms Instructors: http://www.ialefi.com/

International police agencies: http://www.policeinternational.com/, http://www.interpol.int/)

International Web Police: http://www.web-police.org/

Law enforcement agencies and organizations: http://search.officer.com/agencysearch/, http://info-s.com/police-d.html, http://info-s.com/laworg.html

Official Directory of State Patrol and State Police: http://www.statetroopersdirectory.com/

Pakistan law enforcement: http://www.sindhpolice.gov.pk/

Police Federation of England and Wales: http://www.polfed.org/main_frame.htm

Russia law enforcement: http://www.mvdinform.ru/index.php

UK Police Services: http://www.police.uk/

Universal Police Link site: http://www.spinet.gov.sg/pollink/index.html

U.S. and International Government, Military, and Intelligence Agencies: http://www.sagal.com/ajax/

U.S. Department of Justice: http://www.usdoj.gov/index.html

World Factbook of Criminal Justice Systems: http://www.usdoj.gov/bjs/abstract/wfcj.htm

Worldwide Governments on the WWW: http://www.gksoft.com/govt/en/world.html

Web Sites for Specific Law Enforcement Interests and Concerns

Adoption records and searches: http://www.adopting.org/adoptees.html, http://www.birthquest.org

Amnesty International—The death penalty: http://web.amnesty.org/pages/deathpenalty_index_eng?OpenDocument

American Academy of Forensic Sciences: http://www.aafs.org

Background checks: http://info-s.com/background.html

California Department of Justice, Stop Drugs: http://www.stopdrugs.org/identification.html

Center for Civilian Biodefense Strategies: http://www.hopkins-biodefense.org

Center for Disease Control and Prevention (CDC): http://www.cdc.gov/

Center for the Study of Terrorism and Political Violence: http://www.st-and.ac.uk/academic/intrel/research/cstpv/

Central Intelligence Agency—The war on terrorism: http://www.cia.gov/terrorism

CIA Factbook: http://www.odci.gov/cia/publications/factbook/index.html

College and university phonebooks: http://www.uiuc.edu/cgi-bin/ph/lookup?Query=

Community Policing Consortium: http://www.communitypolicing.org/

Corrections resources: http://www.officer.com/correct.htm

Death Penalty: http://info-s.com/deathpenalty.html

DOE Network: http://www.doenetwork.org

Domestic Security in Florida: http://www.myflorida.com/myflorida/domestic_security/index.html

Drug Enforcement Agency: http://www.usdoj.gov/dea/

Executive Office for Weed and Seed: http://www.ojp.usdoj.gov/eows/welcome.html

FBI Crime Scene Safety Handbook: http://www.fbi.gov/programs/lab/handbook/safety.htm

Federal Bureau of Investigations—War on terrorism: http://www.fbi.gov/terrorinfo/terrorism.htm

Federal Emergency Management: http://www.fema.gov

Federation of American Scientists: http://www.fas.org/index.html

Firearms and firearms identification: http://www.firearmsid.com/, http://info-s.com/firearm.html

Florida State University School of Criminology & Criminal Justice: (http://www.criminology.fsu.edu/

Fraternal Order of Police: http://www.grandlodgefop.org/

Fugitives: http://www.fugitive.com/, http://www2.amw.com/amw.html, http://www.fugitivehunter.org/usmostwanted.html, http://www.mostwanted.org/

Fulton County Medical Examiner's Center Atlanta, Georgia—Unidentified deceased individuals: http://www.fcmeo.org/UIDPage.htm

Genealogy: http://www.nara.gov/genealogy/

Henry L. Stimson Center for International Security and Peace—Frequently asked questions: Likelihood of terrorists acquiring and using chemical or biological weapons: http://www.stimson.org/cbw/?SN=CB2001121259

IAGIM—Drug Development Association: http://www.iagim.org/index.htm

International Association of Chiefs of Police: http://www.theiacp.org/

International Association of Women Police: http://www.iawp.org/

International Policy Institute for Counter-Terrorism: http://www.ict.org.il/

JURIST—Terrorism laws and policies: http://jurist.law.pitt.edu/terrorism.htm

Kentucky State Medical Examiners Office: http://www.unidentifiedremains.net/

Locum International Group—U.S. drug identification list: http://www.locum.co.il/druglist/index.php

MEDLINEplus: http://www.nlm.nih.gov/medlineplus/

Michigan State Police: http://www.msp.state.mi.us/persons/remains.htm

Military personnel: http://www.nara.gov/regional/mpr.html, http://www.globemaster.de/faq/locator.html

Missing Persons Cold Case Network: http://www.angelfire.com/mi3/mpccn/index.html

National Association of Chiefs of Police: http://aphf.org/nacop.html

National Center for Women and Policing: http://www.womenandpolicing.org

National Criminal Justice Reference: http://www.ncjrs.org/lewww.html

National Criminal Justice Reference—International subcategories: http:virlib.ncjrs.org

National Highway Traffic Safety Administration: http://www.nhtsa.dot.gov

National Institute on Drug Abuse: http://165.112.78.61/DrugAbuse.html

National Law Enforcement Officers Memorial Fund: http://www.nleomf.com

National Library of Medicine: http://www.nlm.nih.gov/

National Organization of Black Law Enforcement Executives: http://www.noblenatl.org/

New Jersey Domestic Security Preparedness Task Force: http://www.state.nj.us/lps/dsptf/dsptfhome.html

Northern Michigan University Department of Criminal Justice: http://www.nmu.edu/cj/default.htm

Office of Juvenile Justice and Delinquency Prevention: http://ojjdp.ncjrs.org/grants/grants.html

Orange County, CA, Sheriff's Office—Street drug identification: http://www.ocso.com/sid/narcotics/identification/drugs.html

Organized crime and corruption: http://yorku.ca/nathanson/default.htm

Police Executive Research Forum: http://www.policeforum.org/

Police Training Institute at the University of Illinois: http://www.pti.uiuc.edu/

Political: http://www.tray.com/fecinfo/

Prison information and papers: http://www.soci.niu.edu/~critcrim/prisons/prisons.html

Public records: http://www.searchsystems.net/, http://www.publicrecordfinder.com/

Regional Information Sharing Systems (RISS) Program: http://www.iir.com/riss/

Sentencing: http://www.ojp.usdoj.gov/bjs/sent.htm

Serial killers: http://www.crimelibrary.com/serialkillers.htm

Telephone and e-mail information: http://www.anywho.com/

Telephone reverse lookup: http://www.officer.com/research.htm

Training: http://www.fbi.gov/hq/td/academy/academy.htm

Unclaimed property: http://www.unclaimed.org

United Nations: http://www.un.org

University of Albany Hindelang Criminal Justice Research Center: http://www.albany.edu/hindelang/

University of Memphis: http://www.memphis.edu/

Urban Legends: http://www.scambusters.org/legends.html

U.S. Department of Justice Bureau of Justice Statistics: http://www.usdoj.gov/bjs/

U.S. Department of Justice Office of Community Policing Services: http://www.cops.usdoj.gov/

U.S. Department of Justice Office of Justice Grants: http://www.ojp.usdoj.gov/fundopps

U.S Department of State: http://www.state.gov/

U.S Department of State, Office of International Information Programs—Response to terrorism: http://usinfo.state.gov/topical/pol/terror/

Vital records: http://vitalrec.com/index.html

Communication Activities

Copnet—Public services: http://www.copnet.org/pubserv.html

CopsOnLine: http://www.copsonline.com

Police World: http://www.policeworld.net

Police-L: http://www.police-l.org/about.html

Law Enforcement Use of E-commerce

Clerk, Cobb County State Court, Marietta, Cobb County, Georgia: http://www.cobbstatecourtclerk.com/

Collecting traffic fines and fees: http://www.ezgov.com/newsfullpage.jsp?ArticleID = 3684&category = pressreleases

Courts of New Zealand—Fines On-Line: http://www.fines.govt.nz/

Crime and the Internet

FedStats—Statistical agencies: http://www.fedstats.gov/agencies/

FBI Online Child Pornography Innocent Images National Initiative: http://www.fbi.gov/hq/cid/cac/innocent.htm

FBI Uniform Crime Reports & NIBRS Federal Bureau of Investigations: http://www.fbi.gov/ucr/ucr.htm

Internet Fraud Complaint Center: http://www1.ifccfbi.gov/strategy/statistics.asp

Uniform Crime Reports: http://www.fbi.gov/ucr/ucr.htm

United Nations Office on Drugs and Crime: http://www.unodc.org/unodc/crime_cicp_sitemap.html

U.S. Customs CyberSmuggling Center's Child Exploitation Unit: http://www.cbp.gov/xp/cgov/enforcement/investigative_priorities/c3/child_exploitation.xml

University of Maryland Libraries—Sources of international statistics on the Internet: http://www.lib.umd.edu/MCK/STATS.html

Internet Legal Issues

Cybercrime: http://www.cybercrime.gov

Department of Justice Computer Crime and Intellectual Property Section—Electronic commerce: Legal issue: http://www.cybercrime.gov/ecommerce.html

Law Firms

Victoria S. Dennis, *Minnesota State Bar Association*
Judith C. Simon, *The University of Memphis*

INTRODUCTION

Law firms were not one of the first types of organizations to use the Internet extensively, but they have made significant gains in Internet presence in recent years. Researching legal issues was one of the first uses law firms made of the Internet, and the types of resources available for such research have continued to expand. The topics available online are much more extensive than just a few years ago. Eventually, law firms began to have Web sites online to provide contact information to their clients. Now those Web sites have expanded to provide additional services to clients. Law firms have also adopted Internet technology for use in providing internal Web capabilities within their organizations (intranets) and sometimes for links to specific outside organizations (extranets).

Some ways in which the Internet is widely used by law firms are described in the sections below. Although a prominent use of the Internet by law firms is e-mail, this topic is dealt with in detail in other chapters. The first and most extensive section involves ways in which the Internet is used for legal research, which continues to be the capability of the Internet most used by law firms. The second major section describes client services available for lawyers using the Internet. The third section discusses concerns of in-house counsel, as well as other internal concerns related to the Internet, primarily the potential use of intranets and extranets. The final section provides a listing of some major online resources for law firms, with brief identifications of the types of online materials that are available at those sites.

The discussions below attempt to provide an overview of ways that law firms are making applications of Internet technology, with enough details to indicate the range and quantity of resources available.

LEGAL RESEARCH USING THE INTERNET

Use of the Internet to conduct legal research has become much more extensive in recent years as the number of law-related sites has expanded. Professional organizations, government entities, law firms and other legal professions, and educational institutions are examples of suppliers of Web site content.

The ABA's Legal Technology Resource Center conducted a survey of trends in legal publishing, which was reported on its Online Research site (http://www.abanet.org/tech/ltrc/oresearch.html). The findings indicated that progress in the movement by legal publishers from print formats to electronic formats (CD and Internet) had been slow but steady. About 30% of the media production by legal publishers was for Internet products. Prices have increased for products in print media format, which will likely cause a greater movement to Internet media. An initial concern related to the Internet involved reliability of information, but the expectation now is that the quality of Internet products will continue to improve.

A good "first step" when performing legal research on the Internet is the Web site of the user's state bar association. Most states' bar associations have very comprehensive Web sites, with links to sites on specific topics as well as state statutes and rules. There are generally also lists of in-state contact individuals.

For much online research, a nonlegal search engine such as Yahoo! or google.com may suffice to locate the information that is needed. Before information that is garnered from a Web site not sponsored by a reliable source is used, however, the validity of the information must be cross-checked with a reliable source.

In locating a specific attorney or law firm, http://www.martindale.com, the Web site for Martindale–Hubbell, is frequently updated and is very useful. This Web site also contains links to other generalized legal Web sites.

The two Web sites that most recent law school graduates become familiar with and continue to rely upon during practice are those operated by Westlaw and LexisNexis. Westlaw is the larger service, and the more user-friendly of the two, according to the American Bar Association's 2001 Legal Technology Resource Center Survey. LexisNexis directs its services to a broader area than just the legal profession, facilitating not only traditional legal research but also research by accountants, academicians, corporate officers, journalists, librarians, and other professionals. The advent of these two services revolutionized legal research. Attorneys are no longer bound to law libraries; equally important, however, is the fact that information gathered through these online services can be relied upon to be extremely current. Both of these services are available through subscriptions.

There are, additionally, a number of excellent free research Web sites. Some contain general knowledge, such as http://www.nolo.com, which contains a broad spectrum of information in a user-friendly format. This site, although directed somewhat to nonlawyers, has a wealth of knowledge and links. LexisNexis operates a free site, http://www.lexisone.com, that links to much of the basic information, such as cases and forms, that is offered through subscription services. Additionally, this site offers "pay as you go" options, such as a per-case charge for Shepardizing documents (determining the current status of case law or legal authority, finding the most recent decisions, and determining if other cases have dealt with the same factual issues). These types of pay-per-use options have made online research more accessible to sole practitioners and attorneys practicing in small firms, who may not be able to afford subscriptions to a full-service option. Other free sites offering good legal research options include http://www.lawyers.com, http://www.findlaw.com, and http://www.law.com.

Additionally, there are a number of Web sites specific to areas of practice that are a good "first stop" before a broader search-engine Web search. In many cases, these sites can act as a practitioner's "frequently asked questions" page, and consulting them first may make broader searches unnecessary. Several areas of law have a number of excellent sites with information specific to those practices. Some of those topics are described in this section, including

Patent law
Income tax law
Real property law
Family law
Litigation
Alternative dispute resolution
Electronic commerce
Legal ethics

Patent Law

A patent is obtained by an inventor to keep others from producing or using that person's invention for some length of time. To qualify, the invention has to be unique, have a use, and not otherwise be obvious in nature.

Patent law is arguably the preeminent area of law accessible via the Internet. The nature of patents themselves and of those attorneys who practice patent law are a natural fit for creating and maintaining excellent Web sites. Information as basic to patent practice as Article I, Section 8 of the U.S. Constitution and Title 35 of the United States Code are readily available online. The U.S. Constitution states that the U.S. Congress has the power ". . . to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries."

Title 35 of the United States Code contains the main body of law related to patents. For example, the time frame for patents is covered. Until a recent amendment, patents were issued for one nonrenewable time period of 17 years from the date the patent was issued. In June 1995, the term limit became 20 years from the date of application for the patent.

The Patent and Trademark Office administers patent laws in the United States. An examiner in that office reviews each application to determine its qualifications for a patent. The regulations used by the Patent and Trademark Office are available from online sources.

For more information than is available from the U.S. Code, it is necessary to tailor the search very specifically because the volume of information and the level of detail available in this area are enormous. A very good general patent Web site is located at Cornell's Legal Information Institute (http://www.law.cornell.edu/topics/patent.html). Here are some useful links provided on this site's home page (Legal Information Institute, 2002b):

Article I, Section 8 of the U.S. Constitution
Title 35 of the United States Code
Decisions of the Supreme Court and U.S. Circuit Court of Appeals related to patents
Patent database
International Patent News Service (free weekly e-mail notice of patents issued)
World Intellectual Property Organization, including a database of intellectual property laws

For a source of more specific patent information, a good recommendation is PatentLawLinks.com (http://www.patentlawlinks.com/). This site (PatentLawLinks.com, 2002) provides links to access such items as

Case law, including Supreme Court decisions and cases in the Court of Appeals and U.S. District Courts
Patent search engines
Worldwide patent office sites
Patent law firms in the U.S. and in other countries
Intellectual property law
Statutes and regulations
Academic and professional journals
Intellectual property news and mailing lists
Legal search engines, such as LawCrawler
Patent-related forms

Another useful Web site, the U.S. Patent and Trademark Office (http://patents.uspto.gov/), has a substantial index that allows searches for patents (United States Patent and Trademark Office, 2002). Users may also access information on the patenting process and apply online for patents.

Income Tax Law

Income tax law is one of the more form-driven areas of practice, therefore lending itself well to online research. Whether an attorney needs to locate U.S. Code, court decisions, or the most current legislation or to download an arcane form, there are Web sites available to provide the needed information. An excellent Web site providing hundreds of well-described links is at taxsites.com (http://www.taxsites.com/federal.html). This website (Taxsites.com, 2002) has links to such materials as

Tax legislation updates
Summaries of recent tax acts
Tax code and tax regulations
IRS Web site
Related court decisions

These sections include subdivisions for free services and Web-based subscription services including Lexis-Nexis and Westlaw. Another excellent Web site that provides a vast storehouse of information as well as tax forms that can be downloaded, is http://www.irs.gov.

Real Property Law

Real property (more commonly referred to as real estate) law is an area of practice that is more slowly making information available on the Internet. For a variety of reasons, primarily the sheer volume of documents involved in each real property transaction, it is cumbersome to have a great deal of the information available online. Title searches, property tax information, and contractors' liens are areas that may be researched online in some jurisdictions.

The American Bar Association (ABA) site (http://www.abanet.org) has a section on Real Property, Probate, and Trust. It contains links to useful sources such as the *Real Property, Probate and Trust Journal* and *Probate and Property Magazine*. It also has a public information section, with links to a wide range of information on topics in which legal agreements will no doubt occur, such as

Buying or selling a house (including mortgage borrowing and a lengthy family legal guide)
Remodeling a house
Renting a house
Liability issues and insurance

Family Law

Family law became an area of distinct legal specialization around 1960 and has become quite extensive in scope. According to the American Bar Association Section of Family Law Web site (http://abanet.org/family/), areas of interest include

Divorce
Custody
Adoption
Alimony
Child abduction
Federal and interstate legislation
Mediation
Paternity
Genetic engineering

The major areas listed above, as well as the general area of family law, can be found as links on numerous Web sites mentioned in this chapter that support legal research on the Internet.

Litigation

Litigation generally refers to the handling of lawsuits and involves an enormously wide-ranging group of topics. Numerous Web sites on litigation are available. For example, the FindLaw Web site (http://findlaw.com/01topics/29litigation/) has links to litigation attorneys and litigation law firms, litigation practice support and consultants, and litigation expert witnesses (FindLaw for Legal Professionals, 2002). Some of the practice support and consultants links are listed below to provide a general idea of the types of available resources:

Business consultants
Courier/messenger services
Document preparation
Jury consultants
Paralegals
Technology consultants
Trial presentation

Large consulting firms often have specialty legal practices, with information provided at their Web sites. For example, Ernst & Young's Web site (http://ey.com/) includes a link to their Litigation Advisory Services page (Ernst & Young, 2002). This page lists their main services, with explanations and instructions for obtaining additional information. Some of their services include

Fact finding and discovery
Analysis and quantification of damages
Fraud and forensic investigation
Expert witness testimony

A special subtopic of litigation involves Internet or e-commerce issues. In the case of contracts, some Internet contracts are no different from traditional contracts. However, others are unique to the Internet. For example, several cases involving the terms of service have involved "clickwrap" agreements, which are similar to non-Internet "shrinkwrap" agreements and often involve the purchase of software online. Another Internet-related litigation topic has to do with consumer fraud, which is a standard litigation topic but with different circumstances because of the use of the Internet. Many Internet-related cases have involved false advertising and deceptive practices.

Alternative Dispute Resolution

Alternative dispute (ADR) involves ways of settling disputes without going to court. This option is increasingly given consideration, for reasons including the time delays in getting a case to court, as well as the rising costs of litigation. According to the Legal Information Institute Web site on ADR (http://law.cornell.edu/topics/adr.html), the two most common forms of ADR are arbitration and mediation (Legal Information Institute, 2002a). Arbitration is considered a simplified form of a trial, with arbitration hearings lasting only for a few hours. Mediation is a less formal method than arbitration, in which mediators trained in negotiations bring the opposing parties to a meeting to try to work out an agreement that both can accept.

The Legal Information Institute Web site (Legal Information Institute, 2002a) contains links to related information, including

Title 9 of the U.S. Code regarding Federal law supporting arbitration

UN Convention on the Recognition and Enforcement of Foreign Arbitral Awards

Recent decisions on arbitration by the U.S. Supreme Court and the Circuit Courts of Appeals

State statutes dealing with ADR

U.S. Department of Justice, Office of Dispute Resolution

Technical arbitration and resolution

International Chamber of Commerce Index of International Arbitration and Alternative Dispute Resolution Services

Mediation Information and Resource Center

ADR may now be conducted online, at such sites as http://www.mediate.com, http://www.arb-forum.com and http://www.squaretrade.com. As with other topics mentioned in this chapter, additional sites are available related to ADR, but this section provides a sampling of the resources that are available.

Electronic Commerce (E-commerce)

The ABA Legal Technology Resource Center Web site (http://abanet.org/tech/ltrc/) has a link to "Web Development & E-Commerce," with suggestions and resources for law firms to use to develop their own Web sites (ABA Legal Technology Research, 2002). The ABA site has links to topics such as

How to plan, design, and build a Web site

Tips for getting a law firm onto the Web, such as obtaining a domain name and a Web host

Resources, such as product sites

An article on "Web Worries of Dot-com Lawyers" (Beckman and Hirsh, 2000) discussed several concerns of "e-lawyers" related to providing legal services on the Internet. One service that was described involved providing documents on Web sites for a fee and using credit cards for the collection method. Concerns included checking for conflicts before obtaining any confidential information that might be submitted on the form and determining how to keep confidential information secure. The general conclusion was that e-lawyering is expected to continue, and many law firms will need to provide the service to remain competitive.

In addition to considering the possibility of conducting business themselves over the Internet, lawyers are now involved in other legal issues regarding e-commerce. Numerous articles and other resources are available online regarding this subject, which is likely to continue to expand. One concern is related to Internet agreements, such as those that involve approval to have links to another organization's Web site or use its content, logo, etc. Some of the contracts involving the Internet are similar to existing non-Internet practices, but the use of specific Internet-related contracts removes some possible uncertainty, because the Internet has caused new concerns related to topics such as intellectual property and ownership. Sources are available for Internet-specific contracts, as well as other related information of value.

Legal Ethics

The ABA has a Center for Professional Responsibility with an Ethics Department (ABA Center for Professional Responsibility, 2002). According to the Web site (http://www.abanet.org/cpr/ethics.html), which provides links to extensive information on this Center and related issues, the Center studies, develops, and implements model legal and judicial ethics standards. The Ethics Department reports on new developments and drafts potential revisions to the model ethics standards. This department also operates an extensive research service called ETHICSearch, which analyzes ethical dilemmas and helps users in determining appropriate standards and materials to resolve the dilemmas. Additional information on this service is available at this same Web site.

The ABA Center for Professional Responsibility Web site also provides "Materials for Research in Legal Ethics," with links to additional important information, such as

Model Rules for Professional Conduct

Model Code of Professional Responsibility

Model Code of Judicial Conduct

CLIENT SERVICES FOR LAWYERS USING THE INTERNET

Attorneys have been slower than many other businesspeople in developing ways to utilize the Internet to communicate with clients. Because of concerns about client confidentiality, the ethics of "advertising" by posting a Web site, and the concern about a Web site lacking of the proper decorum desired by the profession, as well as justifiable concerns regarding accusations of offering legal advice online, Web sites maintained by law firms have primarily listed the firm s' addresses and telephone numbers, along with the attorneys' names, but have contained very little other information. This is changing, as perceptions about the propriety of doing business on the Internet change.

Although initially law firms developed Web sites to attract new clients, more legal Web sites are now being developed for other purposes, such as attracting recruits from law schools or serving as portals with information on specific areas of law. A significant change has also occurred in the number of legal sites that act as extranets for existing clients, to allow secure document exchanges.

According to findlaw.com, a law firm in Cordova, Tennessee is revolutionizing the way it interacts with clients using the Internet (Bodine, 2000). The law firm makes its Web content accessible for download into a Palm Pilot, and the attorneys utilize the free calendaring service offered at http://calendar.yahoo.com/. Only the person to whom the calendar belongs can view the details of the appointments (a client viewing his or her attorney's calendar would only see that the attorney was busy during a certain block of time). Clients can then see when the attorney is available prior to calling to make an appointment. Because there is sometimes a perception by clients that attorneys are inaccessible, this level of openness in communicating makes the attorney–client relationship more of a partnership and makes the client feel comfortable when hiring an attorney that he or she will be able to contact the attorney when necessary.

THE INTERNET FOR IN-HOUSE COUNSEL AND OTHER INTERNAL ACTIVITIES

In-house counsel have a single-minded responsibility, unique among attorneys, to a single corporate client. In the context of the Internet, this responsibility may be divided into three categories:

Legal issues raised by the company's current use of the Internet

Research into how the company's Internet use can and should develop, and the legal issues raised by this future Internet use

How the practice of law will develop and change with the use of the Internet

The first two categories present similar sets of issues: copyright protection and ensuring that the content of the Web site is accurate and is not biased regarding race, gender, sexual orientation, or religion, or otherwise offensive or unlawful. The company will likely want to have a statement regarding the content of any Web sites to which it links, limiting the company's liability to that content over which it has direct control. If the Web site includes a service for which customers need to register, the Web site should have a license agreement which the customer must agree to by clicking on an "I Agree" button before proceeding further. This is an example of a "clickwrap" agreement. These agreements have been in place for a number of years, and their validity has been upheld in most jurisdictions.

If the company is actually doing business via the Web site, it is also wise to have a "choice of law" statement posted on the Web site. Because Web sites can be viewed worldwide, a potential lawsuit could arise from anywhere in the world. This clause limits the litigation of any issues that arise between the company and the users from use of the Web site to a certain jurisdiction.

Additionally, in-house counsel will want to develop a technology use policy for the employees of the company, as well as any contractors working onsite. Critical to such a policy is a clear statement that the use of the company's resources is for business purposes only, and the employee should have no expectation of privacy. Only with such a policy in place can a company then safely, without fear of an invasion of privacy lawsuit from the employee, ensure that employees are not improperly or unlawfully using the company's resources to access or post content on the Internet.

Another type of decision related to Internet-related activities involves the use of "intranets" and "extranets." Intranets, as the "intra-" prefix suggests, are used for internal Web-based activities in an organization. Accessing needed documents and data using Web technology can be relatively easy, but can also provide some security related to access. An individual group within a firm may have links to documents that they need within the group but may not have access to pages that belong to activities of other groups within the firm. Additionally, a firm's internal library can be made available and regularly updated, along with links to external Internet data. When the firm has a need for secure transmission of confidential information to an outside organization, an extranet can be used for this Web-based connection using Internet technology. The need for privacy and security in the implementation and ongoing maintenance of such sites is extremely important. Because client-related data are vulnerable in ways that can create legal liability, it is crucial that privacy and security issues be addressed not just as a priority, but as a necessity.

The use of intranets and extranets will continue to vary among different types of law practices. The work of the lawyer will be generally the same, but the technology used can be expected to change for some aspects of the work. Although businesses have used Internet technology for intranets and extranets for a number of years, the size of the organization has been directly related to the perceived need for these options. That same consideration is true of law firms. A larger firm would have a greater need for an extensive intranet that would be true of a small firm.

SUMMARY OF ONLINE RESOURCES

A wide range of resources are now available on Internet Web sites for use by law firms. Listed below are some of the most extensive online sources of information, including the site addresses and descriptions of the sites. The topics shown are not a complete listing but are intended to emphasize the quantity of materials now available through use of the Internet.

http://www.abanet.org/home.html

American Bar Association; provides a vast array of resources related to the topics discussed above, plus

- Law school accreditation
- Continuing legal education
- Information about the law
- Programs to assist lawyers and judges in their work
- Initiatives to improve the legal system for the public

http://www.lawforum.net
Law Forum; directory of law-related Web sites; has resources by state and by "type," including

- Law schools
- Bars
- Journals
- Courts
- Movie/film

http://www.lawnewsnetwork.com/
Law News Network; law news and information; has articles and the latest news on various legal topics; has links to practice centers, with sublinks; e.g.

Intellectual Property practice center has these practice center areas:

- Copyrights
- International and state laws
- Media law
- Patents
- Trademarks
- Trade secrets

Tech Law practice center has these practice center areas:

- Biolaw
- Cable/telecom
- Computer law
- Cyber intellectual property
- Cybercrimes
- Cyberspeech
- Domain names
- E-commerce
- Internet regulation
- Licensing
- Security/encryption
- Web sites/linking

http://www.lawguru.com
Legal research and resources; various tools and other resources, such as

- Access to over 500 legal search engines, tools, and databases
- Links to thousands of legal sites
- Daily legal news
- Legal forms
- A library of articles on various legal topics

http://patents.uspto.gov/
Web site for the U.S. Patent Office; contains a "how to . . ." section for finding patents and trademarks, getting patents and trademarks, etc., and allows searches for

- Patents
- Trademarks
- Manual of Patent Classification
- Forms
- Fees

http://www.law.cornell.edu/topics/
Legal Information Institute; has a keyword search capability as well as links to general topics such as

- Enterprise law
- Intellectual property
- Taxation
- Family law
- Employment law
- Criminal law
- Legal education

http://www.patentlawlinks.com/
A Web site for patent attorneys; links to Web sites to access resources such as

- Case law
- Patent search engines
- Worldwide patent office sites
- Statutes and regulations
- Journals
- Legal search engines
- Patent-related forms

http://findlaw.com/
Resources for four categories of Web site users:

1. Legal professionals
   - Legal subjects such as intellectual property, criminal law, litigation, immigration law, and gaming law
   - Continuing legal education
   - Software and technology
   - Laws—cases and codes
   - U.S. federal and state resources
   - Legal organizations
   - Forms
2. Students
   - Law schools
   - Law reviews
   - Law student resources
   - Outlines and exams
   - Employment
3. Business
   - Business formation
   - Finance
   - Intellectual property
   - Human resources
   - Corporate contracts
4. Public
   - Housing
   - Automobile

- Personal injury
- Family
- Work
- Immigration

GLOSSARY

Alternative dispute resolution Involves ways of settling disputes without going to court.

Intranet Used for internal Web-based activities in an organization.

Extranet Used for direct, secure transmission to an outside person or organization through the use of Internet technology.

Litigation Generally refers to the handling of lawsuits, including court appearances at various levels.

Patent Obtained by an inventor for some unique item, to keep others from producing or using that person's invention for some length of time.

Real property Commonly referred to as real estate.

CROSS REFERENCES

See *Legal, Social and Ethical Issues; Patent Law; Taxation Issues; Trademark Law.*

REFERENCES

ABA Center for Professional Responsibility (2002). Legal ethics. Retrieved May 25, 2002, from http://www.abanet.org/cpr/ethics.html

ABA Legal Technology Research (2002). Retrieved May 25, 2002, from http://www.abanet.org/tech/ltrc/

Beckman, D., and Hirsch, D. (June 2000). Web worries of dot-com lawyers. Retrieved May 25, 2002, from http://www.abanet.org/journal/jun00/tkdave.html

Bodine, L. (2000). Marketing your firm on the Net: The next step. *American Bar Association Law Practice Management Magazine*. Retrieved from http://www.abanet.org/lpm/magazine/magarchive_front.shtml/

Ernst & Young Litigation Advisory Services (2002). Retrieved May 27, 2002, from http://ey.com/

FindLaw for Legal Professionals (2002). Litigation. Retrieved May 27, 2002, from http://findlaw.com/01topics/29litigation/

Legal Information Institute (2002a). Alternative dispute resolution. Retrieved May 27, 2002, from http://law.cornell.edu/topics/adr.html

Legal Information Institute (2002b) Patent law. Retrieved May 26, 2002, from http://www.law.cornell.edu/topics/patent.html

PatentLawLinks.com(2002). IP law. Retrieved May 26, 2002, from http://www.patentlawlinks.com/

Taxsites.com (2002) Federal tax law. Retrieved May 26, 2002, from http://www.taxsites.com/federal.html

United States Patent and Trademark Office (2002). Patents. Retrieved May 26, 2002, from http://patents.uspto.gov/

FURTHER READING

Sanders, C. H. (2000). Trends in legal publishing for the millennium: Quality moves to the Internet. Retrieved 25, 2002, from http://www.abanet.org/tech/ltrc/oresearch.html

Legal, Social, and Ethical Issues

Kenneth Einar Himma, *University of Washington*

INTRODUCTION

The creation of the networked world has produced a host of social and individual benefits. The World Wide Web, for example, makes it possible for individuals to access a wealth of information at any time of the day from the convenience of their own homes. E-mail capabilities enable individuals to communicate information from one country to another at a fraction of what it would cost to make an international telephone call. Bulletin boards and discussion lists provide large groups of like-minded persons with a convenient forum in which ideas can be exchanged at any time of the day.

The unique capabilities of these new technologies, however, have also created a host of novel legal, social, and ethical problems. The ability of users to communicate information to large audiences facilitates many legitimate purposes, but it also facilitates many purposes of questionable morality and legality. The fact that a user can, in principle, communicate information via the Web to millions of people around the world creates a variety of problems. Since, for example, the harm to an individual's reputation caused by defamation is a function of how many people receive the defamatory material, the capabilities of the Web increase a user's potential for causing harm by publishing defamatory material. Similarly, since the economic loss to a copyright holder caused by the unauthorized distribution of copyrighted materials is also determined by how many people receive those materials, the capabilities of the Web also increase a user's potential for causing harm by sharing copyrighted materials.

Not surprisingly, use of these new information technologies implicates a wide variety of interests of legal, social, and ethical significance. These interests include privacy, free speech, security, economic well-being, and intellectual property. This chapter discusses some of the important ways in which these technologies have unfavorably implicated these interests.

FREE SPEECH ON THE INTERNET

While the establishment of the World Wide Web has greatly enhanced the ability of ordinary individuals to receive and communicate information and ideas, online speech has begun to pose problems. Unfortunately, the Internet has been used not only to seek information and truth, but also to defame, defraud, sexually exploit, and incite other people. It is not surprising, then, that such conduct has given rise to a number of controversies regarding the scope of the right to free speech.

Introduction: Legal Protection of Free Speech

There is a broad consensus among people in Western nations that citizens have a moral right to free speech that deserves legal protection. A number of nations, including Canada and the U.S., have formal constitutions that explicitly create a legal right of free speech limiting what the state may do in restricting the free flow of ideas. The Canadian Charter (Canadian Charter of Rights and Freedoms, 1982), for example, provides that "Everyone has the following fundamental freedoms: (a) freedom of conscience and religion; (b) freedom of thought, belief, opinion, and expression, including freedom of the press and other media of communication; (c) freedom of peaceful assembly; and (d) freedom of association" (part I, section 2). The First Amendment to the U.S. Constitution (United States Constitution, 1789) provides that "Congress shall make no law . . . abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances" (First Amendment).

While formal constitutions protecting free speech are less common in Europe, European nations also tend to regard the free flow of information among individuals as a fundamental value. Indeed, a recent draft of the proposed Charter of Fundamental Rights of the European Union (Draft of the Charter of Fundamental Rights of the European Union, 2000) includes a provision establishing a right of free speech: "Everyone has a right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers" (chapter 2, article 11). Free speech is, in any event, also protected by the courts in many European nations as an "implied right."

Justifications for Free Speech Rights

Utilitarian justifications for protecting speech emphasize the benefits of free speech to the common good. John Stuart Mill (1989), for example, argues that allowing people the freedom to speak their minds and express their creative abilities promotes human happiness in several ways. First, allowing speech conduces the speaker's well-being by facilitating the development of the speaker's critical faculties. Second, allowing creative expression promotes the ideological and technological betterment of humanity. Third, and most importantly, allowing free debate of ideas increases the likelihood that the truth will be discovered—and knowing the truth is always conducive, in Mill's view, to human well-being.

Deontological justifications argue that there are strict moral limits on the extent to which one person may justifiably interfere with the autonomy of another person. Robert Nozick (1977) argues, for example, that every autonomous moral agent has a "natural" right to liberty that includes the right to express personal views free of coercive interference. Since people have natural rights by virtue of their status as moral agents and not by virtue of their status as citizens of some state, the operation of natural rights is not limited to obligating other individuals; even the state must respect a person's natural rights. Indeed, so critically important are these rights that Nozick believes that the only legitimate function of the state is to provide institutionalized coercive protection of each person's natural rights.

Contract theorists justify a right to free speech in terms of citizen consent. While classical contract theories typically require that such consent be either express or tacit, the most influential modern approach focuses on what citizens would consent to. John Rawls (1999) argues that a just state is bound by those principles that rationally self-interested agents would choose if they had to select principles constraining the government without having any specific information about their own particular preferences, abilities, and social circumstances. In Rawls's view, citizens choosing from behind the "veil of ignorance" would protect themselves from oppression by selecting a principle that establishes a right to personal liberty, which includes a right to free speech. In particular, Rawls believes that citizens would choose a principle (the liberty principle) that grants each agent as much freedom as is compatible with like freedoms for all other agents.

Internet Issues

Pornography on the Web

Pornographic material on the Web is both plentiful and easy to come by. A search of the word "sex" on any mainstream search engine will produce not only links to scientifically and medically useful information on sex, but also links to pornographic Web sites. Entering "http://www.sex.com" into a Web browser will take the user to a portal that provides links to various different types of pornographic material, ranging from standard sexually explicit material to material that appeals to various "fetishes." As is evident, the word "sex" can be used in a variety of Web-based devices to locate Web sites specializing in graphic sexual content.

This raises the worry that children who are emotionally unprepared for pornographic content may accidentally or intentionally encounter it while surfing the Web. While it is fair to say that the effects of pornography on adults remain unclear, child development experts agree that chronic exposure to pornographic materials can have long-lasting and harmful effects on children (Benedek & Brown, 1999). The ease with which such materials can be accessed on the Web creates the possibility of such exposure in children—and its harmful consequences.

Concerns about the potential impact of Web pornography on the emotional well-being of children are shared by legislators on both sides of the Atlantic Ocean. The European Parliament, for example, is on record as strongly advocating the enactment of legislation that would protect children from access to inappropriate sexual material on the Web. In particular, the European Parliament (1997) passed a resolution stating that "minors should be protected as soon as possible against access, via the new networks and services, to material which may harm their physical and psychological development" (Clause D, Resolution on the Commission Green Paper on the Protection of Minors and Human Dignity in Audiovisual and Information Services, 1997).

The U.S. legislature has taken the most aggressive steps to reduce the chance that children encounter pornography on the Web. The Communications Decency Act (CDA), for example, prohibited the "transmission of any comment, request, suggestion, proposal, image, or other communication which is obscene or indecent, knowing that the recipient of the communication is under 18 years of age" and the "knowing [transmission] to a specific person or persons under 18 years of age . . . [of] any comment, request, suggestion, proposal, image or other communication that, in context, depicts or describes, in terms patently offensive as measured by contemporary community standards, sexual or excretory activities or organs" (Communications Decency Act, 1996, 47 U.S.Code section 223(a) and (d)). CDA authorized a prison sentence of up to two years for violations.

The U.S. Supreme Court (the Court) held that CDA is unconstitutional on the ground that its language was overbroad and would restrict the constitutionally protected speech of adults (Reno v. ACLU, 117 S.Ct. 2329, 1997). Crucial to the Court's decision was the observation that the Web is entitled to greater First Amendment protection than television and radio because one must take a series of affirmative steps to view specific content online. Children are far less likely to be accidentally exposed to sexually explicit material on the Web than on television, for example, since they cannot randomly sample Web pages simply by changing channels. While curious children can, of course, seek out such materials, it is the "ease with which children may obtain access to broadcasts" that, according to the Court, "justifie[s] special treatment of indecent broadcasting" (Reno, 117 S.Ct. at 2342).

In 1998, Congress responded by enacting the Child Online Protection Act (COPA), which makes it illegal for commercial Web sites to allow persons under 17 to view sexually explicit materials that are "harmful to minors." COPA's restrictions are narrower than CDA's in three important respects: (1) While CDA applied to the entire Internet, COPA applies only to Web sites; (2) while CDA applied to all sites, COPA applies to only commercial sites; and (3) while CDA restricted "indecent" and "patently offensive" speech, COPA restricts only speech that is "harmful to minors."

The constitutionality of COPA is presently unsettled. Although the Court rejected the argument that the phrase "harmful to minors" is overbroad in Ashcroft v. ACLU (Ashcroft v. ACLU, 122 S. Ct. 1700, 2002), the decision established no more than that the "harmful to minors" standard was not itself constitutionally defective; as the Court put it, "COPA's reliance on community standards to identify 'material that is harmful to minors' does not *by itself* render the statute substantially overbroad for purposes of the First Amendment" (Ashcroft, 122 S. Ct. at 1713). Accordingly, the Ashcroft decision leaves open the possibility that the Court might declare COPA unconstitutional on other grounds. (The Court was favorably impressed by the fact that the definition of the crucial expression "harmful to minors" incorporates principal elements of the constitutional standard for obscenity. COPA defines material as harmful to minors if and only if "(A) the average person, applying contemporary community standards, would find, taking the material as a whole and with respect to minors, is designed to appeal to, or is designed to pander to, the prurient interest; (B) [it] depicts, describes, or represents, in a manner patently offensive with respect to minors, an actual or simulated sexual act or sexual contact, an actual or simulated normal or perverted sexual act, or a lewd exhibition of the genitals or post-pubescent female breast; and (C) taken as a whole, [it] lacks serious literary, artistic, political, or scientific value for minors." The language here incorporates key elements of the obscenity test established by the Court in Miller v. California 93 S.Ct. 2607, 1973.)

Online Hate Speech

There are a growing number of racist, anti-Semitic, and heterosexist Web sites worldwide that advocate, typically in reprehensible language, violent measures to achieve hate-inspired political agendas. White supremacist Web sites, for example, frequently call for "race wars." Extremist antiabortion Web sites feature photographs of women coming to or leaving from abortion clinics. The most infamous of these sites once posted a "wanted list" of abortion doctors and currently features an editorial advocating the arrest (which it describes as "an act of love") of all persons with a same-sex sexual preference. (The infamous Nuremberg Files Web site can be found at http://www.christiangallery.com.)

Online hate speech creates a host of social problems. To begin with, the anonymity of online communications has emboldened racists, anti-Semites, and homophobes to create more hate sites with increasingly egregious content. Further, the worldwide availability of such content enables bigots to find one another with unprecedented ease and to more easily reach people whose educational and economic circumstances make them susceptible to content that identifies an "other" as scapegoat. Finally, many hate sites explicitly encourage violent and terroristic acts to achieve their political agendas.

Different countries have adopted different approaches to the regulation of online hate speech (see Biegel, 2001, for an outstanding analysis of the social and legal issues presented by hate speech). Legislation specifically restricting hate speech is not feasible in the U.S. According to prevailing interpretations of the U.S. Constitution, hate speech does not fall into any of the traditional categories of unprotected speech and hence appears to be protected by the First Amendment. Insofar as online hate is protected by the First Amendment, laws that target online hate sites on the basis of their content are constitutionally impermissible in the U.S.

The U.S. is unique among western industrial nations in this regard. Many European nations, such as Germany, have laws criminalizing certain forms of hate speech. Indeed, the Council of Europe is currently preparing legislation that would prohibit posting or distributing racist or xenophobic material through a computer system (see http://www.coe.int/T/E/Communication_and_Research/Press/Theme_Files/Cybercrime/Index.asp for the Council of Europe's Web site on cybercrime). For its part, Canada has already enacted extensive legislation against online hate. The Canadian Human Rights Act (1996–77) prohibits the communication over computer networks of "any matter that is likely to expose a person or persons to hatred or contempt by reason of the fact that that person or those persons are identifiable on the basis of a prohibited ground of discrimination."

Spam

Unsolicited mass commercial e-mailings ("spam") are ethically problematic because they impose significant costs on consumers. First, spam consumes scarce network resources (e.g., time and disk space) for which recipients must pay. Second, recipients must expend significant quantities of time and energy to deal with these unsolicited mailings; as Richard Spinello (2000, p.63) points out, "If a vendor sends out 6 million messages and it takes 6 seconds to delete each one, the total cost of this one mailing is 10,000 person hours of lost time." Third, ISPs and other consumers are harmed because large quantities of such mailings can overload networks, slowing response rates and even causing downtime. Spam raises special ethical issues because these costs are typically imposed on consumers without their consent. (The proliferation of pop-up advertisements on the Web raises some similar issues. For example, the consumer must expend time and effort to close a pop-up ad; in this respect, pop-up ads are like spam. Even so, there is one fundamental difference between the two: to receive a particular pop-up ad, one must perform a very specific cyberact—namely, direct a Web browser to retrieve the contents of a particular Web site. Receipt of spam, in contrast, does not require any specific act on the part of the user, apart from having an e-mail account.)

Intentionally deceptive spam raises additional ethical issues. Many senders attempt to ensure that recipients view their mailings by concealing their commercial nature. Such practices are ethically problematic not only because they are dishonest, but also because they deliberately attempt to frustrate the intent of consumers who wish to save time online by deleting spam mailings without viewing them.

Western nations agree on the need to restrict spam, but disagree on what sorts of restrictions are legitimate. The European Union recently adopted a general directive that requires spammers to obtain the consumer's consent before sending unsolicited commercial e-mail (Saunders, 2002). While such legislation is not possible in the U.S. because commercial speech is protected by the First Amendment, intentionally deceptive commercial speech can be prohibited. In 1998, for example, the Washington State Legislature prohibited the transmission of unsolicited commercial e-mail to Washington state residents that contains false sender addresses or deceptive subject lines. Washington's "Unsolicited Commercial Electronic Mail Act" was upheld by the state's highest court in 2001.

Filtering Devices

There are a number of products that filter objectionable content online. Individual filtering programs, for example, can be installed on a user's personal computer to block access to Web sites using certain sexually explicit terms or featuring images with a disproportionate quantity of flesh tones. Additionally, an increasing number of Web sites, though still a comparatively small percentage of total Web sites, participate in the Platform for Internet Content Selection ratings system, which suppresses Web sites that rate themselves for mature audiences, as well as Web sites that decline to rate themselves.

The most contentious issue regarding such products involves their use by public libraries. While many persons believe public libraries should install such devices out of respect for the values of the communities they are intended to serve, library professionals and associations frequently oppose their use on two grounds. The first is that filtering programs are imprecise at this point in time: they sometimes fail to block access to pornographic content and sometimes block access to unobjectionable scientific or health-related content. The second is that there is a concern about whether it is ever appropriate for a library to censor content. The American Library Association, for example, takes the position that "A person's right to use a library should not be denied or abridged because of origin, age, background, or views" (American Library Association Library Bill of Rights, Section V [1996]). In this view, any form of censorship is inconsistent with the duty of libraries to ensure the free flow of information to individuals of all ages.

The U.S. Congress has recently entered the filtering controversy by enacting the Children's Internet Protection Act (CIPA), which requires public libraries to use filtering devices as a condition for receiving certain federal funds. A federal district court, however, recently struck down CIPA on the grounds that filtering devices typically screen out protected content. See American Library Association v. United States, 201 F.Supp.2d 401 (2002). The court's holding suggests that library use of filtering devices could be mandated only if such devices filter out no protected material; thus, it leaves open the possibility of mandating use of such devices once filtering technology becomes more accurate. The case has been appealed to the Supreme Court.

INTELLECTUAL PROPERTY

Introduction: Legal Protection of Intellectual Property

Western nations offer a number of different types of legal protection for intellectual property. First, patent law protects a person's interests in his or her own inventions, which includes newly designed useful processes. (The intellectual property statutes of various English-speaking nations are available at the following Web sites: Canada—http://strategis.gc.ca/sc_mrksv/cipo/welcome/welcom-e.html; United States—http://www.law.cornell.edu/topics/topic2.html; United Kingdom—http://www.intellectual-property.gov.uk/.) A patent protects a person's invention by granting the inventor a limited monopoly power over the invention. This allows a patent-holder the right to prevent other firms and persons from making or marketing it. Similar protections ("design patents" in some nations, "designs" in others) are also available for various aspects of a product's ornamental design.

Second, trademark law protects the right of a product or company owner to use marks that distinguish its goods and services from others. Trademark infringement generally occurs when a firm or individual uses a mark that is likely to confuse a reasonably intelligent consumer about the source or sponsorship of a good or service. Many nations allow a firm to establish a trademark simply by using it, but formal registration is frequently available as well.

Third, copyright law protects the original expression of ideas and facts (i.e., the particular form, language, and structure of articulated ideas). It is crucial to note that copyright protection is never given to the underlying idea or facts themselves. Others are free to express the same ideas and facts as long as they do not intentionally duplicate the author's original expression. Originality of expression typically requires the introduction of something new to the world; a person who simply copies the expression of some other person has not produced anything original that can be copyrighted. Copyright protection typically applies to original literary works (including computer programs), musical works, dramatic works, choreographic works, artistic works, sound recordings, and architectural works.

Copyright law typically defines two intellectual property rights. First, it grants to authors an exclusive right to reproduce, modify, distribute, perform, and display the protected work. Second, it grants to authors certain moral rights of authorship. Such entitlements may include the right to claim authorship in the protected work; the right to prevent use of the author's name as the author of any work he or she did not create; the right to prevent use of the author's name as the author of a protected work if it has been modified; the right to prevent intentional distortion, mutilation, or modification of a protected work that would detract from the author's reputation; and the right to prevent intentional or grossly negligent destruction of a protected work.

Most nations include an exception for what are deemed "fair" uses (see, e.g., the "Fair Dealing" exception to Canadian copyright law beginning at section 29 of the Canadian Copyright Act (R.S. 1985, c. C-42) and the "Fair Use" exception to U.S. copyright law beginning at 17 U.S.C. section 107). Copyrighted material may generally be used for news, educational, and research purposes, provided that such material is not directly used for material gain. In determining whether a use of copyrighted material is fair or not, courts may look to the amount of material used relative to the copyrighted work as a whole and to the effect of the use on the market for the work. The last factor is especially important insofar as copyright is intended to protect an author's right to collect the economic value of his or her expression.

Justifications for Intellectual Property Protection

Theorists have produced three main lines of justification for legal protection of intellectual property. The first line is grounded in the Lockean view that persons have natural (or moral) property rights in their bodies and labor. Since the particular sequence of words or symbols chosen by authors to express an idea or fact is the product of their labor, they acquire an exclusive property right to that particular sequence of words or symbols—as long as no one else has a prior right to that sequence. The Lockean argument can thus roughly be summarized as follows: you made it; therefore, you own it.

The second line of justification for intellectual property rights is grounded in the Hegelian view that creators use symbols, words, and sounds to express their personhood. As an expression of personhood, an original creative work realizes and extends the creator's person. Since no one but the authors have a protected interest in their personhood, the authors acquire a protected interest in their expressions because such expressions are, in some vague but morally significant sense, literal extensions of their personhood. Roughly put, this line of argument can be summarized as follows: it is part of you; therefore, it is yours.

The third principal line of justification is utilitarian in character. Protection of intellectual property rights is justified because such protection is conducive to the common good. People will be far more likely to invest time and energy in creating the intellectual products that contribute so much to human happiness, flourishing, and well-being if they are granted an exclusive property right in their original creations. Roughly put, this line of argument is: giving it to you benefits society; therefore, it is yours.

It is notable, however, that intellectual property rights have recently come under fire from various quarters of cyberspace. The most extreme of these opponents is John Perry Barlow, who argues that existing intellectual property laws have no proper application in cyberspace. First, Barlow (1996) believes that cyberspace is a distinct metaphysical reality that lies beyond the proper jurisdiction of any nation: "Your legal concepts of property, expression, identity, movement, and context do not apply to us. They are all based on matter, and there is no matter here." Second, Barlow (1993) argues that information is itself a life-form that, like any other life-form, is entitled to some moral standing. In Barlow's view, information has a morally protected interest in freedom: "information," as he has notoriously put the matter, "wants to be free."

Additionally, a number of theorists worry that the increasing willingness of online users to infringe upon copyright law undermines the social legitimacy of protecting intellectual property in cyberspace. In this line of analysis, which is roughly grounded in social contract theory, the legitimacy of any particular law depends on its being acceptable to those whose behavior it purports to govern. But the growing tendency, especially among younger online users, to reproduce and distribute copyrighted works without permission suggests that they no longer accept the legitimacy of intellectual property protection. Insofar as there is no longer a consensus on the legitimacy of such laws, the contractual basis for protecting intellectual property is deteriorating.

Internet Issues

Domain Names

Every computer on a network has its own Internet protocol (IP) address consisting in a unique sequence of numbers and dots (e.g., 213.57.66.9384) that defines its location on the Web. When a user accesses a particular Web site, the contents of that site are sent from the host server's IP address to the IP address of the user's computer. In effect, then, IP addresses make it possible for networked computers to find each other, enabling users to access the contents of Web sites hosted at other locations on the network.

In most cases, users need not know a complicated IP address to access a Web site. Most Web sites have a natural language domain name (e.g., http://www.sportinggoods.com) assigned to its IP address that permits easier and more intuitive access to its contents. The user simply types in the natural language domain name and the ISP either looks for the corresponding IP address or submits a request to a "root server" that serves as a digital directory associating IP addresses and domain names. Once the ISP has determined the corresponding IP address, the desired site is accessed.

Domain names can be very valuable commodities. An intuitive domain name saves online consumers time and energy: it is much easier to find a site with an intuitive domain name than with a complicated IP address that is difficult to find and remember. The resulting convenience to users can naturally translate into economic benefits; the easier it is to access a commercial Web site, the more likely users are to visit and buy from that site.

Consequently, there have been a number of conflicts over the use and ownership of domain names. Early in the development of the Web, some people registered domain names featuring the trademarked names of large firms in the hope that the firms would buy those names whenever they decided to go online. (Compaq reportedly paid more than $3 million to purchase the domain name http://www.altavista.com from the former owner of Alta Vista Technology after acquiring the company. Until Compaq purchased the domain name, it was forced to use the ungainly http://www.altavista.digital.com.) Though a few such "cybersquatters" made a quick profit for their trouble, courts now treat the practice of speculating on domain names incorporating trademarks as actionable trademark infringement (see, e.g., Panavision v. Toeppe, 141 F.3d. 1316, 1998).

More commonly, a slightly modified version of a popular Web site's domain name is used to capture some of its traffic. One commercial pornographic Web site in the U.S., for example, uses the domain name "http:// www.whitehouse.com." Users who type "http:// www.whitehouse.com" instead of "http://www. whitehouse.gov" into their browsers—presumably a common mistake—will access sexually explicit material instead of the U.S. President's official Web site. By such means, Web site owners can dramatically increase traffic to their sites.

Though such practices can plausibly be characterized as deceptive, they can nonetheless be used for legitimate purposes of free expression. A user who mistakenly types "http://www.gwbush.com" instead of "http://www.georgebush.com" will access a site criticizing George Bush's views and policies instead of his personal Web site. While the commercial use of a domain name similar to a trademarked name can dilute the value of the trademark and is hence unethical, the politically motivated use of a domain name to express legitimate criticism is arguably unobjectionable—as long as users are not likely to be confused about the origin of the site.

Illicit Copying over the Internet

No case better exemplifies the clash between the intellectual property rights of copyright holders and the increasingly libertarian spirit of online users than the proliferation of MP3 file sharing over the Web. The development of the MP3 format was the first significant step in realizing the Internet's latent potential for online dissemination of music files. Earlier technologies offered little incentive to share music files: the files were too large to be uploaded and downloaded quickly, and the sound quality was generally inconsistent. MP3 technology, however, permits the compression of nearly perfect digital reproductions of sound recordings into small files that can efficiently be transmitted from one user to another.

Napster augmented MP3's capabilities by introducing true peer-to-peer (P2P) file sharing. Whereas users of earlier file sharing technologies had to download previously uploaded files from a central Web site or file transfer protocol site, Napster users could simply take music files directly from the computers of other users. Though a central server was needed to keep a searchable list of all the available MP3 files, its purpose was limited to helping Napster users find each other. Since users could share music files online without anyone needing to take the time to upload music files to some central server, Napster made it easier than ever before for large groups of users to share their sound recordings.

Napster's P2P networking capabilities also inhibited the efforts of recording companies to stop reproduction and distribution of their copyrighted materials. When music files had to be transmitted through a central server, recording companies could demand that the server's owner destroy copyrighted files or litigate an expensive civil suit. But because Napster eliminated the need for centralized storage of such files, there was no one entity that could be pressured by copyright holders. Not surprisingly, the music industry viewed Napster as a grave threat to the value of its copyrights.

The conflict came to a head when a group of music companies sued Napster for "indirect" violations of U.S. copyright law. Since Napster's role was limited to enabling users of Napster's MusicShare software to gain access to the hard drives of other users, the company could not be held liable for direct infringements. Instead, the plaintiffs sought to hold Napster liable for contributory infringement (i.e., knowingly assisting others in directly infringing a copyright) and vicarious infringement (i.e., benefiting financially from infringements when it has the ability to supervise and terminate users).

The litigation ultimately proved fatal to Napster. A U.S. federal court issued a preliminary injunction prohibiting Napster from assisting users in sharing copyrighted materials without the express permission of the owners (A&M Records v. Napster, 114 F. Supp. 2d 896, 2000). The court based its injunction on a prediction (as opposed to a final judgment) that Napster would lose at trial because (1) users were deriving an unfair economic benefit from using Napster by saving the cost of the relevant recordings and (2) Napster use was decreasing CD sales among users (Napster, 239 F.3d at 1017). Although the court's decision did not conclusively settle the material legal issues, the injunction effectively functioned as a death sentence for Napster. (Because the court did not issue a final judgment, music sharing technologies and Web sites continue to proliferate. Indeed, http://www.afternapster.com lists 32 file sharing Web sites, many of which improve on the P2P networking capabilities of Napster.)

Plagiarism

The availability of so much information on the Web has been of particular benefit to students. First, students have access to far more written academic materials than ever before; the availability of academic writings on the Web is a welcome supplement to the offerings of school and public libraries. Second, student research efforts are no longer tied to the operating hours of libraries; students can access a wide range of materials at any time from wherever they happen to be with their computers. A student with a cellular modem does not even need to be near land-based telephone lines.

But the Web also makes student plagiarism much more tempting. The Web not only provides easy online access to an abundance of quality writings, but also produces it in a form that is easy to plagiarize. Formerly, plagiarists had to take the time to copy a text sentence by sentence onto a medium that could be turned in as their own; not infrequently, this involved hours of time typing or writing the text. Now plagiarists can take someone else's text with a few keystrokes: They need to do no more than highlight the relevant text, enter the appropriate keystrokes to copy and paste it into a word processing document, and unethically claim it as their own work.

Student plagiarism also raises third-party ethical issues, as there are a number of Web sites offering original research papers for sale. One such Web site boasts: "Our 80 full time researchers are available 24 hours a day, 7 days a week. Order now and we will write your term paper within your specified deadline." (See http://www.term-paper-time.com. It should be noted that the site also states in ironically ungrammatical fashion: "We always urge them not to use the work as their own. Although the work we do is completely original and cannot be found anywhere else on the Web.") While students must obviously accept full moral responsibility for initiating the sequence of acts that culminates in plagiarism, Web sites offering term papers for sale are not beyond ethical reproach since the operators are presumably aware that there is a high likelihood that students will claim the work as their own.

Indeed, these sites seem to invite student plagiarism. After all, the only writing that is appropriately characterized as a "term paper" is writing that is being turned in as coursework; thus, advertising a piece of writing as a "term paper" is not unreasonably construed as a claim that the work is suitable, as written, to be turned in for a grade. If, as seems reasonable, knowingly aiding someone in committing an ethical violation is unethical, then offering term papers for sale in such a suggestive manner is also unethical. (Although there are now a number of Web-based products that assist instructors in detecting student plagiarism, these products will assist only in detecting instances of cut-and-paste plagiarism; they will not help in detecting instances where students have purchased specially written papers.)

INFORMATION PRIVACY

Introduction: Legal Protection of Information Privacy

The various constitutions differ with respect to how much protection they afford to privacy. The Canadian Charter and U.S. Constitution both contain a number of provisions that can be construed as concerned with protecting privacy. Each contains clauses protecting freedom of speech, thought, conscience, religious worship, as well as freedom from unreasonable searches and seizures (see the First, Third, Fourth, Fifth, and Ninth Amendments to the U.S. Constitution; and sections 2, 7, 8, 9, 11, and 13 of the Canadian Charter of Rights and Freedoms). Such protections are reasonably construed as being concerned to establish a zone or sphere of privacy in which a person's movements are protected against state intrusion.

Neither the Canadian Charter nor the U.S. Constitution, however, contains a provision that explicitly defines a privacy right in personal information or data. While the U.S. Supreme Court has grounded a general right of privacy in the "penumbras" of the various protections mentioned above, the constitutional right to privacy in the U.S. has most commonly been cited as a justification for invalidating laws that restrain reproductive freedom (see, e.g., Griswold v. Connecticut, 381 U.S. 479 [1965] and Roe v. Wade, 410 U.S. 113 [1973]). What protections there are in such nations for personal information are defined largely by statute and common law.

In contrast, a recent draft of the Charter of Fundamental Rights of the European Union provides explicit privacy protection of personal information. Article 8 provides that "(1) Everyone has a right to the protection of personal data concerning him or her [; and] (2) Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified." This draft of the Charter explicitly extends privacy protection, as a matter of constitutional right, to personal information.

It is worth noting that European statutory law provides comprehensive protection of a person's interest in information privacy. The European Parliament and Council of October 24, 1995, have issued a general directive governing the processing of personal information (the full text of the directive is available from http://www.privacy.org/pi/intl_orgs/ec/final_EU_Data_Protection.html). Chapter 1, article 1 provides that "Member states shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy, with respect to the processing of personal data." ("Processing of personal data" includes "collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction" [chapter 1, article 2b].) Among other things, the directive limits the purposes for which personal information on an individual may be collected to "specified, explicit and legitimate purposes" and guarantees a right of access of the individual to such data.

In notable contrast, the U.S. has adopted a far more conservative approach to protecting information privacy. Most recently, the U.S. enacted the U.S. Patriot Act in response to the terrorist attacks of September 11, 2001. (The text of the U.S. Patriot Act is available from http://www.eff.org/Privacy/Surveillance/Terrorism_militias/20011025_hr3162_usa_patriot_bill.html. For a critical evaluation of the implication of the Act on information privacy, see http://www.eff.org/Privacy/Surveillance/Terrorism_militias/20011031_eff_usa_patriot_analysis.html.) Among other things, the Patriot Act expands the surveillance capacities of the government in a number of ways. Section 213, for example, allows the government to conduct a search without notifying the subject that a warrant has been executed if "the court finds reasonable cause to believe that providing immediate notification of the execution of the warrant may have an adverse result." Similarly, Section 216 allows the government to monitor an individual's movements on the Web upon a showing that "the information likely to be obtained . . . is relevant to an ongoing criminal investigation." While the expansion of such investigative capacities may arguably further the interests of citizens in security, it also poses a direct threat to their interests in information privacy.

Justifications for Privacy Rights in Personal Information

Consequentialist theories justify privacy rights as necessary for a person's happiness and well-being. James Rachels (1975, p. 331) argues, for example, that protection of privacy rights is justified by our need to control the structure of our social relationships: "If we cannot control who has access to us, sometimes including and sometimes excluding various people, then we cannot control the patterns of behavior we need to adopt . . . or the kinds of relations with other people we will have." Moreover, it is commonly believed that protection of privacy rights is justified by a concern to prevent the embarrassment or offense that would be caused to a person by disclosure of certain facts.

Deontological theories of privacy, in contrast, take the position that private facts about a person are "nobody else's business" regardless of what the consequences of disclosure might be. Privacy should be respected, on this line of reasoning, not only because it is conducive to well-being, but also because persons are intrinsically valuable beings entitled to be treated as autonomous ends-in-themselves. The intrinsic value of each person, then, requires that certain facts be treated as private and subject to the control of that person—even if those facts might turn out to be extremely useful to other people.

Some theorists, however, caution that legal protection of privacy should be narrowly crafted to avoid unnecessarily restricting the free flow of information. Solveig Singleton (1998), for example, argues that the collection and dissemination of consumer information from business to business should not be restricted by privacy protections. Business dissemination of consumer information, in her view, is not morally distinguishable from ordinary gossip. Inasmuch as a legal ban on ordinary gossip would violate the right to free speech, so too would a ban on the dissemination of ordinary consumer information. Thus, she concludes, "[any] country that takes the freedom of information seriously cannot properly prohibit one business from communicating information about real events and real people to other businesses."

Public and Private Information

The general claim that personal information ought to be protected by law does not, by itself, tell us much about how to determine what information about a person deserves legal privacy protection. For example, the altogether plausible consequentialist claim that protection of information is justified by a personal need to control the structure of various social relationships says little about what information about a person ought to be protected by the law. For this reason, general justifications of privacy rights, such as discussed in the last section, represent only a starting point in determining what content privacy law ought to have.

It is reasonable to think that whether a person ought to have a protected privacy right in a piece of information depends in part on the character of that information. Some facts about a person are generally accepted as private facts in which a person has a legitimate expectation of privacy. Since, for example, I am entitled to draw my drapes to prevent people from viewing what is going on in my home, the facts about what is going on my home are private—at least when the drapes are drawn (I can, of course, always voluntarily make private facts public by leaving my drapes open) and my behavior is lawful. Thus, I have a legitimate expectation of privacy in aspects of my behavior that I may rightfully prevent people from viewing; these aspects of my behavior define private facts.

Some facts, however, should be regarded as private in virtue of their intimate character. It is almost universally accepted that certain physical functions, such as those involving the sexual and excretory organs, express private facts because of their felt intimate character. Information regarding a person's physical and emotional health is also widely regarded as private information that she should be entitled to control; indeed, so intimately vital are these facts that medical professionals are charged with a legal duty of confidentiality.

While many privacy issues concern private facts, others are concerned with information of a significantly different character. Some information that is contained in public records concerns matters that most individuals would regard as sensitive. For example, many people are reluctant to make their debt history readily available to anyone who happens to be curious about it—and this is especially so if that history includes a bankruptcy. Likewise, many people who have paid their debt for criminal offenses are reluctant to make their criminal records easily available out of a concern that such information would be used to discriminate against them.

The issue of whether someone has a moral right to control a particular piece of information that ought to be protected by the law thus depends on a variety of considerations. It depends not only on broad theoretical arguments regarding the general justification of intellectual property, but also on the character of the particular piece of information and how that information might be used by other persons. Such determinations present difficult issues of policy and ethics.

Internet Issues

Corporate Use of Personal Information

Many commercial firms collect information from visitors to their Web sites, which is stored in small data files called "cookies" and deposited on the visitors' own computers. These files typically contain information, such as passwords, on-site searches, dates of previous visits, and site preferences, that can be used by the firm to customize the user's experience when she revisits its site. For example, a bookselling Web site might store a list of previous searches on the user's computer so that it can be accessed by the site on subsequent visits to generate a list of books to recommend to the user. This enables the site to provide what it considers to be better service to users by tailoring their Web environment to personal preferences as expressed in the history of visits to the site.

Though the use of cookies thus has a plausible business rationale, it raises a number of ethical issues. Typically, cookies are transmitted from the user's hard drive to the site and retransmitted (possibly with modifications) from the site to the user's hard drive in a way that does not interrupt the user's browsing experience. This means that, in many cases, the user's hard drive is being modified without the user being aware of it—and, more importantly, without user consent. The idea that someone else can, in essence, modify the user's property without user consent raises, to begin with, ethical issues concerning the user's property rights over the contents of the user's computer.

Moreover, some theorists worry that the use of cookies to keep information on the consumer raises privacy issues. As Richard Spinello (2000) put the matter, "cookie technology is analogous to having someone follow you through the mall with a video camera" (p. 111). In both cases, the technology keeps information on where you have gone, what you have looked at, and what you have purchased. If pressed, most people would likely have a difficult time explaining why they find this offensive; after all, this information appears to be significantly different from the sort of information that can be obtained by looking into a person's bedroom or bathroom. But if the amount of press generated by the cookie controversy is any indication, many people experience similar discomfort with cookies. To the extent that one has a legitimate expectation that one's movements in a public mall not be recorded, it can reasonably be argued that one also has a legitimate expectation that one's movements in cyberspace not be recorded.

While it is possible for users to set up their browsers to refuse cookies or to alert them whenever a site attempts to store a cookie, this can cause inconvenience to the user. Refusing all cookies restricts the user's options in cyberspace as some Web sites cannot be viewed without accepting cookies. Setting the browser to ask before accepting cookies can result in frequent interruptions that radically change the quality of the browsing experience. Many users who restrict cookies find that the disutility associated with such frequent interruptions outweighs, at least in the short run, their privacy concerns and restore their browsers to the default setting that allows for unrestricted cookies.

Though there is thus a sense in which users who decline to configure their browsers to refuse cookies can be presumed to consent to cookies, such consent is of questionable ethical significance. If the initial choice between A and B is not an ethically acceptable one, then the fact that a person voluntarily chooses A does not necessarily entail that the person consents to A. For example, the fact that I voluntarily choose to give a robber my money if my only other choice is being shot does not entail that I have, in any ethically significant way, consented to give the robber my money. Consent is ethically significant only to the extent that it is rendered in an antecedent choice situation that is ethically acceptable. Thus, if the choice between accepting cookies and not being able to browse a Web site efficiently is not an ethically acceptable choice to unilaterally impose on a user, then the fact that the user chooses to accept cookies does not entail ethically meaningful consent.

More troubling to privacy advocates than the data kept by any one firm, however, is the possibility that it could be combined with the information of other firms to create a comprehensive file about a user. To continue Spinello's analogy, this is analogous to having your movements in every store and mall recorded by a video camera and then keeping all those recordings in one central location that can be accessed by other persons. The more information about an individual that is centrally located and available for use by other persons and firms, the more likely it is to strike individuals as involving a breach of their privacy.

Notably, there are economic forces pushing in that direction. Businesses realize that consumer information is a valuable commodity and have evinced a growing willingness to sell it. Information about a consumer's buying and browsing habits can be used to tailor advertisements and mailings to the consumer's particular tastes and preferences, arguably serving both the consumer and the firm. It is not surprising, then, that trading in information itself is becoming an increasingly profitable venture—not only for firms specializing in information commerce, but also for ordinary firms specializing in other areas—and hence increases the likelihood that businesses will compile comprehensive files of personal information on individuals. (Online privacy services, such as Trust-E and BBB Online Privacy Program, rate various Web sites according to whether they agree to disclose their policies regarding the collection and dissemination of personal information by businesses. Trust-E is located at http://www.truste.org/, while BBB Online is located at http://www.bbbonline.org/.)

State Databases

State databases raise a different set of issues since they contain only information that is "public" in an ethically meaningful sense. In most cases, what is at issue is the disposition of information that the public has a right to collect through its official state representatives. Thus, the information contained in state databases, if there legitimately, is information to which the public has some sort of antecedent claim. This distinguishes privacy issues involving state information from privacy issues involving corporate use of information that is not public in this sense.

Though state information is a matter of public record, privacy advocates believe that governments should take strong steps to protect the privacy of driver's license numbers, birthdates, official identification numbers (e.g., social security numbers), and other identifying information frequently used to access sensitive information about a person. This kind of information is uniquely subject to abuse: all that an identity thief in the U.S., for example, needs to obtain credit cards in another person's name is that person's social security number and date of birth. With those two pieces of information, an identity thief can inflict long-term damage to a person's financial health and credit record.

The online dissemination of information already available to the public also raises privacy concerns. As noted above, privacy advocates are opposed to posting a person's public records online on the ground that such information can be used for discriminatory purposes. The availability of criminal records on the Web, for example, increases the likelihood that employers will discriminate against persons with criminal histories who have paid their debts to society. (The ready availability of such information on the Web could increase the probability of recidivism. If employers unfairly refuse to hire persons with criminal records, they are likely to reoffend.) Indeed, even a person's marital history can be used to discriminate against a person. A landlord might refuse to rent to an older person who has never been married out of a suspicion that he or she might be gay. It is the possibility of such discrimination that, in this line of reasoning, requires granting a protected privacy interest in information that is admittedly public in one sense.

Such records have, of course, always been available to the public, but the risk of misuse increases with the ease with which these records can be obtained. As a general matter, most persons are not willing to incur the inconvenience of visiting a courthouse and asking for a person's criminal or marital records. Since posting such information on the Web eliminates such inconvenience, it dramatically increases the likelihood that people will seek such information and, a fortiori, the likelihood that people will abuse it. For this reason, privacy advocates oppose posting this sort of public information on the Web.

These privacy concerns are exacerbated by the amount of information that might be made available on the Web. When it comes to privacy, the qualitative difference between what is and is not ethical is sometimes a matter of quantity. As Spinello's mall analogy suggests, many people believe that it would violate a person's legitimate interests in privacy to have his or her movements in a mall recorded and publicized. In relating this analogy to the issue of whether public records should be made available on the Web, it is crucial to emphasize that information about a person's movements around a mall is, in some sense, "public" information that is freely available to whoever might happen to be there.

Such worries are further exacerbated by the ease with which various records might be obtainable on the Web. Here it is worth noting that a number of companies sell software that purports to enable people to obtain, among other things, public records that include criminal history, debt history, real property acquisitions, and marital records. (One Web site boasts of a program, which it calls "the Internet's best selling spy software" that will allow persons to find "driver's records, lawsuits, criminal records, asset identification, . . . tax liens, . . . and court documents" (http://www.oddworldz.com/landoh34/learn.html).) The idea that so much sensitive information about an individual can be obtained with just a few keystrokes makes many people rightly uncomfortable because of its susceptibility to being misused. Accordingly, privacy advocates argue that the easy availability of such information poses a significant threat to an individual's legitimate interests in information privacy.

Encryption Programs and Public Policy

A user's privacy can be violated online in yet another way. Ordinary means of communicating over the Internet are surprisingly insecure. E-mail messages are typically routed through many servers en route to their final destination. This raises the possibility that such messages could be intercepted and read by persons other than the intended recipient. For example, hackers or even system administrators could breach a user's privacy rights by reading the user's confidential e-mail.

One means for preventing these violations of privacy is the use of encryption programs. The most popular encryption programs function by means of an electronic binary "key" that maps strings of linguistic symbols into unintelligible code that can be deciphered only by someone who has the key. Senders and recipients who share a viable key, then, can communicate privately by means of encrypted messages.

While privacy advocates generally oppose restrictions on such technologies, some legislators and citizens worry that more sophisticated encryption systems can be used to facilitate threats to national security. Since, for example, encryption software using a 128-bit key is virtually unbreakable, terrorists could use a system with a 128-bit key to communicate their plans via e-mail with only a negligible risk of discovery. Many regard this as an unacceptable risk to bear for the sake of online privacy in a post-September-11th world. Consequently, some officials favor legislation that prohibits the export of more sophisticated encryption software without providing state access to the keys.

SEARCH ENGINES

Search engines are indispensable tools for sifting through the approximately 8 billion pages now on the Web (Suzukamo, 2002). (In 1999, S. Lawrence and C. L. Giles estimated that there were 800 million Web sites [Lawrence & Giles, 1999]. If these estimates are accurate, the number of sites on the Web is 10 times as large as it was three years ago.) If users know which sites contain the information needed and the precise URL addresses of those particular sites, users can simply enter the addresses into their browsers and it will access those sites. But if, as is more often the case, users do not know where to go to find the information needed, they must rely on a search engine to find appropriate Web sites. Navigating the Web without the use of a search engine can be a time-consuming process that fails to produce appropriate sites.

A search engine imposes order on the Web by creating a large database that contains an index of each page's URL along with a fairly substantial list of keywords describing its contents. When the user submits particular keywords to a search engine, the engine returns a ranked list of URLs containing those keywords. Search engines structure the Web by, in effect, characterizing each indexed page in terms of its content (as indicated by the appropriate keywords).

Search engines, then, determine which pages users are likely to visit in two ways. First, search engines make available only those pages they have indexed—and there is currently no engine that has indexed every page on the Web. (Google, one of the most comprehensive search engines, states that it has indexed approximately 3 billion Web pages; see http://www.google.com/help/features.html.) For many people, a page that is not listed on any of the search engines does not exist; as Lucas D. Introna and Helen Nissenbaum (2000, p. 170) put the matter, "to exist is to be indexed by a search engine."

Second, users are more likely to visit only highly ranked pages. Anecdotal evidence suggests that users are likely to visit only the top 10 to 20 pages (or "hits") returned by a search engine. If a low-ranked page exists for users, its existence is far less substantial than that of a highly ranked Web site.

Insofar as the criteria that determine the inclusion and ranking of pages determine how frequently a Web site is visited, they have two consequences of ethical significance. First, they determine to what extent a particular publisher's Web-speech is received. Second, they determine what content is ultimately available to persons seeking information on the Web. Given these critical effects, one can reasonably argue that search engines ought to employ ranking and inclusion criteria that satisfy both technical and ethical standards.

Most search engines employ criteria that do not explicitly favor any particular class of publishers over another for reasons unrelated to merit. Web sites are usually either manually submitted to search engine editors who decide whether they are suitable for indexing or are retrieved by "spiders" that crawl the Web automatically indexing pages. A page's rank is usually determined by either the number of times a particular keyword appears in the page or the number of other pages that contain links to the page. (Some newer search engines attempt to evaluate a page's "authority." Teoma, for example, ranks a site based on the number of same subject pages, as opposed to general pages, that reference it. The goal is to be able to identify pages that are regarded as "expert" among a particular subject community. See http://www.teoma.com.)

But some search engines use economic criteria that favor firms willing to pay. Some engines, for example, allow a Web site to pay for expedited indexing—a process that can otherwise take months. Some engines permit firms to buy advertising linked to keywords; any search using a particular keyword will turn up a screen with an advertisement from a company that sells a related product or service. And some search engines have gone so far as to allow firms to bid on their top rankings.

Critics of economic criteria have raised two distinct ethical worries. First, such criteria discriminate unfairly against Web sites that are unable or unwilling to pay to be indexed or favorably ranked. Second, as Introna and Nissenbaum (2000, p. 179) put the point, economic criteria compromise the ideal of the Web as a public good insofar as the Web "fulfills some of the functions of other traditional public spaces—museums, parks, beaches, and schools" and thereby contributes to the common good. On this view, selling influence on the Web is as ethically problematic as selling influence in a museum or school: in neither case should the common good be sold to the highest bidder.

HACKERS AND SECURITY

There are a growing number of well-publicized incidents in which hackers obtain unauthorized entry into a firm's or state agency's servers. Some of these incidents involve comparatively innocuous exploration of a network's structure; in such cases, hackers look around and leave without altering the system. Others involve the commission of computer pranks; one such famous incident involved an insulting message left by hackers on the New York Times Web site. Yet others involve the commission of cyberterrorism that threatens national security, as when a hacker breaks into a government network that stores classified material, or individual well-being, as when a hacker breaks into a corporate server and takes credit card and bank account numbers. (Some people distinguish "hacking" from "cracking." Cracking, unlike hacking, involves a malicious purpose: the intent is to gain entry to a network to cause harm or damage. In contrast, hacking is motivated primarily by curiosity.)

The ethical quality of such behaviors seems quite easy to characterize. While the more malicious of these acts involve very serious ethical transgressions because of the harm they are intended to cause, all seem morally objectionable because they constitute an electronic form of trespass onto the property of another person. To obtain unauthorized entry into some other person's network seems, from an ethical perspective, straightforwardly analogous to uninvited entry onto the real property of another person. Such trespass is widely regarded as morally wrong, regardless of whether it results in damage or harm, because it violates the property right of the owner to control the uses to which the owner's property is put and hence to exclude other people from its use. Similarly, hacking into someone else's network is wrong, regardless of whether it results in damage or harm, because it violates the property right of the owner to exclude other people from the use of the owner's server.

Many hackers, however, reject the analogy with land-based trespass on the ground that some hacking activity can be justified in terms of its social benefits—at least when it results in no damage or harm to innocent persons. Electronic trespass, they point out, contributes to increasing our technological knowledge in a number of ways. First, by gaining insight into the operations of existing networks, hackers develop a base of knowledge that can be used to improve those networks. Second, the very break-ins themselves call attention to security flaws that could be exploited by malicious hackers or, worse, terrorists. Thus, electronic trespass is distinguished, according to proponents, from other forms of trespass in that it is inevitably conducive to public benefit.

Certain hacking activities have also been defended as a form of free expression in two different ways. First, the permissibility of benign break-ins appears to be a consequence of the claim that "information wants to be free." If it is true, as an ethical matter, that all information should be free, then security measures designed to keep hackers out of networks are morally objectionable on the ground that they inhibit the free flow of information. (For a critical discussion of this claim, see Spafford, 1992.) Second, some writers have argued that benign break-ins can be defended as a form of protest or political activism ("hacktivism"). According to this reasoning, such incidents express legitimate outrage over the increasing commercialization of the Web. Politically motivated hacking, according to these writers, should be permitted as long as it results in neither harm nor profit (Manion & Goodrum, 2000).

CONCLUSION

While the advent of the Internet Age has empowered ordinary citizens in novel ways, it has also created a number of equally novel ethical and social problems. Indeed, as is generally true, the very capabilities that increase a person's ability to promote public and private interests also increase a person's ability to harm other people. To the extent that, for example, the Internet enables ordinary citizens to reach a worldwide audience, it dramatically improves their ability to propagate unethical and dangerous ideas.

We should expect that the social and ethical issues of the Internet and e-commerce are no less complicated than the remarkable technology that engenders them. For every argument that invokes traditional ethical and social values in defending a position about online behavior, there is a countervailing argument that resists the application of these values to a technology that seems so radically different from what has preceded it. The debates on these fascinating issues will fill the pages of academic and popular publications for the foreseeable future.

GLOSSARY

Consequentialism The class of ethical theories that determine the moral goodness or badness of an action entirely in terms of its consequences.

Cookies Small files deposited by a Web site on a user's computer for the purpose of enabling the Web site to track the user's preferences.

Copyright Legal device that grants an exclusive right to the holder to reproduce and distribute fixed, original expression.

Cyberterrorism Hacking activity that attempts to harm innocent persons and thereby create a general sense of fear or terror among the population for the purpose of achieving a political agenda.

Deontologism Ethical theories that take the position that the moral goodness or badness of some actions is determined, not by their consequences, but by their intrinsic features.

Domain names Natural language phrases (e.g., http://www.sportinggoods.com) that are associated with Web sites' IP addresses.

Encryption The translation of e-mail messages into code that cannot be deciphered and read by unintended recipients.

File transfer protocol Device that allows a user to transfer files from a user's personal computer to a central network server and conversely.

Filters Programs designed to prevent a user from accessing Web sites with content that is deemed inappropriate.

Hackers Persons who attempt to gain unauthorized entry to network servers. Hacking is usually distinguished from "cracking" in that the latter, unlike hacking activity, is intended to cause harm to innocent persons.

Hacktivism Hacking activity motivated by a desire to express a political view or agenda. For example, a hacktivist might target a corporate Web site as a means of protesting the increasingly commercial character of the Web.

Intellectual property Mental and abstract entities considered as property. Intellectual property includes music, expression of ideas, and designs.

MP3 files Digital files using a format that permits the compression of nearly perfect digital reproductions of sound recordings into small files that can efficiently be transmitted from one user to another.

Patent Legal device that grants inventors monopoly power over their inventions.

Peer-to-peer file sharing A file sharing device that allows Web users to directly access files stored on the computers of other Web users.

Spam Commercial e-mail sent to a user either without the user's consent or against the user's wishes.

Trademark Legal device that grants a firm an exclusive right to use a mark that distinguishes its goods and services from those of other firms.

Utilitarianism A consequentialist moral theory that holds that the goodness or badness of an action is determined entirely by its consequences on well-being, happiness, the number of preferences satisfied, or pleasure in the community.

CROSS REFERENCES

See *Copyright Law; Cybercrime and Cyberfraud; Cyberlaw: The Major Areas, Development, and Provisions; Digital Divide; International Cyberlaw; Internet Censorship; Patent Law; Privacy Law; Taxation Issues; Trademark Law.*

REFERENCES

American Library Association Library Bill of Rights (section V) (1996). Retrieved April 3, 2003, from http://www.ala.org/work/freedom/lbr.html

Barlow, J. P. (1996). *A declaration of the independence of cyberspace*. Retrieved April 3, 2003, from http://www.eff.org/~barlow/Declaration-Final.html

Barlow, J. P. (1993). *A taxonomy of information*. Retrieved April 3, 2003, from http://www.eff.org/~barlow/EconomyOfIdeas.html

Benedek, E., & Brown, C. (1999). No excuses: Televised pornography harms children. *Harvard Review of Psychiatry*, 7(4), 236–240.

Biegel, S. (2001). *Beyond our control? Confronting the limits of our legal system in the age of cyberspace*. Cambridge, MA: The MIT Press.

Canadian Charter of Rights and Freedoms, Fundamental Freedoms (part I, section 2) (1982). Retrieved April 3, 2003, from http://laws.justice.gc.ca/en/charter/

Canadian Human Rights Act (section 13) (1976–77). Retrieved from April 3, 2003, http://laws.justice.gc.ca/en/h-6/29417.html

Communications Decency Act (1996) 47 U.S. Code Section 223. Retrieved April 3, 2003, from http://caselaw.lp.findlaw.com/scripts/ts_search.pl?title=47&sec=223

Draft of the Charter of Fundamental Rights of the European Union (chapter 2, article 11) (2000). Retrieved April 3, 2003, from http://www.europarl.eu.int/charter/pdf/text-en.pdf

The European Parliament (1997). *Resolution on the commission green paper on the protection of minors and human dignity in audiovisual and information services* (COM(96)0483-C4-0621/96). Retrieved April 3, 2003, from http://www.gilc.org/speech/eu/ep-minors-resolution-1097.html

Google (2002). *Features.* Retrieved April 3, 2003, September 9, 2002, from http://www.google.com/help/features.html

Kornblum, J. (1998, August 11). *Compaq buys AltaVista domain.* Retrieved April 3, 2003, from CNET News.com http://news.com.com/2100-1023-214326.html?legacy=cnet

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matter. *The Information Society, 16*(3), 169–186.

Lawrence & Giles (1999). Accessibility and distribution of information on the Web. *Nature 400,* 107–109.

Manion, M., & Goodrum, A. (2000). Terrorism or civil disobedience: Toward a hacktivist ethic. *Computers and Society 30,* 14–19.

Mill, J. S. (1989). *On liberty.* Cambridge, UK: Cambridge Univ. Press.

Nozick, R. (1977). *Anarchy, state, and utopia.* New York: Basic Books.

Rachels, J. (1975). Why privacy is important. *Philosophy and Public Affairs, 4*(3), 323–333.

Rawls, J. (1999). *A theory of justice* (rev. ed.). Cambridge, MA: Harvard Univ. Press, 1999.

Saunders, C. (2002, May 31). EU OKs spam ban, online privacy rules. *Internet News* Retrieved April 3, 2003, from http://www.internetnews.com/IAR/article.php/1154391

Singleton, S. (1998, January 22). *Privacy as censorship: A skeptical view of proposals to regulate privacy in the private sector* (Cato Institute Policy Analysis No. 295). Retrieved from http://www.cato.org/pubs/pas/pa-295es.html

Spafford, E. (1992). Are computer hacker break-ins ethical? *Journal of Systems Software 17,* 41–47.

Spinello, R. (2000). *CyberEthics: Morality and law in cyberspace* (p. 63). Sudbury, MA: Jones & Bartlett.

Suzukamo, L. B. (2002, September 7). Have you Googled yet? *Seattle Times,* C7.

United States Constitution, First Amendment (1789). Retrieved April 3, 2003, from http://caselaw.lp.findlaw.com/data/constitution/amendment01/

Library Management

Clara L. Sitter, *University of Denver*

INTRODUCTION

Library and information science (LIS) leaders have been among the most enthusiastic professionals to embrace the Internet. The development of the World Wide Web in the mid-1990s forced librarians to examine their roles in the information environment.

The Internet is a source of information as well as a platform for delivery. In addition, librarians use the Internet as a tool for the management of both information resources and general library functions.

Internet sources of information provide an obvious complement to library collections. The advantages of Web-accessible materials over print, CD-ROM, or tape versions are many. They can be updated frequently, accessed remotely, and searched electronically.

APPLICATIONS OF THE INTERNET IN THE MANAGEMENT OF LIBRARY RESOURCES

Information resources for library collection development are subject to a cycle of activities. Steps in the continuing cycle include identification, selection, acquisition, organization, access, utilization, and evaluation.

Internet resources being viewed as a source of information implies that they are subject to the same cycle. Internet sites identified as valuable and recommended by librarians are generally selected, linked, promoted to users, and eliminated when no longer relevant. The subject is complicated because Internet access in most libraries is open to *all* sources and the distinction between "recommended" and "available" resources is not always clear to library users.

The management of collections is aided by Internet use. Each activity uses Internet applications. Identification of possible resources, both free and fee-based, is the first step in the collection development process.

Identification

Many types of information are available from the Internet, e.g., indexes, abstracts, full-text journals, and electronic books. In addition there are many types of reference tools including dictionaries, encyclopedias, almanacs, and handbooks. Reference materials are most useful via Internet access when the sources are continually updated. Libraries pay for access to many Web-based reference materials.

Indexes and Abstracts

Online database searching had its beginning in the early 1970s. Dialog offered the first publicly available online research service in 1972. Online databases began with dial-up modems to vendors providing access to commercial (fee-based) databases on a pay-per-search basis. Database vendors offered a variety of resources such as bibliographic indexes, abstracts, and directories. Trained library professionals typically did mediated searches with charges passed on to users. Fees were based on an initial access charge plus per-use/per-minute charges. Web access replaced dial-up service as it became available. Now most libraries pay for fixed-fee subscriptions and permit users to do their own searching at no cost. Database vendors continue to provide an important service allowing libraries to search on a pay-per-search basis for resources they are unable to provide by subscription.

FirstSearch, OCLC's online database package, was first available in 199l and is now used by nearly 20,000 libraries. Many other vendors market individual databases and packages of databases to libraries and consortia.

Full-Text Journals (e-zines, e-journals, webzines)

E-zines (electronic magazines, or e-magazines), e-journals, and Webzines are terms used for electronic publications with no print counterpart. In addition, there are Web versions of print journals.

Access to full-text journals expanded during the 1990s. User demands prompted vendors to expand the number of databases, journal coverage, and retrospective years covered. Libraries developed local area networks on which they loaded CD-ROM or tape resources. Libraries invested large sums of money in technology and infrastructure to support these networks, in addition to the cost of materials. This provided library access, expanded later to institution access, and finally to remote access. Institutions were restricted by the limitations of their technology and the rising cost of materials. The Internet eliminated the problem of technology limitations and shifted the constraints to budget choices.

Full-Text Journal Projects

A number of library-based special projects began in the mid-1990s with the goal of providing access to quality full-text journals available online. Examples include JSTOR and Project MUSE.

JSTOR started with a 1994 Andrew W. Mellon Foundation grant to the University of Michigan to provide access to core scholarly journals in the social sciences and humanities, beginning with the earliest issues in the 1800s. One year later it was established as a nonprofit organization (http://www.jstor.org).

The idea for Project MUSE was born in 1995 at Johns Hopkins University with the Milton S. Eisenhower Library and Johns Hopkins University Press (JHUP) collaboration plan to offer the full text of JHUP scholarly journals via the World Wide Web. In 1999, MUSE published online 46 JHUP titles in the humanities, the social sciences, and mathematics, available for institutional subscriptions either as a package or as individual titles. In 2000 they began adding journals published by other university presses and within two years the project became a collaboration of 29 nonprofit publishers offering more than 220 full-text, peer-reviewed journal titles (http://muse.jhu.edu).

A "linking relationship" between MUSE and JSTOR was announced in 2002. The arrangement provides navigation between the two databases beginning with 18 journal titles held by both. In 2002 MUSE reported a 65% overlap in the two customer bases. The arrangement helps both sets of users by providing easy access to gaps in holdings of each.

Electronic books (E-books)

The Gutenberg Project started computer-accessible titles in 1971 when Michael Hart was given an operator's account with a large amount of computer time by the operators of the Xerox Sigma V mainframe at the Materials Research Lab at the University of Illinois. His idea was to produce public domain editions using replicator technology so that once an item was stored in a computer it would be available to anyone and everyone in the world. The Gutenberg Project (http://www.promo.net/pg/history.html) began its focus with the United States Declaration of Independence and reached 5,000 titles in April 2002. The goal is to offer 10,000 titles by the end of 2003.

Commercial offerings of full-text books were quick to follow the full-text journal explosion. Commercial e-book ventures started in the late 1990s and began marketing individual and packages of contemporary books to libraries. One of the most visible, netLibrary (http://www.netLibrary.com), claiming to be the world's premier provider of electronic books (eBooks), was founded in August, 1998, in Boulder, Colorado. Three years and millions of dollars later, the company had digitized nearly 40,000 titles but was financially unable to continue. OCLC purchased netLibrary in January 2002 and added eBooks to their growing list of services to libraries. MetaText (http://www.MetaText.com), a division of netLibrary, hosts and manages Web-based digital textbooks. XanEdu, a division of ProQuest Information and Learning, purchased MetaText in August 2002. Textbook service will undoubtedly continue to grow as distance learning opportunities expand and face-to-face class technology demands more interactive teaching using electronic tools.

E-books have had mixed success. Complaints include eyestrain, tedious navigation, and reading device performance. Advantages, particularly appropriate for student textbooks, include frequent updating, highlighting, font size manipulation, and searching options. Technology improvements will encourage the acceptance of this format for books. A number of libraries, primarily academic and public, are purchasing and promoting e-books. Some libraries circulate e-book readers, whereas others permit only electronic access. Improved technology may determine the success of the e-book format.

Other Reference Tools

Electronic resources including dictionaries, encyclopedias, directories, handbooks, books of quotations, and statistical resources are available in libraries and homes today. Many library e-resources are available by subscription and may include a print copy. During the 1980s CD-ROMs supplemented the print resources but within 10 years the norm was access to Web resources with the print edition as a bonus. Web sources of library materials are desirable because they provide frequent updates, remote access, and shelving economy. Issues related to the use of electronic Web-based library resources include licensing agreements, nonmediated use of resources, and identification of users.

Selection

The selection process is enhanced by the use of a variety of Web-based resources including Global Books in Print, WorldCat, and publisher Web sites. Bookstore vendor sites such as Borders/Amazon (http://www.amazon.com) and Barnes & Noble (http://www.bn.com) can aid in the selection process. A number of large local bookstores, such as the Tattered Cover in Denver (http://www. tattered-cover.com), have Web sites and are particularly helpful for locating local and regional materials.

Bibliographic Verification

Titles and holdings are verified easily through Internet connections. Subject searches in library catalogs, as well as identifying area holdings of duplicate titles, are accomplished with a single search.

Vendor Services

Online vendor catalogs and services to support collection development officers in libraries make heavy use of the Internet. Essentially all vendors have Web sites with useful information online. Orders are submitted electronically.

Acquisition

Integrated Acquisition Modules

The acquisition of library materials has been streamlined with the use of integrated acquisition modules. Orders may be placed electronically and the library catalog can be updated to show that the item is on order. Electronic MARC records are transmitted to provide full cataloging for the online catalog. On-order status is updated when the item is received.

Online Service Options

License agreements developed with the shift from buying electronic products to paying for access. Increasing percentages of materials budgets are committed to online resources. The licensing of electronic information can be a complex issue involving elements such as user identification and information ownership as well as service and access charges. Libraries and systems are forming subscription groups to get better pricing for electronic products. Statewide funding for public access to general databases is developing. The state of Alaska, for example, provided funding for all citizens of Alaska to access a package of databases beginning in 1998.

Organization

The organization of information, including the creation of catalog records, is an example of one of the earliest applications of electronic resources. Machine-readable cataloging (MARC) was developed in the 1960s. Bibliographic utilities were created in the 1970s to aid librarians in the creation of surrogate records for library materials. The major utilities are

OCLC (Online Computer Library Center, http://www.oclc.org/home) has more than 41,000 members. In 2001 OCLC merged with WLN (Western Library Network), a longtime bibliographic utility serving members primarily in the Pacific Northwest, Alaska, Canada, and Australia.

RLIN (Research Libraries Information Network) serves many special library members. RLG (Research Libraries Group, http://www.rlg.org) is a not-for-profit membership corporation with more than 160 universities, national libraries, archives, historical societies, and other institutional members.

A-G Canada is a computer-based bibliographic network offering its database and services to a variety of Canadian libraries and also to a few libraries in the northeastern United States. A-G Canada was formerly known as Utlas International, and then ISM/LIS (Information Systems Management/Library Information Services).

Bibliographic utilities offer access to standard catalog (MARC) records to ensure bibliographic control. Early access was available through phone lines and modems. Web access has increased the efficiency of the service. OCLC began offering Internet access to records in 1995. OCLC has provided leadership in the establishment of standards for library materials as well as standards for encoding records for electronic resources, archival materials, and other formats.

Most public and school libraries use the Dewey Decimal Classification (DDC), whereas academic (higher education) and special libraries more often use the Library of Congress Classification (LCC). Both systems classify materials by subject. OCLC provides members with classification tables in electronic format with crosswalks relating the two systems.

Cataloging electronic formats prompted discussions by catalogers, resulting in the development of Dublin Core metadata elements. The elements are characteristics of electronic resources including such things as title, creator, subject, date, and language. The Dublin Core Metadata Initiative (DCMI) is an organization dedicated to promoting the adoption of metadata standards and developing specialized vocabularies for describing resources that enable more intelligent information discovery systems (http://dublincore.org).

Access

Union catalogs for library systems, districts, and affiliates have been available for many years. Early formats for union catalogs included computer printouts, ROM readers (looped microfilm), and microfiche. By the 1990s many libraries had migrated to online public access catalogs (OPACs) for patron use. These were first made available by loading magnetic tape updates quarterly or annually onto local area networks. Most library systems now permit daily updates by library staff.

OPACs are the norm for public and university libraries as well as many school and special libraries. State and regional systems often provide links to libraries within the area. An example is ACLIN (Access Colorado Library Network, http://www.aclin.org). Consortium agreements encourage the creation of union catalogs by related institutions.

The integration of library catalogs for geographic, interest, or service areas increases access to information resources. An example, Prospector, is a unified catalog of 16 academic, public, and special libraries in Colorado and Wyoming. The catalog provides records for more than 13 million items that can be accessed by a single search.

Integrated online library systems continue to improve. Library users embrace new features such as added content in library catalogs, linking to online resources, and expanded search capabilities. Libraries continue to customize functionalities in their systems. Trends in automated library systems are reported annually in the April issues of *Library Journal*.

WorldCat, a product of OCLC, is the largest union catalog in the world with nearly 50 million unique MARC records in more than 400 languages representing more than 41,000 members from nearly 100 countries. As

WorldCat continues to grow, the concept of a global library becomes more imaginable.

Reciprocal borrowing has been encouraged by the adoption of one-card systems for states or regions. Geographic lines of service are blurred for the user. Service areas will expand as states continue to develop cooperative agreements.

Utilization

Circulation

Systems managing the circulation of library materials can provide much information for both user and library. Many systems permit users to check their personal records online, renew materials, and reserve books. Self-checkout is also a feature in many automated systems. Almost any library function that is automated can also be accessed through the Internet if libraries choose to make access available. Generally circulation records for individuals are not kept after the materials are returned.

Resource Sharing

Sending resources across system and geographic boundaries increases as access is facilitated. Viewing worldwide records and the expansion of cooperative borrowing create the illusion of "one world, one library"—but not without concerns. Local libraries keep personal information about library borrowers. Most libraries protect patron privacy and resist the temptation to form large patron databases beyond the local library system. The protection of patron records will undoubtedly become an issue as one-card agreements develop.

Interlibrary loan service (ILL), a process for sharing resources by loaning materials library-to-library, has been common practice since the beginning of the 20th century. The first ILL guidelines were established in 1917. In the beginning multiple copies of paper requests were mailed from library to library in search of the needed items. Materials were sent through the mail to the requesting library. Until electronic records were available this process could take weeks.

In the early days a few vendors offered full-text articles for a portion of their indexed resources, resulting in elaborate local area network configurations for periodic updates. In some cases patrons shuffled appropriate discs to access the full text. Subscriptions were expensive. Electronic access to library catalogs and availability of online indexes multiplied ILL requests rapidly. Until databases began offering full text articles by way of the Internet the demand on interlibrary loan offices was immense. Even with patron access to many full-text articles, support for ILL service is a major financial commitment for most libraries.

Document delivery is a term implying the purchase of information sources from commercial suppliers when ILL is not available or when faster service is required. An early example of a document delivery service, Uncover, developed by CARL (originally Colorado Alliance of Research Libraries) scanned journal articles and then made them available by faxing "on demand" for a price. Ingenta purchased Uncover in 2000 and continues to provide service to libraries.

Programming

Library programming has changed since the Internet has been available. The ability to create Web sites, post documents, promote programs, and send distributed messages expands the ability of libraries to communicate with users as well as to learn what similar institutions do for programming. Plans for the promotion of books and reading are generously shared. Successful ideas are easily replicated for use in other systems. Specialized discussion lists and Web sites facilitate the recycling of successful programs for library users of all ages. This is a popular practice for youth summer reading programs.

Electronic distribution lists impact professional communication of ideas related to programming, problem solving, and other work-related concerns. Discussion lists for librarians number in the hundreds. Each association division, function area, and special interest group has at least one discussion list. The American Library Association alone offers nearly 200 lists (http://lp-web.ala.org:8000/guest/main). Most librarians subscribe to multiple lists.

Service

Virtual reference is one of the hot topics in continuing education and professional literature. The interest will likely continue as patron demands become stronger. This is an area for possible outsourcing or collaboration. Asynchronous e-mail reference has been a service of libraries for a number of years, but the delay in response is unsatisfactory for many users. Ask Jeeves and Refdesk.com, free round-the-clock resources, prompted librarians to think about offering virtual 24/7 reference service. Early projects included California's Metropolitan Cooperative Library Service (July 2000) and the Library of Congress Collaborative Digital References Services (CDRS) following a few months later.

In June 2002, the Library of Congress and OCLC launched Question Point (http://www.questionpoint.org/), a collaborative reference service developed with input from participating members of the Global Reference Network, a group of libraries and institutions worldwide that are committed to digital reference. Virtual reference is being developed in a number of different configurations by all types of libraries.

Instruction

In-house tutorials as well as links to outside instructional modules are common on many library Web sites. Libraries of all types offer instruction through workshops, classes, point-of-use tutorials, handouts, help screens, and one-on-one contacts. The Internet is often both the subject of instruction as well as the tool for teaching. Helping library users become wise consumers of information is an important responsibility for today's information professional.

Colleges and universities have a strong focus on instruction. The Association of College and Research Libraries (ACRL) has established a number of standards and guidelines including "Information Literacy Competency Standards for Higher Education" (http://www.ala.org/acrl/ilcomstan.html). Academic librarians traditionally called their instructional service "bibliographic instruction" or simply BI. The term lingers although the

ALA Library Instruction Roundtable (LIRT) changed the terminology to simply "instruction" in the 1990s.

Public libraries spend an increasing amount of energy instructing users in Internet and database searching. Literacy is a strong focus for many public libraries. The Public Library Association (PLA) projects reflect the priorities for many public libraries (http://www.ala.org/pla).

School libraries continue a strong focus on "literacy" skills. The American Association of School Librarians (AASL) continues to take a leadership role in providing resources for school library media specialists (http://www.ala.org/aasl/index.html). Many K-12 libraries offer Internet accessible tutorials for students. External links to well-done instructional packages are found on library Web pages. Librarianship is a generous profession when it comes to sharing expertise as well as resources.

Evaluation

Statistics on the circulation of specific titles as well as subject areas in library collections can be valuable when making collection development decisions. Library automated systems can provide access to much more circulation information than most libraries utilize. In general, libraries make a practice of using statistics rather than specific user information in collection development. The extent to which libraries use circulation information is a policy decision made by individual libraries. Much valuable information can be derived from the user database and the profile of library users. Most libraries have resisted tying circulation records to specific patron records to protect the privacy of the users. Sending e-mail to users alerting them of books they might want to check out, the amazon.com model, might be viewed as a service. But doing so at the cost of invasion of privacy is a serious tradeoff.

Collection development in libraries has improved because of electronic records. The Western Library Network (WLN), together with Pacific Northwest member libraries, developed a "conspectus" approach to collection analysis beginning in the 1980s. As electronic manipulation became possible WLN could run an analysis of age, depth, and breath of individual collections. In addition, collections could be compared with core title lists or holdings in other institutions. Paper reports gave libraries documents that identified strengths and weaknesses in their collections. Although the initial analysis was not Internet-based, follow-up collection development Internet activities have developed. Examples include locating out-of-print resources, comparing similar collections, and accessing electronic records. WLN merged with OCLC in January 2001. Collection development options will continue to expand and will be available through Internet access.

The Internet allows anyone and everyone to become an author and a publisher. Quality and quantity have no restrictions. The evaluation of Internet materials and the filtering of public and school library Internet workstations continue to be a public concern. The Children's Internet Protection Act (CIPA) is an example of the strong interest of the public in Internet materials. Background on CIPA and the position of the American Library Association (ALA) is available at http://www.ala.org/cipa

APPLICATIONS OF THE INTERNET IN THE MANAGEMENT OF LIBRARY FUNCTIONS

In addition to the unique applications of the Internet to library resources and user services, the Internet enhances traditional management functions such as planning, organizing, staffing, directing, controlling, reporting, and budgeting. Enhanced communication and user access to information are most important applications of the Internet in the management of library functions.

The applications of the Internet in library management may not seem different from applications in other management units but the attitude and approach may be distinctive. Librarians, and information providers in general, are extremely cooperative and collaborative professionals. Electronic communication was accepted early in its development as a way to collaborate beyond library walls, system networks, and political borders.

Planning

The planning aspect of library management is a process leading to the development of statements or documents addressing library philosophy and values, vision, mission, goals, objectives, actions, activities, policies, and procedures. This process is not unlike planning processes in many institutions, although it may be more complex due to the range of user characteristics, the variety of information packages, and the differences in information sources.

The planning process generally includes an environmental scan of internal and external factors. Internal factors may include personnel, services, and resources such as facilities, collections, budgets, and technology. External factors include influences such as political, economic, social, and technological issues (PEST). An analysis of strengths, weaknesses, opportunities, and threats (SWOT), along with internal and external factors, is useful in the planning process. Data to support both PEST and SWOT analyses are available from electronic sources on the Web.

Internet resources facilitate library planning in a number of ways including easy access to information. Access to U.S. Census records, as well as state and local community data, provides invaluable information about the user and nonuser bases of libraries serving the public. Online U.S. Government documents describing K-12 and higher education (e.g., the National Center for Educational Statistics, http://www.nces.ed.gov) provide resources for school and academic libraries. Special libraries, profit and nonprofit, access various elements of community analysis information as well.

Vision, Mission, Goals, and Objectives

Libraries began creating Web pages early in the development of the Web. Vision and mission statements are regularly posted on Web sites. Preparation of these documents for Web access prompted renewed interest in using them as public relations tools.

Strategic and Long-Range Planning

The Internet is used for e-mail discussions, transfer of draft statements, and collaborative creation of strategic and long-range planning documents. Library Web sites provide a means for public access to these and other institutional documents.

Policy-Making

Decision makers within library systems sometimes make policies based on norms in like institutions. Internet access to examples of policies provides a foundation and argument for adopting particular positions as well as providing a thorough analysis of users.

Forecasting and Decision-Making

Access to information is critical for forecasting and decision-making for library administrators. Library and Internet resources provide access to private and public information on all topics relevant to decision-making.

Organizing

Organizing as a function of library management is generally an internal activity. The Internet provides a tool for communicating the organizational structure when libraries choose to post their organizational charts or structures on their Web sites.

Organizational Structure

Internet technology has increased options for libraries in areas including resources, communication, and services. The increased complexities of library management, along with other external factors, have resulted in a trend toward nontraditional organization. Hierarchical structure and organization by function such as public services and technical services is no longer used in many libraries or may be supplemented by cross-functional teams. Flat organizational charts, participatory management, and collaborative groups may be displayed graphically as round, spoke, star, or other creative illustrations of the relationships. Electronic communication helps the structures work.

Staffing

Library management staffing issues include recruiting, training, retaining, evaluation, policies and procedures, and legal issues. Access to information and resources via the Internet has resulted in a greater awareness of job opportunities. Librarians in general have strong professional networks for service that may also be used for recruiting staff. Librarians conscious of professional development are likely to change jobs and institutions multiple times throughout their careers.

Recruiting

A broader search for employees has been possible since the Internet became available for free or inexpensive job postings. Links from professional associations or state libraries to joblines in state and regional affiliates provide easy access for job seekers. Jobs posted on library Web sites provide easy communication as well. Both employees and employers can compare salaries, job descriptions, and qualifications.

Still and moving pictures provide additional information about job settings for remote applicants. Web-based portfolios prepared by applicants will likely become standard for job seekers. Video conferencing, Web cams, and Web sites facilitate a broader recruitment base.

Training

Currently local training is still predominately face-to-face but prepared modules available on the Internet are growing. Generic training and continuing education opportunities through Internet access are regularly offered. Specific training, e.g., from OCLC, is offered via webcast presentations viewed from workstations in employee offices. Employee support for continuing education, training opportunities, and job enrichment are motivating factors for professional and support staff. Identification of appropriate sources, access to good technology, and budget allocations for time and registrations are important success factors. Associations, graduate schools, and interest groups sponsor downlinks for special continuing education opportunities.

Retention, Evaluation, and Policies and Procedures

The Internet applications relative to retention, evaluation, and policies and procedures are information-based and communication-oriented. Employee access to online evaluation forms, policies, and procedures helps keep staff informed. Updates are easily made. Public access to personnel documents gives potential employees insight into the institution. E-mail distributed to employees provides instant communication for updates and changes.

Legal Issues

Access to discussions of legal issues as well as to the laws themselves is an obvious application of the Internet. Web access to proprietary legal databases is a help to administrators as well as to library users.

Directing

Motivation

Access to Internet resources may indirectly provide motivation within an organization. Simply keeping up with technology, resources, and communication available via the Internet has been an incentive to staff to update technology skills, learn new resources, and manage volumes of e-mail messages.

Leadership

Strong leaders are needed in the library information science field. Because leaders in the profession are aging, rapid advancement will be possible for the new generation of information professionals. Remote mentors, online training, and continuing education modules will support developing leaders.

Communication

The Internet makes communication easy by using distributed e-mail, discussion lists, and electronic access to documents. The problems with electronic communication result from the absence of face-to-face communication including facial expressions, eye contact, voice tone, and other body language, which may result in occasional misunderstandings. Smaller PDAs (personal digital assistants), wrist-size computers, and wireless technology will facility Internet connections.

Controlling

Standards

Many standards have been developed for library services and collections. Competency standards have been developed for librarians. Examples include the Special Libraries Association (SLA), http://www.sla.org/content/SLA/professional/meaning/competency.cfm; the American Library Association (ALA), http://www.ala.org/work/standards.html; the Medical Library Association (MLA), http://www.mlanet.org/; and the American Society for Information Science and Technology (ASIST), http://www.asis.org/AboutASIS/professional-guidelines.html

Reporting and Budgeting

The availability of the Internet as a tool for communication to the service base, the public in many cases, has resulted in a wider distribution of the business of libraries. As libraries spend larger proportions of materials budgets on access it will be important to communicate other library values to decision makers.

AREAS OF FOCUS FOR INFORMATION PROFESSIONALS IN LIBRARY MANAGEMENT

The opportunities and challenges for information professionals are enormous. Many of the Internet-related challenges facing information professionals relate to collections and users. In addition there is concern for the future of the library information science profession itself.

Collections

Fee-Based Resources

Library users do not always recognize the differences between free and fee-based library materials. The Internet as a delivery vehicle for purchased resources makes the access seamless to many library users. As consortia expand, access to fee-based databases will be even smoother. Communication with users will be more important to ensure continued funding.

Budgets

Increasingly larger portions of library budgets will be invested in "access" rather than ownership of materials. The public may begin to question the value of walk-in libraries. Libraries must continue to examine their roles in the lives of information seekers as well as recreational readers.

Digital Libraries

By definition, digital libraries offer materials in digital format. An example, Jones *e*-global Library, a subsidiary of Jones Knowledge, was one of the earliest to be developed. Jones University was the first higher education institution accredited on the basis of a digital library. Jones *e*-global Library was designed to support the students in distance programs of Jones University but it expanded to provide electronic resources as a supplement to all types of libraries and users. When web-based resources became widely available to libraries the need for the packaged services of digital libraries declined. Commercial digital libraries had a short life.

The creation of a digital library across institutions presents one of the greatest challenges for information professionals. The Digital Library Federation (http://www.clir.org/diglib/about/strategic.htm) and the Library of Congress (http://www.loc.gov), along with library leaders from all types of libraries, may collaborate to develop standards, create tools, identify content, and negotiate licenses.

Copyright

Digital copyright issues will continue to be a concern. See information on the Digital Millennium Copyright Act at http://www.ala.org/washoff/dmca.html. Librarians are champions of copyright protection for authors and creators of information resources. At the same time, they must balance creator rights with public access to information by opposing per-use charges for electronic information (http://copyright.ala.org/internet.html).

Preservation

Technology will continue to be a challenge, particularly for the preservation of digital records, as advances are made and migration is difficult. Digitization projects continue to expand to include more archival and museum resources, offering opportunities for stronger collaboration with these related institutions.

Users

Access

Issues of easy access to electronic information, equal access for all citizens, and preservation of intellectual freedom all relate to the defense of a democratic society. ALA's Office of Intellectual Freedom (http://www.ala.org/alaorg/oif "Access to Electronic Information, Services, and Networks: An Interpretation of the Library Bill of Rights") provides guidance for professionals (http://www.ala.org/alaorg/oif/electacc.html). The overload of Internet information, the gap between the "haves" and "have-nots," and the demand for Internet filtering will continue to be challenges and political issues for information professionals.

Information Literacy

Teaching information users of all ages to be smart information consumers will become even more important as libraries become more virtual and patrons are more independent searchers. Information professionals in all types of libraries share the responsibility for instruction.

Privacy

As technology becomes even more sophisticated, librarians will have to weigh the advantages of knowing user preferences with the trade-off of invading patron privacy. Librarians may find themselves among the few to defend privacy as the public rushes to give up rights in the name of national security. "Privacy; an Interpretation of the Library Bill of Rights" (http://www.ala.org/alaorg/oif/privacyinterpretation.html) is a resource for information professionals involved in developing policies related to user privacy.

Following 9/ll Congress passed legislation to expand the powers of law enforcement for the purpose of security. See http://www.ala.org/washoff/patriot.html for information on USA PATRIOT Act (Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism). These actions raise important issues of privacy.

Distance Delivery and Disintermediation

As students and workers telecommute, the role of all types of libraries becomes more focused on access than on users. The challenges of licensing electronic resources, providing virtual reference, and keeping up with technology may be more easily addressed than the lack of human contact. The increased distance and reduction of face-to-face time between librarians and users will be an adjustment and a loss for both. Librarians will be challenged to find ways to continue the connection with patrons.

The TEACH Act (Technology, Education and Copyright Harmonization Act) of 2002 clarified the terms and conditions for higher education institutions to use copyrighted materials in distance education.

Technology

Librarians will continue to embrace new technologies as they become available. It has become cost-effective for libraries to cooperate and collaborate in technology issues including resource management, continuing education, policy-making, legislative positioning, and public relations. PDAs (personal digital assistants) with Internet and cell phone capabilities, wrist-size computers, wireless networks, and countless other devices are finding library applications. [NOTE: The font size for this paragraph is 10 point!]

Providers

LIS Professionals

The need for library information science professionals is great. The professional work force is graying. For the past few years there has been a shortage of school library media specialists, youth services librarians, and catalogers. Now there is a growing need for reference librarians and technology experts.

Professional Certification

Librarians have always been committed to lifelong learning. Plans have been proposed for public library administrators to receive additional training and education courses beyond the MLS or MLIS to be certified public library administrators (CPLA); see http://www.pla.org/projects/certification/certification.html. The Association of Specialized and Cooperative Library Agencies (ASCLA), the Public Library Association (PLA), and the Library Administration and Management Association (LAMA) established standards and processes with formal announcement in 2002. The proposed areas of competencies include nine core areas: (1) budgeting and finance, (2) fundraising, (3) library building, planning, and maintenance, (4) organization and personnel management, (5) technology, (6) building alliances and networking, (7) strategic planning and marketing, (8) serving diverse populations, and (9) current Issues in public library management. A general foundation in library and information science is necessary for all information professionals, but specialized knowledge is needed in many areas in order to do the job well. Library administration is the first area identified. Much of the training will be available on the Internet through online classes and workshops.

Education

The Internet has impacted library education. Curricula are more technology oriented, courses may be all or partially online, and schedules may be nontraditional. The master's degree is generally regarded as the first and terminal degree for the practicing librarian.

The Association for Library and Information Science Education (ALISE, http://www.alise.org), with the W.K. Kellogg Foundation, conducted a study of education in library and information science resulting in the KALIPER (Kellogg-ALISE Information Professionals and Education Renewal project), published in 2000 (http://www.alise.org/publications/kaliper.pdf). The report identified six trends including investments in information technology.

National meetings, the Congress on Professional Education (COPE), in 1999, 2000, and 2003 addressed professional education (http://www.ala.org/congress). Accreditation of master's programs, now granted by the American Library Association, is moving toward an independent agency. Practitioners and educators continue to discuss core competencies.

Library professionals in general enthusiastically defend core professional values. Examples of the values of information professionals can been seen in the Code of Ethics of the American Library Association (http://www.ala.org/alaorg/oif/ethics.html), as well as other associations of information professionals. The important values include resource sharing, equity of access, intellectual freedom, literacy and learning, copyright protection, patron privacy, information preservation, and respect for diversity. These values are observed in the interaction of librarians and users but also carry over into the general management functions of administrators. Regardless of the Internet revolution a commitment to professional values will remain firm.

GLOSSARY

Access vs. ownership Paying to use materials rather than buying them. This concept is a major shift for library collection development.

Bibliographic control The process of using standard descriptive and identifying characteristics to create a surrogate record for each information item.

Bibliographic utility A company offering cooperative cataloging support for members. Members contribute new records and download existing ones, resulting in a large union catalog.

Controlled vocabulary The preferred subject term for retrieval of information in electronic databases. Comparable to subject headings for library catalogs.

Crosswalks A program or instrument to show equivalent values in multiple schemes such as classification [Dewey Decimal (DDC) to Library of Congress (LC)] or standard cataloging records [MARC to Dublin Core].

Digital library An organization that provides the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. [Paraphrased from the Digital Library Federation's (DLF) working definition, http://www.clir.org/diglib/about/strategic.htm]

Dublin Core (DC) Dublin Metadata Core Element Set. A set of standard metadata elements describing characteristics of an electronic information package. The work began at a workshop in Dublin, Ohio, in 1995. See http://dublincore.org/

E-reference materials Electronic resources including indexes and abstracts, bibliographies, reference tools, and full text books.

Information package A broad term for recorded information in any form such as print, graphic, electronic, video, or music.

MARC (MAchine Readable Cataloging, http://lcweb.loc.gov/marc/umb) Standard descriptive data for catalog records for library materials encoded with machine-readable format.

Metadata Data about data; descriptive elements of an information package encoded with a standard format such as MARC, Dublin Core, or GILS (Government Information Locator Service) to create machine-readable records. Essentially they are a set of elements for describing electronic resources for electronic library records.

OCLC (Online Computer Library Center) The largest bibliographic utility in the world (http://www.oclc.org/home). In addition to cooperative cataloging other services include support for acquisitions, serials, interlibrary loan, online databases, and e-books. OCLC began as the Ohio College Library Center in 1967 and changed its name in 1981 to reflect wider membership.

Relational database A database with information from searchable fields of records stored in separate files (e.g., name, title, subject) to facilitate searching.

Surrogate record A record representing the characteristics (descriptive and identifying elements) of an information package.

Union catalog A combined catalog with records of more than one collection.

Virtual library Digital formats make up the collections of a virtual library. The terms digital library and virtual library are often used interchangeably, though technically they are not the same.

Virtual reference service Reference service provided to remote users by using real-time virtual reference systems software.

WorldCat The combined catalog of OCLC member libraries. WorldCat is the largest union catalog in the world with nearly 50 million unique MARC records in more than 400 languages representing more than 41,000 members from nearly 100 countries.

CROSS REFERENCES

See *Databases on the Web; Digital Libraries; Research on the Internet.*

FURTHER READING

Evans, G. E. (2000). *Developing library and information center collections* (4th ed.). Englewood, CO: Libraries Unlimited. Retrieved 2000 from http://www.lu.com/getpage.cfm?file = textbook2.html & userid = 77731466. [Supplement]

Evans, G. E., Ward, P. W., and Rugaas, B. (2000). *Management basics for information professionals*. New York: Neal–Schuman.

Libraries and the Internet (2001) *CQ Researcher, 11,* 465–488.

Liu, L.-G. (1996). *The Internet and library and information services; Issues, trends, and annotated bibliography, 1994–1995.* New York: Greenwood Press.

Lynch, C. (2000). From automation to transformation: Forty years of libraries and information technology in higher education. *EDUCAUSE Review, 35,* 60–68.

Reitz, J. M. (2003). *Dictionary of library and information science.* Westport, CT: Libraries Unlimited.

Reitz, J. M. (2002) *ODLIS: Online dictionary of library and information science*. Retrieved February 25, 2003, from http://vax.wcsu.edu/library/odlis.html#B

Sherman, C., and Price, G. (2001). *The invisible web; Uncovering information sources search engines can't see*. Medford, NJ: Information Today. [See also http://www.invisible-web.net]

Shuman, B. A. (2001). *Issues for libraries and information science in the Internet age*. Englewood, CO: Libraries Unlimited.

Smith, K. W. (ed.) (1998). *OCLC 1967–1997: Thirty years of furthering access to the world's information*. New York: Haworth Press. [Available at http://www.oclc.org]

Stueart, R. D., and Moran, B. B. (2002). *Library and information center management* (6th ed.). Englewood, CO: Libraries Unlimited. [Support site at http://www.lu.com/management]

Taylor, A. G. (1999). *The organization of information*. Englewood, CO: Libraries Unlimited. [Supplement available at http: // www.pitt.edu / ~agtaylor / courses/glossup.html]

Tenopir, C., and Ennis, L. (2002) A decade of digital reference, 1991–2001. *Reference & User Services Quarterly. 41,* 264–73.

Linux Operating System

Charles Abzug, *James Madison University*

OVERVIEW

Linux is probably the best all-around operating system in use in the world today. Linux has a number of competitors, most of which are commercial products, like Microsoft Windows in its various manifestations, IBM's OS/2, Sun Microsystem's Solaris, SGI's IRIX, and Novell's NetWare. In contrast to these other operating systems, the development of Linux is carried out not by paid employees operating in an industrial environment, but by a small army of dedicated and enthusiastic volunteers, many working evenings and weekends during their spare time. Both the mode and the culture within which Linux was nurtured are responsible in large measure for the superiority both in usability and in performance attained by this operating system. In order to explain this attainment, it is necessary first to describe what is an operating system and what is the role of the operating system in the functioning of a computer. In addition, it is necessary to convey the uniqueness of Linux in terms of both how it was originally developed and how it continues to evolve.

What Is an Operating System?

The operating system is the most important piece of software in a computer. There are two principal points of view for looking at an operating system, that of the computer system and that of the user and programmer.

The View of the Computer System

From the point of view of the computer system, the operating system is the software responsible for mediating the control of all three families of the computer's resources: hardware, software, and data. The hardware resources controlled by the operating system include one or more central processing units or CPUs, the main memory (often referred to as random access memory or RAM), the keyboard and display, and also any devices that may be present. These devices include communications facilities, such as a network card or modem, as well as all storage devices, including magnetic disks ("hard" disks, floppy disks, and various proprietary removable disk cartridges, such as Zip, Jaz, and Bernoulli), CDs or DVDs, and tape drives. Software resources of the computer include, in addition to the operating system itself, a variety of utilities that may be bundled with the operating system by the vendor, as well as any separate commercially purchased or custom-programmed software applications. Applications software, at a minimum, on a computer intended for use as a personal workstation usually includes a word processor, e-mail handler, spreadsheet, Web browser, and database manager. Applications software on a server may include a robust database manager specifically designed for simultaneous access by multiple users, or some other heavy-duty application. In a corporate environment, computers are typically purchased *en masse,* and both the hardware configuration and the installed software are usually identical or nearly identical for a group of desktop systems purchased together as a single purchasing action by a commercial or governmental organization. The data resident on each system, however, are what distinguish that system from all others that may be otherwise identical. The data make each computer different from every other system in the organization. Thus, in summary, the operating system is the highly complex software that controls all three types of resources of the computer: hardware, software, and data.

The operating system is much more, however, than merely the manager of the resources of the computer system. In addition to mediating the control of the disparate resources of the system, the operating system also provides the interface with which the user interacts with the computer, both to pass on the expressed wishes of the user to the computer and to communicate to the user information about what the system is doing, as well as reports regarding the status of requests made by the user, and warnings and messages regarding errors and malfunctions that the user needs to know about.

The View of the User and of the Programmer

So far, we have considered the operating system principally from the point of view of its role in controlling the computer. The other point of view is that of the user or of the programmer. There are several services provided both to the user and to the programmer by the operating system. These include

execution of programs at the user's request;

performance on behalf of the program of input/output operations, thus saving the programmer from having to program such operations in detail;

interaction with devices, thereby relieving the programmer of having to program in detail in accordance with the needs of the individual device;

reading and writing of files of various types;

handling of communication from the user to the machine, typically via either a command-line interface (CLI) or a graphical user interface (GUI);

communication between processes;

detection and reporting to the user of errors occurring in system operation;

communication to the user of status reports, including report of the successful completion of tasks requested by the user;

a set of rules for writing function or subroutine calls that can be invoked from an application program and that are carried out on its behalf by the operating system (application programming interface, or API); and

provision of an environment conducive to the development by programmers of applications to carry out useful work on behalf of the user, as well as to the testing, debugging, and maintenance of application programs.

Design Goals for an Operating System

There are four principal goals that the developer must meet in the design of the operating system. First among these is efficiency of use of the resources of the system. The resources of the computer system are expensive, and the user therefore is eager to see the resources put to use efficiently, as well as to get his work accomplished in a timely manner. Second in importance is ease of use of the system. Users come in a great range of levels of sophistication of computer skills. The entire spectrum of users, from the almost totally computer-illiterate to the most highly skilled computer professional, should all be able to exercise an adequate level of control over the operation of the system so as to obtain from it a high level of service. A third goal is that the resources of the computer system be adequately apportioned to the various tasks to be accomplished so that each task progresses at a speed and efficiency appropriate to its importance relative to other tasks. In particular, if two or more users share the machine, then each user must be allocated an appropriate share of each of the various resources of the system. Finally, the security needs of the system must be adequately addressed. This last concern is particularly important for a system whose use is shared by multiple users.

Operating Environments

There are several different environments in which computer systems function. In extreme circumstances, a particular environment may require the services of a special-purpose operating system whose design is optimized to serve the special environmental needs of that system particularly well. In most cases, however, a general-purpose operating system can be developed that meets the needs of a broad range of uses, and also which includes sufficient flexibility to enable the system manager to configure the system on installation to meet the operational needs anticipated. A good general-purpose operating system will also provide a range of facilities and services that can be adapted and tuned as necessary so as to enable the system to continue to meet the needs of its operational environment even as these needs change over time.

Broadly speaking, the operational environment of a computer system falls into one of three categories. The simplest environment is that of the single-user-at-a-time. Most desktop computer systems, including those of higher-performance capability also known as workstations, fall into this category. The old DOS that came with the original IBM PC (IBM DOS) as well as with most PC clones produced by other manufacturers (MS-DOS) was a single-user operating system. DOS was not only a single-user but also a single-tasking operating system. That is, the user could have the computer do only one thing at a time, and had to wait until the current task was completed before assigning a new task. Single-user operating systems with more sophistication were developed, in which the user is able to initiate several tasks that run concurrently, thus allowing a more efficient use of computer resources. When a disk-read or disk-write operation is required, for example, it is usually necessary for the read/write head of the disk drive to travel to some location overlying one particular circular track selected from the many that exist on each of the disk platters. This operation, known as a disk head seek, is incredibly slow by computer standards, occupying an amount of time in which the CPU, were it free and unencumbered, could perform thousands or even millions of operations. Thus, while the disk drive is executing its head seek, the CPU could be doing something else that is not dependent upon the completion of the disk-read or -write operation. In an old-fashioned single-tasking system, this does not happen; instead, the CPU must wait for the disk-read or the -write operation to complete before resuming operation. In the large-scale computer (minicomputer or mainframe), multitasking operating systems were developed as early as the 1960s. At the smaller scale of the personal computer, several operating systems that enabled the single-user microcomputer also to do multitasking were developed in the 1980s. These included Mac OS for the MacIntosh, several variants of Microsoft Windows, and IBM's OS/2. Several UNIX variants were developed for the PC as well, which were even more capable.

Next in complexity, after the single-user multitasking environment, is that characterized by multiple simultaneous users. This environment is characteristic of most minicomputers and mainframes, as well as of servers. Operating systems with facilities that enable their use in a server or multiuser environment include the server versions of Windows NT, Windows 2000, and Windows XP, as well as Novell NetWare, the server version of IBM's OS/2, DEC's VMS, various mainframe operating systems, and a host of proprietary variants of the UNIX operating system, including IBM's AIX, Hewlett–Packard's HP-UX, Silicon Graphics' IRIX, and HP–Compaq/DEC's ULTRIX.

Historically speaking, one of the earliest capabilities to be developed in operating systems was the ability to handle batch jobs. This development took place prior to the development of multiuser, multitasking operating systems. In a batch environment, one or more tasks are submitted to be run on the system. Each of these runs to completion, and the results are collated and left for later, offline examination by the user. Modern systems with multiuser, multitasking capability may also be provided with the capability to run batch jobs. Usually, the tasks submitted as batch jobs have no need for ongoing monitoring or supervision by the user. Examples of such jobs are regular, complete system backups and periodic scanning of the entire system for computer viruses. Such a job can either be scheduled to run at a time when the user is not present, so as not to compete for system resources with tasks of greater urgency, and thus interfere with the performance of those tasks for which the user is anxiously awaiting results. Alternatively, the batch job can be run in background at a reduced priority level, so that it runs only when a task to which the user is attending will not be slowed or delayed. Operating systems in the more robust category often have the capability to run batch jobs in addition to handling multiple tasks and multiple users.

Finally, there is the real-time environment, where the computer is used to control something whose on-time performance is critical. Examples of real-time environments include the flight-control and navigational computers of aircraft and spacecraft; various military environments, such as the command and control computers used in single-vessel and integrated fleet defensive systems, like the U.S. Aegis system; aircraft and antimissile weapons-systems control computers, like the U.S. Patriot antiaircraft and antiballistic-missile system; systems used in medical diagnosis and treatment; systems used in factory or shop floor automation, including robotics, large-scale distributed computer systems used to control transnational telecommunications systems; and the computers used to control nuclear power plants. In several of these cases, a general-purpose operating system can be used that provides, besides the multiuser and multitasking features found in many modern operating systems, additional facilities suitable to the more exacting needs of a real-time environment. However, in those instances where the real-time needs are particularly severe, only a special-purpose operating system will do. In such extremely constrained circumstances, it is usually cost-effective to have a dedicated, special-purpose computer that does nothing else but see to the needs of the so-called hard real-time environment.

What Linux Is All About

Linux is a UNIX-like group of operating systems, which was developed as a cooperative effort by a loosely knit team of capable and enthusiastic programmers who volunteered their services. The programmers on the original development team were associated with the Free Software Foundation's GNU project. The activities of programmers working on the GNU project were augmented by Linus (pronunced "Lee'-nuhs") Torvalds and a small number of associates working with him. This latter group contributed the operating system's kernel. The name Linux, which properly applies to the kernel alone, is derived from Linus Torvalds' first name. The Linux kernel was originally targeted at the Intel ×86 family of processors. However, it now comes in several varieties, each of which runs on a particular hardware platform. The various incarnations of the kernel intended for the different hardware platforms are all integrated with the pre-existing GNU software to form a complete operating system. Quite a number of hardware platforms are currently supported. The various members of the Linux family are all multiuser, multiple-concurrent-process operating systems featuring time-sharing, but also supporting batch processing in background. They differ from each other only in whatever way is necessary to support the particular hardware architectures on which they run. The look-and-feel of all these systems is very similar.

Linux is open-source software. The source code is bundled together with the executable code and the documentation, and all can be had without charge. Should the user require the code and manuals on CDs, then these are available for purchase at a very modest charge to cover the media, duplication, packaging, and distribution costs only. The software is universally usable by anyone for any purpose with absolutely no licensing fees.

The Linux user is free to modify any or all of the source code, and then to recompile the modified code, to meet the needs of his system. If the modification is potentially of use to other members of the Linux community, then he is encouraged to share with the community the changes made. Furthermore, anyone at all who enhances the software is welcome to distribute the enhanced version, provided that the provisions of the GNU General Public License are upheld, preserved, and transmitted onward together with the new code and documentation, thus preserving the users' freedom to read, study, modify, and enhance the software.

Why Linux Is Important, and Why It Can Be Useful to You

The importance of Linux is due to a number of factors. Some of these are technical and some nontechnical, but particularly related to the Linux culture.

Performance and efficiency are superb. In particular, users who find themselves often frustrated by the notorious hourglass encountered in the various Windows environments find themselves warming up to Linux as they encounter a snappier and more responsive system.

Reliability is certainly among the best in the industry. Linux systems are extraordinarily stable, with several server systems known to have been running continuously for several years without ever having to reboot. The organization "The Linux Counter" (http://counter.li.org) encourages Linux users to register as well as to provide information voluntarily regarding their installation of Linux, and especially its performance. The data displayed recently show that there is one machine that has been running Linux continuously for 1000.3 days (i.e., nearly three years). This machine happens to be not an Intel architecture machine, but a Hewlett–Packard/Compaq/DEC Alpha. However, there is a list of the top ten machines reported to be up continuously on Linux, and the other nine, all of which consist of Intel hardware (one 80486 and the rest in the Pentium family) have all been running for more than 500 days. In fact, the average running time reported for Linux is an astounding 37.7 days. The contrast between the ultra-high reliability of Linux and the frequently appearing "blue screen of death" of the Windows world is particularly striking.

Linux complies with the Portable Operating System Interface Standard (POSIX). More than 95% of the code is written in the programming language C, and thus it can be ported relatively readily to any new hardware architecture that might be developed.

Linux already runs on a huge variety of hardware, ranging from desktop systems to mainframes (details presented below). Over the entire range of hardware supported, there is a consistency both of user interface and of programmer interface. Thus, the user does not need to learn a new operating system when switching to a new hardware platform.

There is a standardized structure to the Linux file system, which results in a uniform location of critical files over both different hardware platforms and different Linux distributions. The Linux community recognized relatively early on that the originally different file structure used by the various vendors in their Linux distributions was chaotic, and so a Filesystem Standard (FSSTND) was developed to bring the situation under control. Subsequently, a superior Filesystem Hierarchy Standard (FHS) replaced FSSTND.

Development and evolution is vastly different for Linux from what it is for proprietary operating systems. Since they are volunteers, Linux developers do not have to answer to either a marketing department or a profit-oriented corporate bureaucracy. A proposed change or a reported software error is posted by the interested party on the Internet, where it is likely to be seen very quickly by any number of Linux enthusiasts. If one of them deems it to be worthwhile, then either the change is implemented or the bug gets fixed, whichever is applicable, and detailed information on the change is then also posted. There is a spirit and a culture of cooperation, and therefore anyone who programs an improvement in the code, even though primarily intended for his/her own private use, will usually also post it on the Internet so that other users may also benefit from it if they wish. All of the improvements are then collated for integration into the next release of the kernel software. This is in marked contrast to the philosophy encountered in the commercial software-development world, where the attitude taken towards the users is typically, as described by Stallman (1999), "If you want any changes, beg us to make them." In addition, in the commercial environment users must often pay for upgrades with no assurance that the bugs that they have reported have already been fixed rather than still sitting in the queue waiting to be worked on.

New modules can be linked in to the kernel, and old modules delinked, while the system is running. In most cases, it is not necessary to bring down the system for reboot.

The operating system is highly flexible in use. The user can exercise a high degree of control from a single window over several jobs, including both foreground and background processes.

Linux incorporates features of both major streams of the UNIX world: AT&T UNIX and Berkeley Standard Distribution. For example, there are several shells (CLIs) available under Linux. The first popular UNIX shell was the Bourne Shell. The "Bourne Again Shell" (bash) is an improved version of the original UNIX Bourne Shell from AT&T UNIX. The developers of Berkeley UNIX produced the "C Shell," which differs in several key aspects from the earlier Bourne shell. Linux offers the TC Shell (tcsh), which is an improvement on the C Shell. The "Z Shell" (zsh) is a hybrid, and contains several features of both bash and tcsh, and of the Korn Shell as well. Several additional shells are also available. Some of the more valuable facilities once unique to one or another of the primordial shells have now been incorporated into one or more additional shells, thus to some extent reducing the differences among them in facilities offered. For example, the C Shell of Berkeley UNIX introduced a *history* facility, which had not been present in the original Bourne Shell. History retains some number of most recently issued commands. The retained commands can then be re-executed, either in the identical form to which it had originally been issued or in altered form. Bash incorporates a history facility, thus in some small measure reducing the number of features that distinguish bash from tcsh. Although the facilities that originally distinguished certain shells have been incorporated under Linux into other shells as well, thus diminishing the uniqueness of the feature set offered by each separate shell, the original differences in command syntax between the various shells have been retained.

Several additional UNIX features are also supported by Linux. Redirection of both input and output is a simple means of executing a program and providing the input from a file instead of from the console or standard input device (keyboard), and of having the program's output written to a file instead of to the standard display device (usually either a video display tube or a flat-panel liquid-crystal display). Many other operating systems also allow input and output to be redirected to files, but redirection in other operating systems can be a cumbersome affair (e.g., in VMS). Filters are also prevalent in Linux, as they are in UNIX. These are programs designed specifically to process a stream of input data and to produce a stream of output data. Filters often intervene to process output data from one program before the data are input into another program. Linux also supports both hard links and soft links to facilitate the sharing of files. A hard link points to the file's physical structure on the disk, whereas a soft link connects references to the file through the directory structure. A large number of UNIX utilities are also supported. These include about 15 extremely helpful utilities that support either simple tasks, including

date (writes the current date and time to the standard output device);

echo (specifies a text string to be written to the standard output device); and

lpr (sends a file to the printer);

or routine file manipulation and disk management tasks, including

cat (concatenates files and displays the concatenation on the standard output device);

cd (changes the present working directory to a specified new location);

chgrp (changes the group ownership associated with a specified file);

chmod (changes the permissible access mode for a file);

chown (changes the ownership of a file);

cp (makes a duplicate copy of one or more files, either in the same directory or in a different, specified directory);

df (prints on the standard output device either the number of free disk blocks or the size in kilobytes of the free disk space, and the number of free files);

dd (copies files between devices, converting between different block sizes when so specified);

ln (creates either a hard link or a soft link to a file);

ls (lists the contents of a directory);

mkdir (creates a new directory);

rmdir (deletes specified directories);

mv (moves and renames files and directories); and

umask (sets the default access permission bits for all subsequently newly-created files and directories).

There are also several well-known complex or special-purpose utilities. These include

awk and grep (used for searching for patterns in a file);

sed (a batch editor);

finger (provides specific information regarding authorized users of the system);

make (used principally to update a set of executable programs after modifications are made to the source code); and

pine (used both to receive and send news and e-mail).

Shell script is the term used in the UNIX culture to denote what is usually termed a *command procedure* in other operating systems. A command procedure or shell script is a list of operating-system commands that may contain program-like features. If there is a distinct ordered list of operating-system commands that the user needs to execute repeatedly, for example, immediately after every log-in or immediately before every log-out, then most operating systems have a facility for recording the list of commands in a file, which can then either be executed automatically upon log-in or log-out, or can be invoked by the user through the issuance of a single command that results in the execution of the entire contents of the batch file, which can contain as few as one operating-system command or as many as thousands. In most cases, both interactively typed operating-system commands and prerecorded batch command procedures (like the *.bat files of MS-DOS/PC DOS and the *.cmd files of Windows) are executed through interpretation. That is, each command in turn is examined, parsed, and translated in turn into a series of executable machine instructions. Several of the Linux shells, however, provide a facility for the user to write command procedures called *shell functions,* which are retained in memory in partially parsed (preprocessed) form. This allows the command procedure to be executed more speedily when called, as it is not necessary to do all of the work of parsing every time the command procedure is executed. Note that the shell function can be stored in memory either in bash or in zsh. In addition, zsh has a facility called autoload that allows the user to specify a shell function that is not immediately loaded into memory but is, instead, stored on disk, thereby economizing in the use of memory. Only when the shell function is actually invoked is it copied from disk to memory, thus economizing both in the setup time and in the encumbrance of memory, which can be significant if there are many shell functions seldom used.

Additional features from the UNIX world supported in Linux include two very powerful editors, vi and emacs, as well as a wide variety of utilities.

Support is provided for a huge variety of peripheral devices and adapter cards.

Ample tools and facilities are provided for software development, including, in addition to make, the *Concurrent Versions System* (CVS) and the *Revision Control System* (RCS). These utilities are very useful not only for developing new software, but also for taking existing software and both configuring and installing it to run either on a different hardware architecture or on a different operating system.

Several emulators are available to enable the running of programs designed for other operating systems. The operating systems supported include Mac OS for the MacIntosh (executor), the old DOS operating system that ran on the early Intel hardware (dosemu), the Windows environment (wine and wabi), several flavors of UNIX, including

AT&T's System V Release 4, the University of California at Berkeley's Berkeley Standard Distribution (BSD), and Santa Cruz Operation (SCO) UNIX. In addition, GNOME and KDE provide Windows-like environments for running modern Windows applications under Linux, and Plex86, distributed by Debian, allows other operating systems designed for the Intel 80 × 86 hardware environment to run under Linux on Intel 80 × 86 hardware. There is also a commercial product called VMWARE that comes in versions that allow other operating systems to run under Linux. Note that several of these operating-system emulators require licenses as well as executable code for the operating-system environment being emulated.

Note that Linux is currently reported as being installed and used in a total of 183 distinct countries, pseudo-countries, and other geographical areas, such as the Faeroes Islands (http://counter.li.org). The total number of users is currently estimated as being somewhere between 3 and 24 million, with a "best guess" of 18 million (http://counter.li.org/estimates.php).

HISTORY OF LINUX

Linux is not an isolated phenomenon. It arose as the culmination of what can be considered in computer science terms to be a rather lengthy history of development. Understanding what were the antecedents of Linux and how it came into being provides substantial insight into what Linux is all about.

MULTICS

The Linux story begins with the development in 1965 of an operating system called MULTICS (*Mult*iplexed *I*nformation and *C*omputing *S*ervice). MULTICS was a joint effort undertaken originally by MIT and General Electric (GE). Shortly after the work was begun, GE decided to get out of the computer business, and sold off its computer operation to Honeywell, which thus became the successor to GE. Subsequently to initiation of the project, AT&T's Bell Laboratories joined the development team, thus increasing the team's membership from two to three. However, in 1969 Bell Labs withdrew from the MULTICS project.

UNIX

Two members of the Bell Labs team that had been working on the MULTICS project were Dennis Ritchie and Ken Thompson. There was a particular computer game that they had enjoyed playing that had run on the MULTICS system, and they were frustrated that due to their employer's withdrawal from that project they were no longer able to play their favorite game. There was an unused Digital Equipment Corporation PDP-7 computer available in their laboratory, so Thompson developed a simple operating system to enable them to port the game to the PDP-7. He originally wrote the operating system in PDP-7 assembler in 1969, and gave it the name UNICS as a play on the name of the MULTICS system they had previously been working on. Subsequently, the spelling was changed to UNIX.

As it was originally developed, UNIX was not portable. It was closely linked to the PDP-7 hardware for which, and in whose assembly language, it was originally developed. A need for portability was soon recognized, however, and in order to facilitate the achievement of portability, Thompson developed a new programming language. The new language was derived from a pre-existing language called BCPL, and it was given the simple name "B." The new language provided, unlike most previous languages available at that time, such as COBOL and FORTRAN, an unusual level of access to the hardware of the machine. Subsequently, Ritchie developed another language derived from "B," which he called "C." Then Ritchie and Thompson together rewrote UNIX in "C" in 1973. They did this to make it easily possible to port most of the operating system to other hardware, as most of the software could be compiled and executed on any computer for which a "C" compiler was available. Only a relatively small part of the code of the kernel, which had to be written in assembly language, would have to be rewritten in order to allow execution on the new hardware.

AT&T recognized that UNIX had some fine qualities, although initially they appeared to be oblivious to its commercial potential. Therefore, it was initially distributed free of charge. Consequently, there developed early on a UNIX culture in which contributions of code for use in UNIX were made from all over the world. AT&T belatedly realized that UNIX had commercial potential. They therefore claimed it as their intellectual property and undertook a very far-sighted strategy to continue popularizing the operating system. They licensed it to academic institutions at very low rates. As a result, several generations of computer science students were exposed to UNIX at a very early stage in their careers.

According to the Bible, at one point in its development the human race enjoyed a common language: "Now the whole world had one language and a common speech" (Genesis 11:1). The UNIX world at one time enjoyed the same homogeneity. Just as in general human civilization, so, too, for UNIX this state of affairs did not last very long. The University of California at Berkeley started its own separate dialect of UNIX, which featured the C shell, a CLI that was decidedly different from the Bourne shell, which had become popular early on in the UNIX world. There were other differences that also distinguished the University of California at Berkeley's BSD. AT&T tried to control and standardize UNIX around their evolutionary dialect, which eventually became "System V Release 4" (SVR4). Furthermore, individual hardware vendors started getting into the UNIX business, and one of the first things a hardware vendor tries to do is to differentiate its product from that of its competitors to facilitate its goal of selling more of its hardware. Thus, not only did the two principal streams of UNIX diverge from each other, but also each one developed into multiple dialects. One of the major advantages of having a single standardized operating system in use on different hardware platforms is that the user does not have to undertake a major personal training program every time he migrates to a new hardware platform. If the UNIX dialect in use on each platform is different, however, then an appreciable portion of the advantage of having a common operating system is lost.

Another problem afflicting the UNIX world was the set of licensing restrictions. Commercial UNIX was supplied without source code, as is the usual practice with fully proprietary operating systems, such as Hewlett–Packard/Compaq/DEC's VMS, the various members of Microsoft's Windows family, IBM's OS/2, Apple's Mac OS, and various other proprietary operating systems used on PCs, workstations, minicomputers, and mainframes. Not only is the source code not supplied, but also usually the license provisions explicitly forbid the user from disassembling the software, that is, from generating an assembly-language source code program by constructing it from the executable machine code. If the user were to have access to an assembly-language source code program, then he would be free to study the program, attain an understanding of how it works, and possibly modify the program, either to fix program errors or to introduce new functionality not present in the original version. Having access to the original source code is more valuable in this regard than a disassembled program, because the original source code usually contains identifiers (names of variables and of methods, functions, or procedures) that are meaningful and that hint at the function carried out by the item identified. In addition, the original source code usually includes embedded comments put there specifically to help the person reading the code to understand the design concept as well as the function or purpose of constants, variables, methods, functions, procedures, or individual lines or groups of lines of assembly code. A disassembled program, on the other hand, will usually assign identifiers arbitrarily, and therefore the identifier names provide no hint whatsoever at the functionality of the identified items. Also, of course, a disassembled program will include absolutely no explanatory comments.

In following a policy of both not revealing the source code and also explicitly forbidding the user from disassembling the executable code, the operating-system vendor typically keeps the user dependent upon him both for fixing any errors that may be present and for making any needed functional improvements. The vendors' policy is eloquently summarized by Stallman (1999): "If you want any changes, beg us to make them."

The history presented here of the roots of Linux in the UNIX world has necessarily been very brief. Much more extensive coverage of the history of UNIX can be found in the separate chapter in this volume covering UNIX.

Richard Stallman, the Free Software Foundation, and the GNU Project

Stallman had watched the UNIX environment deteriorate from one of cooperative development, with a lot of sharing of code and ideas and a spirit of community, to an insular environment in which individual vendors competed with each other and foreclosed any possibility of user involvement in the development of the operating system or of mutual cooperation to fix and improve. In reaction to this development, Stallman founded the Free Software Foundation (FSF) in 1984 and initiated the FSF's major activity, the GNU Project. He quit his job in the MIT Artificial Intelligence Laboratory, both so that he could devote himself fully to the project and so that his employer would not be able to claim the developed system as its intellectual property.

The principal project undertaken by the Free Software Foundation was the development of a UNIX-like operating system that would be totally free of proprietary code and thus would be open, giving programmers the ability to study and modify the code at will and, in particular, to share in the benefit made by improvements contributed by others. The overall name for the project was "GNU," a name that stands recursively for "GNU is Not Unix." Development started on proprietary UNIX systems, but as various code modules were developed, these were swapped one by one in exchange for their proprietary cousins, and eventually the entire system was implemented in nonproprietary open-source ("free") code.

The major goal of the GNU project was to produce a UNIX-like operating system totally as an open-source product. Nevertheless, several decisions were made that exerted strong influence on the technical aspects of the project. Three of these are especially worthy of note. First, although 32-bit microprocessors were relatively new in 1984 when the project was initiated, it was decided that 16-bit processors would not be supported. Second, the traditional UNIX emphasis on minimization of memory usage was dropped, at least for utilization of up to 1 Mbyte. Finally, wherever possible dynamically allocated data structures were used, in order to avoid the problem of arbitrarily determined size limits. These decisions have had far-reaching effect on the efficiency and performance of GNU/Linux.

The various parts of GNU were developed over a period of several years, leaving the kernel for last. The concept adopted for the GNU kernel was that it would be based upon the Mach microkernel. Mach was originally developed at Carnegie-Mellon University, and was subsequently continued at the University of Utah. The major advantage of Mach over previous UNIX and UNIX-like kernels was its capability to support loosely coupled multiple processors. The GNU kernel was therefore based upon a collection of servers, in the terminology of the project a "herd of GNUs." It was therefore given the name "HURD." (One of the delightful features of the Linux culture is the sense of humor that pervades it.) This was a functionally very useful concept, but it did have a distinct downside. Development of the GNU kernel was very much drawn out in time because of the difficulty involved in debugging the asynchronous multithreaded servers inherent in the concept that must pass messages to each other (Stallings, 2000). This extended development time for the GNU kernel was the sole remaining obstacle to fielding the entire GNU operating system. Thus, the time was ripe for Linus Torvalds to step in with his Linux kernel. This had a less ambitious design concept than the HURD kernel being worked on by the GNU stalwarts, but although differently conceived, it was very well matched to all of the extra-kernel elements of the GNU project. It came along at just the right time to be plugged into the main body of GNU software and thus to complete the free and open-source UNIX-like operating system that was the goal of GNU.

Andrew Tanenbaum and MINIX

Due to the evolution of UNIX into a number of disparate proprietary dialects, as well as its ever-increasing complexity and the absence of source code, UNIX had evolved into a system that was no longer well suited to teaching about operating systems. Two well-known faculty members responded to this situation by designing UNIX-like operating-system kernels that were intended to meet pedagogic needs. Robert Comer developed XINU (note that XINU is UNIX spelled backwards), and Andrew Tanenbaum produced MINIX. These were both available at low cost, and with complete disclosure of source code, thus lending themselves well for use by students. The student could study and understand all of the code, and could also produce his own modifications. From a pedagogical perspective, this results in a very good grasp by the student of how an operating system works.

Linus Torvalds and the Initial Development of Linux

A graduate student in computer science named Linus Torvalds studied operating systems in 1991 at the University of Helsinki in Finland, using Andrew Tanenbaum's textbook and his UNIX-like kernel MINIX. Torvalds was not satisfied, however, both with the MINIX kernel, because of its limited capability, and with the MINIX file system. He therefore decided to try to develop a more capable as well as a pedagogically useful UNIX-like kernel on his own. Torvalds recognized, however, that this was a major undertaking, particularly in light of the modest amount of experience that he had had with operating systems at that time (he was then a relatively junior graduate student). Therefore, when he started work on the new kernel in April 1991, he posted a notice in a widely read news group to inform others of his new project and to solicit their collaboration. The solicitation was successful, the code was posted, corrections were made and also posted, and the first kernel was released publicly in October 1991. Originally, this was just a fun pastime for computer geeks. However, the code evolved into a mature form, and eventually a stable version of the kernel was produced. The first release of the stable kernel took place in March 1994.

Why Is Linux So Much More Popular Than Its Progenitors?

Linux greatly exceeds in popularity its immediate ancestor, Minix, as well as most flavors of UNIX, from which it is derived. It is much more robust and capable than Minix, probably due in large measure to its multiauthored origin together with the "Copyleft" provisions that were so critical in encouraging voluntary contributions to the development of the code. Also its magnificent stability and the uniformity of look and feel across platforms are probably the major reasons, in addition to the openness of the source code, for the headway made by Linux against the various proprietary flavors of UNIX.

Origin, Appropriateness, and Proper Pronunciation of the Name Linux

Torvalds originally used the name Linux only privately, and attempted publicly to name his new kernel Freax. An associate, Ari Lemmke, who ran the FTP site used originally to distribute the code, didn't like that name, and therefore decided to use Linux instead. Thus, Linux was born.

Strictly speaking, the name Linux should be applied only to the kernel, which is the particular contribution made by Linus Torvalds and collaborators. The great majority of the code included in the various distributions of Linux belongs not to the kernel but to other, noncontinuously memory-resident parts of the operating system, such as those comprising the GNU project, which was nearly complete except for the kernel before Linus Torvalds came on the scene. Stallman (2000) has contended that the most appropriate name for the total software package *should* really be GNU/Linux, reflecting the fact that the contribution of GNU code to the entire enterprise is greater than that of Linux, and that the name Linux belongs principally just to the kernel. Despite the attractiveness of Stallman's point, however, most of the world community has come to accept the name Linux for the entire package. This may be not completely fair to the many programmers who made immense contributions of time and effort to the total body of the software both in pre-Linus Torvalds days and subsequently, and in particular to Richard Stallman himself, who almost single-handedly developed and promoted the concept of free software, together with the Copyleft that has been crucial to its success, and who also initiated the GNU project, nurtured it, and led it to success by dint of substantial personal sacrifice as well as incisive leadership. Nevertheless, the world has accepted the name Linux for the entire package, including the part contributed by GNU, and therefore the name Linux is retained here.

What is the proper pronunciation of Linux? There are two principal streams of thought. Some computer folk use a pronunciation based upon the Anglicized pronunciation of Torvalds' first name. Native speakers of English commonly butcher the pronunciation of other peoples' languages, including the pronunciation of foreign names. Thus, the typical pronunciation of the name Linus by most speakers of English is "Lye'-nuhs." This pronunciation is normally used, for example, both for the name of the character in Charles Schultz's *Charlie Brown* comic strip and for the name of the two-time winner of the Nobel Prize (chemistry and the Nobel Peace Prize), Professor Linus Pauling of Cal Tech. The sound of Linus as pronounced by most speakers of English is similar to the English pronunciation of the Persian name Cyrus (Sigh'-ruhs). The equivalent pronunciation of the name of the operating system, Linux, would thus be "Lye'-nuks." However, in Swedish the correct pronunciation of Linus is "Lee'-noos," and Linus Torvalds himself uses the pronunciation "Lee'-nooks" for the name of the operating system. According to Eric S. Raymond's *Rampantly unofficial Linus Torvalds FAQ* (http://www.catb.org/~esr/faqs/linus/), although Linus Torvalds comes from Finland, his native language is not Finnish but Swedish, there being a considerable minority of native Swedish speakers in Finland. Finland has two official languages, Finnish and Swedish, and residents of localities that have substantial native-Swedish population are entitled by law to carry out their communications with government entities in Swedish (http://virtual.finland.fi/finfo/english/finnswedes.html).

This author is strongly of the opinion that English-speaking people should pay foreigners the courtesy of at least attempting to pronounce their names authentically, and is therefore a strong advocate of the pronunciation "Lee'-nooks." The modern miracle of the Internet allows us all to hear the actual voice of Linus Torvalds pronouncing the name Linux (ftp://ftp.kernel.org/pub/linux/kernel/SillySounds/english.au).

GNU and Copyleft, and Their Impact on the Evolution of Linux

In setting up the Free Software Foundation, Richard Stallman very creatively adapted the provisions of copyright law to his purpose of assuring both (a) that the source code would perpetually remain open and (b) that a vendor would be strictly prevented from making derivative works from the free software and making these derivative works proprietary, and then legally exerting control over the users and enticing them away from the free product toward the vendor's proprietary derivatives. To accomplish these goals, Stallman copyrighted the software, and required, as a condition for obtaining a license to use it, that all derivative works made from Free-Software-Foundation-produced software and subsequently copied and distributed, be covered by the identical copyright. There could be no license fee required for the use of the software, although a modest charge could be made to defray the cost of distribution media. Also, it was required that the source code either be distributed in its entirety together with the executable code or be made readily available for no more than a modest charge for reproduction. Deliberate obfuscation of the source code is prohibited, as is restriction on further modification and derivation of new software (http://www. opensource.org).

A friend of Stallman's, Don Hopkins, sent him a letter around 1984/5, on the envelope of which he had scribbled, "Copyleft—all rights reversed." This summed up very nicely Stallman's use of copyright law to accomplish the opposite of what copyright normally does. Instead of restricting the user's right to the software, it guaranteed completely unobstructed right. Therefore, Stallman adopted Hopkins' term, "copyleft," and used it to refer to the GNU Public License Agreement, or GPL. This remains the standard license under which all GNU software and the Linux kernel are distributed today.

Worldwide Cooperation and Support: A Grassroots Movement

Linux is not just an operating system; it is a culture. The concepts promulgated by Richard Stallman under the aegis of his brainchild, the Free Software Foundation, have become the bywords of the Linux movement. There is a small army of enthusiasts who enjoy the camaraderie and the easygoing humor of the Linux community. Anyone who has the technical smarts, as well as the time and will, may participate. The Linux community is a true meritocracy. No one needs to know, nor does anyone care, what is your race, religion, nationality, or sexual orientation, what you look like, or how you dress. Whether you have a pleasant or an obnoxious personality makes very little difference. Unlike the situation in the commercial world, there are no supervisors or executives to suck up to; all that counts is the quality of your contribution. In addition, the key to its astonishing success has been the Internet. As Linus Torvalds himself has said, "One of the most important and unique facets of the Linux development project has been the effect that feedback (mostly via the Internet) has on development: feedback accelerates development dramatically" (foreword to Sobell, 1997). Also, "Linux wouldn't be what it is today without the Internet and the contributions of an incredible number of people" (Sobell, 1997).

HARDWARE ARCHITECTURES SUPPORTED

In the commercial environment, in most instances an operating system is designed for a single hardware architecture. In fact, the kernel of the operating system, which is the part that must reside in memory starting immediately after boot-up and continuing until system shutdown, is highly specific to the hardware architecture, so much so that the name "UNIX," for example, although widely thought of as a single operating system, is, in reality, a family of operating systems, since the kernel that implements UNIX on each and every different hardware platform is different from every other UNIX kernel. This despite the fact that the different versions of UNIX implemented on two particular hardware architectures may have a totally common look and feel.

In the UNIX world, in fact, in addition to the various kernels, each specific to a particular hardware architecture, there are also several varieties of UNIX, of which the major variants are AT&T's SVR4 and the BSD. The look and feel of these major variants are similar, but definitely *not* identical.

In the case of Linux, again we are really dealing not just with one single operating system but rather with a family of operating systems. However, Linux comes in only one flavor, the look and feel being nearly identical across all hardware architectures. The kernel is standardized for each hardware family. Linux runs on quite a number of hardware architectures, including the Amiga, the Apple–IBM–Motorola PowerPC, the Atari, the DEC/Compaq/Hewlett-Packard Alpha, the Intel 32-bit processor family (consisting of the 80386, the 80486, Pentium, Pentium II, Pentium III, Pentium 4, Celeron, and Xeon), the MIPS 10000 series, the Motorola M68000 series, the IBM System/390 zServer iSeries architecture, and the 64-bit architectures produced by Intel (Itanium) and AMD (Sledgehammer).

DISTRIBUTIONS OF LINUX AND VENDOR SUPPORT

A *distribution* of Linux consists of a package containing all of the software necessary to both install and run Linux. There is an Internet site that lists many Linux distributions (http://www.linuxhq.com/). A recent search on this site revealed 164 distributions, some of which are specifically geared to particular languages. There are, for example, distributions listed for the Arabic language (http://www.haydarlinux.com/), as well as for Hebrew (http://ivrix.org.il/). There are seven principal Linux distributions for English-speaking users. One of these, Debian, is produced by a nonprofit corporation called "Software in the Public Interest, Inc." The Debian Linux distribution is produced through the contribution of volunteers. Other commonly used Linux distributions are commercial, and are listed in alphabetical order, together with the percentage of users reported to have installed each distribution, according to the Linux Counter (http://counter.li.org): Caldera (percentage not reported), Connectiva (1.2%), Corel (percentage not reported), Mandrake (20%), Red Hat (30%), Slackware (12.1%), and SuSE (11.7%). The noncommercial Debian distribution is reported to be installed on 13.2% of users' machines. Thus, the five most commonly used distributions account collectively for more than 85% of the installed base of machines. On May 30, 2002, four corporations that produce Linux distributions. Caldera, Connectiva, SuSE, and Turbolinux, announced that they were combining forces to attain a uniformity of their distributions. They are naming their uniform version UnitedLinux.

THE FUTURE OF LINUX

At present, there are two principal factors limiting the expansion of Linux. First is the limited number of mainstream software applications that run under Linux. In the words of Linus Torvalds (foreword to Sobell, 1997), "Linux has the same problem all other operating systems have had: a lack of applications . . . . In order to be a real driving force, Linux needs to have more applications, and those applications need to be readily available with wide distribution and low price." However, he is optimistic regarding the future, as he also states, "So far Linux ports of various applications have followed DOS/Windows pricing (less expensive) rather than the UNIX pricing (more expensive), so I'm reasonably hopeful we can have the types of applications that will help Linux grow." In fact, several vendors have released versions of their mainstream applications ported to Linux. The number of such applications is still modest, but is growing.

The other major factor limiting the expansion of Linux is the natural hesitancy of corporate IT managers to rely upon major software that is not a product of a monolithic "reputable" vendor. The corporate world feels more comfortable when dealing with entities that can be sued if something goes wrong with the software. The Linux culture doesn't fit into the corporate mold. Some IT managers, nevertheless, have listened to their technical experts and have taken the leap to Linux. There is a modest, but growing foothold occupied by Linux in the commercial world. Augmenting this process are two packages that allow Windows applications to run under Linux: KDE and GNOME. Unfortunately, not all Windows applications run successfully in one of these environments. If either of them should develop to the point where nearly everything runs, or if a new windowing environment for Linux were to appear incorporating an improvement in functionality providing universal compatibility with all Windows applications, then that could have a major impact in expanding the acceptance of Linux.

Where will *Linux* be in another year, or in 2, 5, 10, or 20 years? Prediction is a notoriously unreliable endeavor in the computer industry, probably even more so than in general society. The further out we look, the harder it is to predict with any level of accuracy. For example, Bill Gates is supposed to have declared in 1981 that 640 Kbytes is as much memory as anyone would ever need. In addition, it is said that Ken Olson opined in the 1960s or 1970s that there would never be computers in individual households. These are both very intelligent people who each possessed an excellent grasp of the computer industry at the time they made the pronouncements for which they are famous. The reason that their prognostications were so far off the mark is that developments took place in the industry that were simply unpredictable at the time they spoke. As a rule, the shorter the outlook, the more straightforward it becomes to make a credible prediction. Looking at Linux, certainly in the short term, that is, for the next one or two years, we can expect to see continued growth at a rate comparable to what has occurred in the past few years. Linux enthusiasts expect and hope that this growth will continue beyond the immediate future into the medium- and far-term future as well. Although there are some who expect Linux eventually to unseat Windows from its current position of industry dominance, a more balanced view would not consider that to be a high-likelihood event. The basis for the expectation of accelerated growth is the expected continuation of at-best mediocre performance, insufficient reliability, and erratic operation of Windows, coupled with the traditionally sluggish responsiveness by the Windows software vendor in the fixing of bugs. This latter, in particular, is generally thought to be due perhaps in part to the attitude of the company (let the user find the bugs for us, and let him come to us on his hands and knees begging us to fix them), but is probably due in greater measure to the limited ability of the software producer to manage a project of enormous complexity consisting of millions of lines of code, while also trying to keep up with changes in the hardware environment and increasing demands by users for added capabilities in the software. However, it must be borne in mind that the deeper the inroads that Linux makes in Windows' share of the market, the greater the incentive this gives to Microsoft to become competitive! No one can predict where that will lead. Will Microsoft change its long-standing policy and release the source code to allow critical examination, as well as custom modification, by the programming community? Will they solicit suggestions for code changes and for bug fixes, perhaps offering to pay handsomely for sound code contributed by users? Will Microsoft surprise the world by itself adopting the Linux kernel, thoroughly rebuilding the rest of its operating software to maintain the look and feel of Windows, while achieving the benefits of the open-source movement for continued evolution of the kernel? Alternatively, will some other commercial firm perhaps come up with a new operating system that will be technically competitive with Linux, while also having the advantages of roots in the commercial environment and complete compatibility with Windows applications that would make it competitive with Windows? I certainly do not mean to suggest that a commercial challenger is likely to arise to Windows' dominance of the marketplace, but surprises and revolutions in the technological arena certainly do occur, and such a possibility cannot be glibly ruled out. The high-tech marketplace is littered with the corpses of formerly giant firms that had dominated major segments of the industry, and that subsequently either abandoned the computer marketplace and limited themselves to other areas of business, went out of existence entirely, or have been swallowed up by later or once-smaller upstarts. Names like General Electric, Honeywell, Control Data Corporation, and Digital Equipment Corporation come readily to mind. No one can foretell the future of Microsoft.

EXTENDING YOUR KNOWLEDGE ABOUT LINUX

Stallings (2001) presents an excellent exposition of the general principles of Operating Systems. Moody (1997) contains a very readable introductory essay on Linux, while Stallman (1998, 1999, 2000) presents a concise, well-written, and authoritative summary of the history and evolution of Linux. Sobell (1997) is probably the best overall comprehensive introduction to the practical use of Linux. McCarty (1990), Kofler (1999), and Sarwar et al. (2002) are also good source materials on how to work with Linux, though briefer than Sobell. Guidance on the installation, configuration, and administration of a Linux system is provided in Red Hat (2003a, 2003b). Peterson (2001a, 2001b) provides a more comprehensive coverage of the details of working with Linux, while Bovet & Cesati (2001) describes the kernel internals in considerable detail.

ACKNOWLEDGMENTS

I thank my son, Mordechai Tsvi Abzug, for several conversations about Linux in which I bounced some ideas off him and obtained valuable feedback that contributed to the content of this chapter. I also thank Dr. Mal Lane, head of the Computer Science Department at James Madison University, for giving me the opportunity to teach Operating Systems.

GLOSSARY

Application software Programs usually of commercial origin that effect the end-goals of the user.

Architecture The fundamental design structure of a computer system.

Assembly The process of generating, based on a computer program written in assembly language, object code or executable code which consists principally of a list of machine-language instructions. See also compilation.

Assembly language code Source code that bears a one-to-one correspondence to the hardware instruction set of the computer. See also executable code.

Batch job A task requested to be accomplished by the computer, which can be executed without any further input from or interaction with the user.

Command-line interface (CLI) A facility for the user to communicate his wishes to the computer by typing specific commands a line at a time. See also graphical user interface (GUI).

Compilation The generation, from a source code program written in a high-level language, first of an assembly-language program that describes how the high-level language instructions are to be implemented as a sequence of machine language instructions specific to the hardware of the physical computer, and then the machine language instructions corresponding to the intermediate assembly-language expression of the program. (*Note:* Sometimes the machine-language instructions are produced directly, bypassing the assembly-language step.) See also assembly.

Device A physical apparatus that is part of the hardware of a computer. Usually, a device either stores data on a computer in form sufficiently robust that the data are retained even if the electrical power is turned off, or mediates in the transmission of data between the computer and the external world.

Disassembly A process by which a set of executable machine-language instructions is analyzed and rewritten as a list of assembly language source code instructions. See also assembly.

Executable code A list of instructions both written in machine language and containing all memory locations explicitly specified, and therefore directly executable by the computer. See also source code and machine code.

Graphical user interface (GUI) A facility for the user to communicate his wishes to the computer principally by pointing with an appropriate device and in a designated manner either to pictorial elements displayed on an appropriate two-dimensional surface or to lists (menus) of various options appearing on the display medium. See also command line interface (CLI).

Kernel The portion of the operating system that must reside continuously in the primary memory of the computer from the time of completion of boot-up until system shutdown.

Machine code A program expressed as a list of machine instructions directly executable by the computer, except that some of the references to memory locations might not yet be completely specified. See also source code and executable code.

Multitasking An environment in which several tasks can be executed concurrently on behalf of the computer user, exploiting time periods when one task must remain idle while waiting until a critical event occurs before that task can be resumed, and performing other tasks in the meantime.

Multiuser An environment in which a computer system can serve the needs of several users concurrently by interleaving the execution of tasks for different users.

Open-source A special program development environment in which, in contrast to the commercial software environment, the source code for the software is freely distributed, thus enabling the user community freely to modify the software either to fix bugs in the code or to enhance or alter the functionality of the software to meet the individual needs of different systems.

Portability The ability of software originally written with the intention of being executed in a particular hardware-combined-with-operating-system environment, known as a platform, to be transformed and transported for execution on a different platform.

Real-time A relatively unusual environment in which a computer system operates under defined constraints for responding to specific external events. For example, a collision warning system in an airplane must generate a warning and deliver it to the pilot sufficiently speedily as to enable him to take appropriate action to avoid the impending collision. In a real-time system, the issuance of an error message instructing the user to close down some applications in order to free up excessively encumbered memory space is usually precluded.

Shell A particular CLI having explicitly defined properties. Shells available in the Linux environment include the Bourne Again Shell (bash), the TC Shell (tcsh), the Z Shell (zsh), and the Korn shell (ksh), and generally contain facilities for program-like operations.

Shell script A file containing a sequence of commands addressed to the operating system that facilitates the repeated execution of the included commands without their having to be laboriously retyped each time they are executed.

Source code A complete list of instructions that describes how a computer program is to be executed, written in a precisely defined programming language that is relatively easy for a human reader to understand, yet is also readily converted by appropriate special software into a set of machine-language or executable instructions.

Utilities Software provided together with the operating system, which provides commonly used functions that can greatly enhance the ability of the computer system to provide service to the user.

CROSS REFERENCES

See *Open Source Development and Licensing; Unix Operating System.*

REFERENCES

Bovet, D. P., & Cesati, M. (2001). *Understanding the Linux kernel* (1st ed.). Sebastopol, CA: O'Reilly.

Kofler, M. (1999). Linux (2nd ed.). Reading, MA: Addison-Wesley. [*Note:* This book includes two CDs containing part of the Red Hat Linux distribution. Red Hat has produced several distributions since this version, but in any case this earlier distribution is certainly adequate to get started learning Linux, and it is always possible to update later.]

McCarty, B. (1990). *Learning Red Hat Linux*. Sebastopol, CA: O'Reilly. [*Note:* This book includes a CD containing part of the Red Hat Linux distribution. Red Hat has produced several distributions since this version, but in any case this earlier distribution is certainly adequate to get started learning Linux, and it is always possible to update later.]

Moody, G. (1997). The greatest OS that (n)ever was. *Wired,* Issue 5.08, August 1997. Retrieved April 4, 2003, from http://www.wired.com/wired/archive/5.08/linux_pr.html

Petersen, R. (2001a). *Linux: The complete reference* (4th ed.). Berkeley, CA: Osborne/McGraw–Hill. [Comes with significant parts of several Linux distributions, including Caldera ("OpenLinux eDesktop and eServer"), Red Hat, and SuSE on several compact disks.]

Petersen, R. (2002). *Red Hat Linux 8: The complete reference, DVD Edition*. Berkeley, CA: Osborne/McGraw–Hill. [Comes with all of of the Red Hat Linux distribution on DVD and all Red Hat manuals.]

Red Hat (2003a). *Red Hat Linux 9: Red Hat x86 Installation Guide*. Retrieved April 1, 2003, from http://www.redhat.com/docs/manuals/linux/RHL-9-Manual/pdf/rhl-ig-x86-en-9.pdf [Note that Red Hat also has installation guides for the Hewlett–Packard/Compaq/DEC Alpha, for the Intel Itanium, and for IBM's eS-erver zSeries of System/390 mainframes, in addition to the Intel x86 processor series.]

Red Hat (2003b). *Red Hat Linux 9: The official Red Hat Linux getting started guide*. Retrieved April 4, 2003, from http://www.redhat.com/docs/manuals/linux/RHL-9-Manual/pdf/rhl-gsg-en-9.pdf

Sarwar, S. M., Koretsky, R., & Sarwar, S. A. (2002). *Linux: The textbook* (1st ed.). Boston, MA: Addison-Wesley Longman. [*Note:* This book includes a CD containing part of the Mandrake Linux distribution.]

Sobell, M. G. (1997). *A practical guide to Linux*. Reading, MA: Addison-Wesley. [*Note:* This is a *superbly* written book. It delves very thoroughly and very systematically into the technical aspects of working with Linux. Linus Torvalds learned about UNIX and was inspired to work with it through a predecessor to this book. Probably the best all-around work on Linux available today.]

Sobell, M. G. (2003). *A practical guide to Red Hat Linux*. Reading, MA: Addison-Wesley.

Stallings, W. (2001). *Operating systems: Internals and design principles* (4th ed.). Upper Saddle River, NJ: Prentice–Hall.

Stallman, R. (1998). *Linux and the GNU Project*. Retrieved April 4, 2003, from http://www.gnu.org/gnu/linux-and-gnu.html

Stallman, R. (1999). The GNU Project. In C. Dibona, M. Stone, & S. Ockman (Eds.), *Open sources: Voices from the open source revolution*. Sebastopol, CA: O'Reilly. Retrieved April 4, 2003, from http://www.gnu.org/gnu/thegnuproject.html

Stallman, R. (2000). *What's in a name?* Retrieved April 4, 2003, from http://www.gnu.org/gnu/why-gnu-linux.html

FURTHER READING

History and Development of Linux:
http://www.wired.com/wired/archive/5.08/linux.html
http://counter.li.org/ (an organization that compiles statistics on Linux usage)

Open-Source Initiative: Open-Source Definition:
http://www.opensource.org
http://www.gnu.org

Linux Headquarters:
http://www.linuxhq.com

Linux Documentation Project:
  http://tldp.org/
Linux Distributions:
  http://www.fokus.gmd.de/LINUX/LINUX-distrib.html (describes several distributions)
  http://www.caldera.com (Caldera distribution)
  http://en.conectiva.com/ (Connectiva distribution)
  http://linux.corel.com/ (Corel distribution)
  http://www.debian.org (Debian distribution)
  http://www.mandrakelinux.com/en/ (Mandrake distribution)
  http://www.redhat.com/ (Red Hat distribution)
  http://www.slackware.com/ (Slackware distribution)
  http://www.suse.com (SuSE distribution)
Linux Kernel Archives:
  http://www.kernel.org
Linux Journal:
  http://www2.linuxjournal.com/

Note that there are also various Linux users' groups (LUGs) that are excellent forums for dissemination of information.

Load Balancing on the Internet

Jianbin Wei, *Wayne State University*
Cheng-Zhong Xu, *Wayne State University*
Xiaobo Zhou, *University of Colorado at Colorado Springs*

INTRODUCTION

The explosive growth of the Web technology in the past decade created great opportunities for posting data and services on the Internet and making them accessible to the vast online users. Due to the unprecedented scale of the Internet, popular Internet services must be scalable to support up to millions of concurrent client requests reliably, responsively, and economically. These scalability and availability requirements pose great challenge on both processing power and networking communication capacity. Internet services have become an important class of driving applications for scalable computer systems. In particular, a cluster-based architecture is gaining momentum. The architecture deploys a cluster of networked server nodes that work collectively to keep up with ever-increasing request load and provide scalable Internet services, and load balancing is a key integration that distributes the client request load between the servers for scalable and highly available services. The architecture is often referred to as a Web cluster or a distributed Web server. This chapter gives a survey of state-of-the-art load-balancing strategies on the Internet.

Internet services are applications accessible to remote clients on the Internet via ubiquitous networking protocols. For scalability and availability, large-scale Internet services are often run on Web clusters. According to Hennessy and Patterson (2003), the Google search engine used more than 6,000 Linux/Intel PCs and 12,000 disks in December 2000 to serve an average of almost one thousand queries per second as well as index search for more than one billion pages. A most recent survey of large-scale Internet services also revealed that all the representative services were running on a cluster of hundreds or even thousands of servers at more than one data centers (Oppenheimer & Patterson, 2002). Internet services on a Web cluster rely on the load-balancing technology to distribute client requests between the servers efficiently and transparently.

The concept of load balancing is not new. It has long been developed as a prominent technology in both distributed systems and parallel computers. However, load balancing on the Internet distinguishes itself by its new Internet workload characteristics and unique quality of service (QoS) objectives. In parallel computers, load balancing is to assign parallel tasks of a job onto (virtual) processors with the objective of balancing the processors' workload and meanwhile minimizing the interprocessor communication cost due to the task dependences (Xu & Lau, 1997). Its overall goal is to reduce the execution time of parallel tasks. In traditional distributed systems, load balancing is to schedule the execution of independent jobs with the objective of efficient utilization of system-wide computational resources. It is often implemented in a form of load sharing—ensuring no processor is idle while there are jobs waiting for services in other processors.

By contrast, load balancing on the Internet is to distribute the Web service requests between the client, server, and network for enhancing the level of QoS in terms of the service responsiveness, scalability, and availability. Figure 1 shows a typical information flow infrastructure of the Web services. To access Internet service via a URL address like http://www.yahoo.com, which contains the domain name www.yahoo.com, a client first contacts its local domain name system (DNS) to get the IP address of the domain name. If its local DNS does not contain the IP address, the domain name resolution request goes upward to the root name server until the service's authoritative DNS is reached. The resolved IP address of the domain name is usually stored in the local DNS for reuse in the future. With the IP address, the client sends subsequent service requests to the server. The requests may go through a client proxy, Internet service provider (ISP)

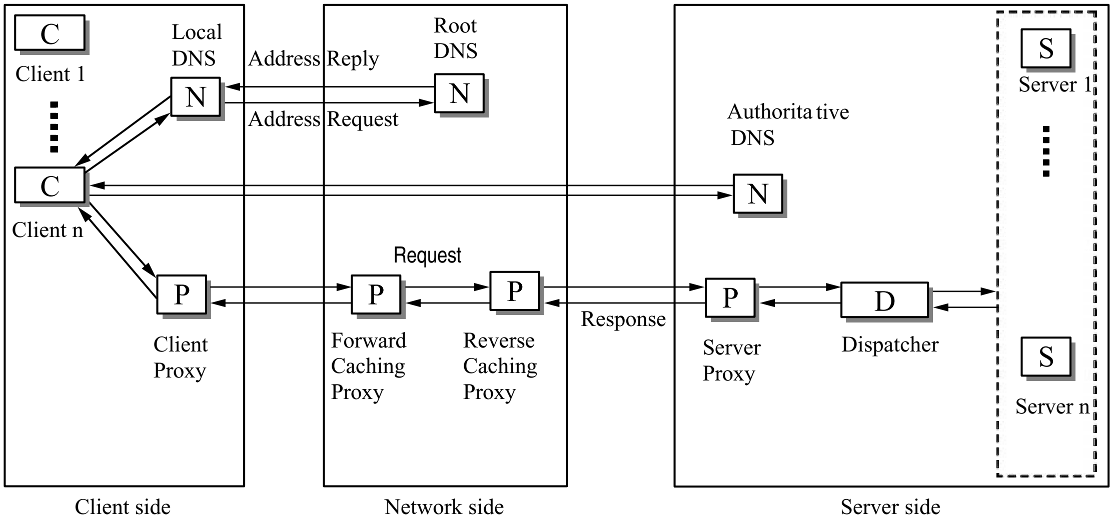


Figure 1: A typical information flow infrastructure of an Internet service.

proxies in the network, and a server proxy. The IP address may be a virtual IP address of a Web cluster. In this case, it is the dispatcher in front of the cluster that decides the actual server in the back-end to serve the incoming requests.

Two primary performance metrics related to QoS are capacity and latency. Capacity refers to the number of concurrent requests that the server can support without causing significant queueing delay. Latency is the service time, as perceived by the client. It is the request processing time in the server, including its possible queueing delay, plus the request/response transmission time in the network. These two metrics are related. A high-capacity server would cause less queueing delay and consequently lead to quick responses to the requests.

For the objective of QoS, load-balancing techniques on the Internet usually make use of multiple servers to process the client requests in parallel. The servers can be either distributed locally at the same site or globally distributed at different geographical locations. We refer to the techniques on locally distributed servers as *server-side load balancing* (or server load balancing) and the techniques on globally distributed servers as *network-side load balancing* (or network load balancing). Server-side load balancing aims at scaling the server processing capacity and communication bandwidth with the increase of the number of participating servers. Due to the scale of the Internet, service transmission time on the network is a significant performance factor, in particular in streaming applications. Wide deployment of broadband networks helps alleviate the problem to some extent. On the other hand, fast networking entices service providers to create services eager for more bandwidth. Network-side load balancing is to move the services closer to clients so as to reduce the network transmission delay on the Internet.

Load balancing on the Internet is also concerned about the workload distribution between servers and the clients. For example, e-commerce services often use mobile codes like JavaScript to verify integrity and validity of user inputs at the client side. This not only avoids back-and-forth data transmission between the server and the client, but also reduces the server-side processing workload. We refer to this type of load balancing as *client-side load balancing* (or client load balancing). A primary objective of client-side load balancing is to improve the quality of Internet services by migrating certain tasks from server to client. Recent advances in peer-to-peer computing further the concept of client-side load balancing by allowing clients to help each other toward the objective of high QoS.

Load-balancing mechanism aside, another important aspect of load balancing is policy. Internet services exhibit different workload characteristics from traditional parallel and distributed applications. Unlike jobs in distributed systems and tasks in parallel computing, the basic scheduling unit of load balancing on the Internet is client request. The requests are nonuniform in size because they may need to access different data and cause different types of operation. Although the requests from different clients are independent, they are actually related if they are about to access the same data. The requests from the same client may belong to the same session and must be handled together. Due to these distinct workload characteristics, many different policies of load balancing have been developed. This chapter surveys representative policies under different load-balancing mechanisms. It starts with a description of Internet traffic characterization, followed by a classification of load-balancing strategies.

WORKLOAD CHARACTERISTICS OF INTERNET SERVICES

Load balancing on the Internet is largely determined by workload characteristics of the Internet services. Two important Internet services are Web and streaming applications.

Web Applications

Web traffic, carried by HTTP protocol over TCP, is one of the dominant components of Internet traffic. It is closely tied to contents of Web pages and to dynamics of the TCP transmission protocol. Past studies of Web workloads concentrated on conventional information provider sites and found several important characteristics common to the sites: target file types, file size distributions, file popularity distributions, self-similarity in Web traffic, reference locality, and client request patterns. An early study summarized the Web traffic characteristics as follows (Arlitt & Williamson, 1997):

- Images and HTML files together account for over 90% of the files transferred. The majority of HTTP requests for Web pages are smaller than 500 bytes. HTTP responses are typically smaller than 50 Kbytes and the median transfer size is small (e.g., less than 5 Kbytes).
- Ten percent of the files account for 90% of requests and bytes transferred. The file popularity distributions are Zipf-like. The file size distributions are *heavy-tailed*. The file transfer size distributions are also heavy-tailed, although not as heavy-tailed as the file size distributions.
- The user request patterns are exponentially distributed and independent (e.g., Poisson distributions). However, significant traffic variance (burstiness) is present on a wide range of time scales. The aggregate traffic generated by many users of the Web has been shown to exhibit *self-similarity*.

The characteristics of Web traffic have a significant impact on choosing load-balancing strategies. It is often assumed that the service time of a request is proportional to the size of its requested file. Simple strategies, like random and round-robin algorithms, are good enough when the requests are uniform in size and independent (Kwan, McGrath, & Reed, 1995). To take advantage of temporal locality in requests, locality-aware distribution strategies make server selection based on the requested content (Pai et al., 1998). In the case that the file transfer size distribution is heavy-tailed, a size interval task assignment with equal load (SITA-E) performs excellent in terms of both mean response time and mean slowdown, provided the system load is not too high (Harchol-Balter, Crovella, & Murta, 1999).

Recent years witnessed a number of major changes in the Web applications (Arlitt, Krishnamurthy, & Rolia, 2001; Padmanabhan & Qiu, 2000). The most notable ones include the following: Web pages are shifting from static content to dynamic content; e-commerce becomes one of the major Web applications; and continuous media are increasingly gaining interests. Dynamic Web pages are generated by programs that run on servers every time the corresponding pages are accessed for producing a different page for every access. Simple examples are the pages that contain displays of a visitor counter or the current date and time. Responses of search engines, stock-quote sites, and personalized Web pages, which are generated on the fly from databases, are typical examples of dynamic pages. Although dynamic pages provide a far richer experience for users than static pages, generating dynamic pages on the fly imposes additional overhead on server resources, especially on CPU and disk I/O. These changes of the workload characteristics pose a challenge to existing load-balancing techniques in performance. Some of the strategies are even no longer applicable. For example, a size-based strategy may not work for dynamic contents because the service time for dynamic content due to its size is unknown . The potential for caching requested files declines and some requested files are even noncacheable because they cause dynamic page generations.

Interest in e-commerce has grown substantially in the past several years (Arlitt et al., 2001; Menasce, Almeida, Fonseca, & Mendes, 1999). E-commerce workloads are composed of sessions. A session is a sequence of requests of different types made by a single customer during a single visit to a site. During a session, a customer can issue consecutive requests of various e-commerce functions such as browse, search, add to the shopping cart, register, and pay. Different customers may exhibit different navigational patterns and hence may invoke the different functions in different ways and with different frequencies. Menasce et al. (1999) present analysis techniques of workload and performance modeling as well as capacity planning for e-commerce sites.

The key QoS issues on e-commerce sites include responsiveness, availability, and security. For responsiveness and availability, it is increasingly common for e-commerce sites to include the capacity of load balancing to avoid bottlenecks or overloading servers. Load-balancing strategies must be session-oriented. That is, they must direct all consecutive and related requests of a session from a given customer to the same server. This is known as session integrity (also referred to as sticky connections). For example, a customer may add an item to the shopping cart over a TCP connection that goes to server 1. If the next connection goes to server 2, which does not have the shopping-cart information, the application breaks. This required coordination in serving of related requests is referred to as client affinity (affinity routing). Hence, the round-robin load-balancing strategies are not applicable in the stateful e-commerce applications while locality-aware request routing mechanisms provide a convenient way to support session integrity.

For security, e-commerce transactions are often encrypted via the secure socket layer (SSL) protocol. Processing SSL transactions puts an extra load on server resources. This means customer's transaction requests and browsing requests have different requirements for the server-side resources and different impacts on the revenue of the sites. On the other hand, an e-commerce merchant tends to share servers handling transaction requests with other merchants in a Web cluster. Load-balancing strategies must deal effectively with these issues. For example, Wolf and Yu (2001) proposed a family of load-balancing algorithms to dispatch various types of customer requests to different servers. The servers are partitioned into two categories, sharable for the public requests (such as browse requests) and not sharable for private requests (such as transactions).

Streaming Applications

The Internet is seeing the gradual deployment of streaming applications, such as audio over IP, video conferencing, and video-on-demand (VoD). These streaming applications generate traffic with characteristics and requirements that differ significantly from traffic generated by conventional Web applications.

Audio and video streams are continuous and time-based. The term "continuous" refers to the user's view of the data. Internally, continuous streams are represented as a sequence of discrete elements (audio sample, video frame) that replace each other over time. They are said to be *time-based* (or *isochronous*) because timed data elements in audio and video streams define the semantics or "content" of the streams. Timely delivery of the elements is essential to the integrity of the applications. Hence, streaming applications have stringent timing and loss requirements. However, unlike typical Web pages, streaming media does not necessarily require the integrity of the objects transferred. There can be some packet loss, which may simply result in reduced quality perceived by users.

Continuous streams are often bulky in transmission. This is especially so for video of a reasonable quality. Media servers that support streaming applications need to move data with greater throughput than conventional Web servers do. Typical audio stream rates remain on the order of tens of kilobytes per second, regardless of the encoding scheme. Video streams span a wide range of data rates, from tens of kilobytes per second to tens of megabytes per second. The server disk-I/O and network-I/O bandwidth can be the resource bottleneck. The characteristics of encoded video (frame size and frequency) vary tremendously according to the content, the video compression scheme, and the video-encoding scheme. In addition, streaming traffic usually consists of a control part and a data part, potentially using different protocols like RTSP and VoIP, while conventional Web traffic is homogeneous in this case. Hence, a traffic manager must parse the control channels to extract the dynamic socket numbers for the data channels so that related control and data channels can be processed as a single, logical session.

Compared to the studies of Web workload characterization, few insightful studies of streaming workloads exist. Past studies of streaming workloads found several common characteristics (Chesire, Wolman, Voelker, & Levy, 2001):

- The video popularity distributions are Zipf-like. The skew parameter a ranges from 0.271 to 1.0.
- Request arrival patterns are nonstationary. For example, in some movie-on-demand systems, the rush hour is usually around 9 PM.
- Users tend to preview the initial portion of a video to find out whether they are interested.

These characteristics and requirements of streaming applications impose a great impact on caching and load balancing. For example, the video-browsing pattern suggests that caching the first several minutes of video data should be effective. However, this partial cache notion is not valid for Web object requests. The large video size and the typical skews in video popularity suggest that several requests for the same object coming within a short time scale be batched and a single stream be delivered to multiple users by multicast so that both the server disk I/O and network bandwidth requirements can be reduced.

Due to the media popularity distributions, data placement methods can also be crucial to load balancing in distributed streaming servers. There are essentially two complementary techniques: striping (i.e., data partitioning between servers) and replication. The main advantages of data striping are high disk utilization and good load-balancing ability. However, a wide-range data striping can lead to high scheduling and system expansion overhead. The cost is especially high when the servers are to be deployed in a geographically distributed environment. Replication tends to isolate servers from each other for scalability and reliability. Nevertheless, it can lead to load imbalance produced by the uneven distribution of media popularity, because the servers storing hot media files can be overwhelmed by client requests.

TAXONOMY OF LOAD-BALANCING STRATEGIES

Internet services are applications that are, as shown in Figure 1, realized through collaborations between three parties: the service provider (i.e., server), the delivery network, and the service requester (client). Accordingly, load-balancing techniques can be deployed in each of these three sides to distribute and deliver the client request load between the participating components in the information flow path. Load-balancing strategies can also be classified according to the type and amount of information about the flow path they take into account. The third categorization dimension of load-balancing strategies is their implementation layer in the network protocol stack.

Load Balancing in the Server, the Network, and the Client Sides

A scalable Internet service is often provided by a group of servers either connected by a local area network (LAN) or distributed over a wide area network. We refer to the locally distributed servers that collaboratively provide the same Internet services on a single location as a Web cluster. Load balancing in a Web cluster focuses on the distribution of client requests between its servers for the purpose of increasing the service responsiveness, scalability, and availability. Three major distribution approaches exist: DNS based, dispatcher based, and cooperative server based. Since all Internet services are advertised by their URLs, which include their domain names, the DNS-based approaches rely on the server-side authoritative DNS server to make load-balancing decisions during the process of domain name resolutions. This process allows the authoritative DNS server to implement different load-balancing policies to select an appropriate server for each client request. The DNS-based approach can be easily scaled from a Web cluster to globally distributed servers and there would be no bottleneck risk in the DNS server. However, the DNS-based approach can only control a small fraction of address-mapping requests due to the presence of caching in the information flow path and cannot directly control subsequent requests because of the address caching in client browsers and other name servers along the path between the client and the authoritative DNS server. Moreover, although the authoritative DNS server can detect failed servers and exclude them from the list of available servers, clients may still send them requests because of the cached IP addresses.

By contrast, the dispatcher-based approach relies on a load-balancing dispatcher in front of the Web cluster to distribute the client requests. However, the dispatcher itself can be a potential performance bottleneck, especially when it needs to process all responses from the servers. This centralized load-balancing strategy is mostly applicable to Web clusters and is hardly extensible to wide area network environments.

The cooperative server-based approach uses a two-level routing mechanism, in which the authoritative DNS server initially distributes client requests to servers based on a simple policy like random or round robin. On the receipt of a request, a server may handle it by itself or redirect the request to another one according to the states of the servers. Unlike the dispatcher-based approach, this kind of approach allows all the servers to participate in making decisions on load balancing. It avoids single failure points and potential performance bottleneck and offers good scalability. On the other hand, request redirections may cause an increase of response time in certain situations.

The objective of load balancing on the network side is to reduce the service delivery overhead by deploying server replicas (or surrogates) close to its prospective clients. Network-side load-balancing techniques can be integrated with caching proxy servers and content delivery networks (CDN). The caching proxy servers store Web resources in anticipation of future requests. It therefore can reduce the network transmission latency and server-side overhead. With the support of CDN, service providers duplicate their services in geographically distributed surrogates around the world (network edge). On the receipt of a client request, the CDN makes use of DNS-based load-balancing strategies to direct the request to the most appropriate surrogate according to the client geographical locations, the CDN topological information, states of the servers, etc. Therefore, DNS-based approaches are also included in the network-side load-balancing category.

On the client side, there are also load-balancing techniques intended to relieve the server workload incurred by the client requests. To access a service provided by more than one server, a smart client can direct its requests to an appropriate replica if the client has some state information about the replicas. In e-commerce services, mobile codes like JavaScript are widely used to verify integrity and validity of the client inputs at the client side to avoid the back-and-forth communications between the server and its clients.

State-Blind versus State-Aware Load Balancing

The second characteristic of load-balancing strategies is the state information considered by the strategies. The state information about a client request includes the client IP address, port number, physical location, request content, request priority, and round-trip time between the client and the server. Client state information like session ID and cookies facilitates client-affinity scheduling to ensure that the requests from the same client in a session are handled by the same server in a Web cluster. The state information about a server includes its static configuration and dynamic workload (e.g., the number of active connections and response times).

State-blind policies assume little knowledge about the state information, except static configuration parameters. Since they do not rely on content-specific information or dynamic workload information about the servers, their implementations incur little run-time overhead. However, the state-blind policies may make poor assignment decisions due to highly variable client requests and the change of server workloads. State-aware policies, in contrast, tend to make adaptive decisions in response to the request content information and/or the workload change. The state-aware policies can be further classified into three categories: client state aware, server state aware, and client and server state aware.

Load Balancing at Different Network Layers

The third classification dimension of the load-balancing strategies is the network layer of their implementations. This implementation issue has a great impact on the performance and efficiency of load-balancing strategies because it determines the possible information to be used by the strategies and the related overhead. Two major implementation strategies are layer-4 (TCP layer) and layer-7 (application layer) implementation with respect to the OSI network stack.

In a layer-4 implementation, information like IP address and TCP port number can be used to select a server for each request. However, it is limited to content-blind load-balancing policies because this approach cannot get any application-specific information in this layer. By contrast, a layer-7 implementation can take full advantage of the information obtained from the lower layers. For example, a session identifier can be used for session affinity scheduling in e-commerce services; request-specific contents from the application layer can be used for content-aware scheduling.

In summary, we present our classification of the load-balancing strategies in Figure 2. It also contains representatives of each category in both academia and industry. Details of the strategies are reviewed in the next three sections.

SERVER-SIDE LOAD BALANCING

DNS-Based Load Balancing

Recall that an Internet service is often advertised by its URL. As shown in Figure 1, a domain name resolution

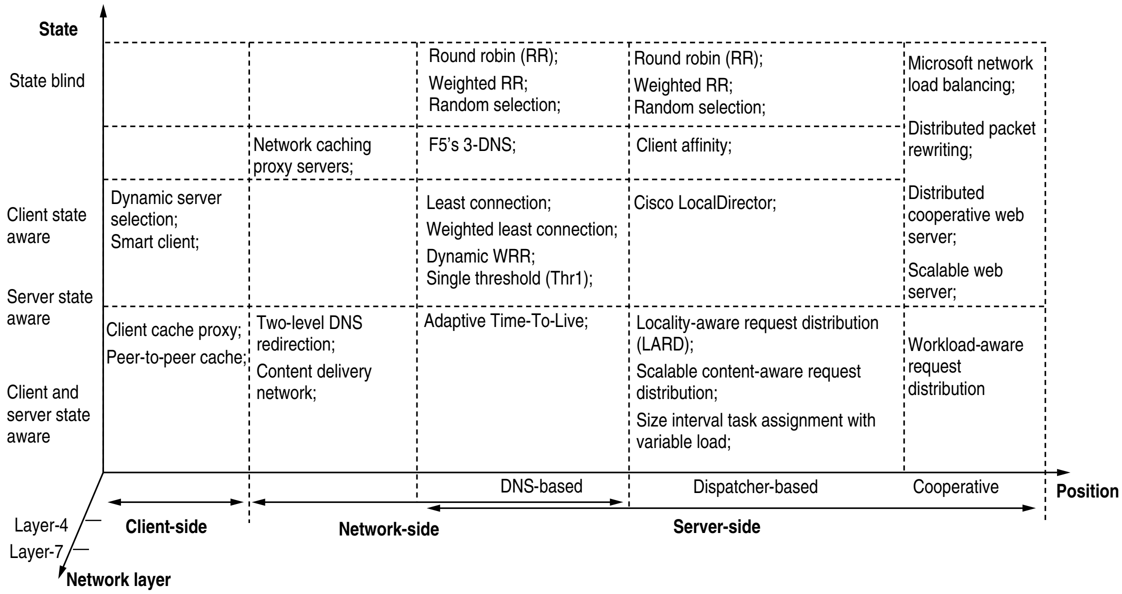


Figure 2: Taxonomy of load-balancing strategies on the Internet.

is carried out by the DNS servers along the path from the client browser to the service authoritative DNS server. DNS-based load-balancing strategies rely on the authoritative DNS server to map the domain name to an appropriate server IP address for each client request. To reduce the response time for future requests, the resolution is usually cached on the intermediate DNS servers and browser cache. Each resolution has a property of time-to-live (TTL) set by the authoritative DNS server. It determines a valid period of the resolution in the caches.

State-Blind Load Balancing

The simplest DNS-based load-balancing approach in a Web cluster is round robin (Kwan et al., 1995). In this approach, the DNS server assigns the IP addresses of the cluster servers in a round-robin manner to the clients' address–mapping requests. It uses a circular server list and a pointer to the last selected server to make the decision. That is, if the last assignment was to server S_i, then the current request will be assigned to $S_{(i+1)\mathrm{mod}\,n}$, where n is the number of servers. Random assignment is another simple state-blind approach, in which the DNS server assigns requests to the servers randomly. Statistically, given a large number of requests, this approach can guarantee an equal distribution of client requests to the available server IP addresses.

The round-robin approach can be easily extended to clusters with heterogeneous servers. A variant of the round-robin policy is weighted round robin (WRR). It assigns requests to the servers according to the relative weights of the server capacities. The weights can be calculated in terms of static server configuration information like CPU speed, memory size, and network bandwidth. We refer to this as *static WRR*.

Client State-Aware Load Balancing

The client state-blind approaches can be extended to take into account some client-specific information in load-balancing decisions. F5's 3-DNS Controller (F5 Network, n.d.) implemented a strategy based on proximity-based address mapping for geographically distributed servers. By comparing the IP address of the client with the IP addresses of the available servers, the controller can determine the proximity of the servers and assign the client's requests to the nearest server.

Server State-Aware Load Balancing

There are also DNS-based approaches that utilize server-side state information for load balancing. A *least-connection* policy allows the DNS server to assign requests to the server with the least number of active connections. The server workload information can be either periodically reported by the servers or polled by the DNS server on demand. A variant of the least-connection policy is weighted least-connection policy for clusters with heterogeneous servers. Denote W_i and C_i as the server weight and the number of active connections, respectively. In the weighted least-connection policy, the DNS server will assign the next client request to the server with minimal C_i/W_i.

Another example of the server state-aware policies is *dynamic WRR*. In this policy, the DNS server calculates the server weights periodically in terms of their dynamic workload such as the number of active connections or measured response time, and assigns each incoming request to a server in the same way as WRR. A critical question is how often does the DNS server need to collect the server workload information? An alternative to periodical workload information collection is an asynchronous alarm mechanism, in which the servers keep evaluating

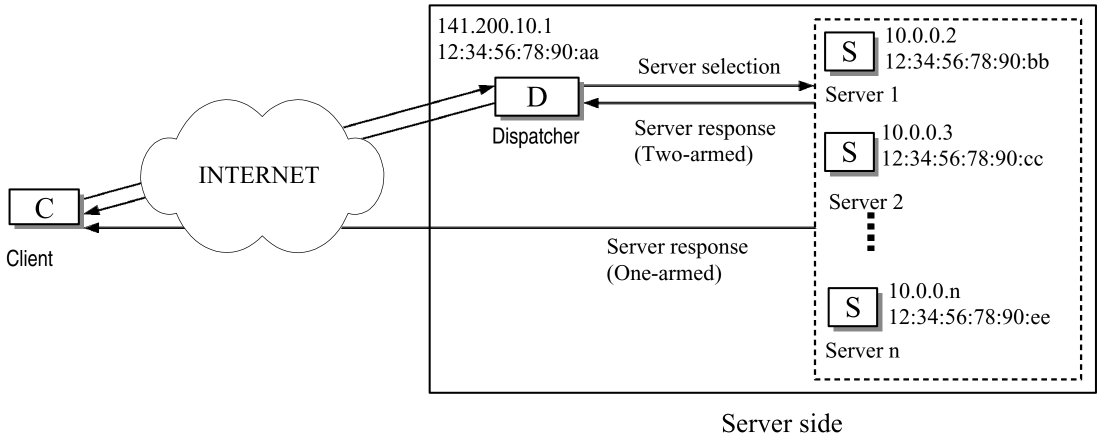


Figure 3: A dispatcher-based load-balancing infrastructure.

their resource utilization and send the DNS server alarms only if their utilization has exceeded a predefined load threshold. The DNS server excludes any alarmed server from the available server list until the server load is back to normal.

Client and Server State-Aware Load Balancing

There are DNS-based load-balancing approaches that take into account the state information about both sides. An example is *adaptive TTL* (Colajanni, Yu, & Cardellini, 1998). It uses hidden load weight, which is the number of incoming requests from different client domains, to distribute requests. The DNS server assigns client requests to the available servers in the round-robin manner and associates a client-specific TTL value to each request. The client-specific TTL is set in terms of the hidden load weight of the client domain as well as the server processing capacity. It is known that the hidden load weight increases with the TTL value. The TTL should be set inversely proportional to the hidden load weight in order to amortize the hidden loads from different client domains while being proportional to the server processing capacity.

In summary, the DNS-based load-balancing approaches are simple to implement because they do not require modifications of the existing network switch. Their main drawbacks are as follows:

- Unbalanced load distribution. As we see from Figure 1, the DNS-based approaches allow clients to contact the server directly based on their cached name resolutions. This is prone to unbalanced load distribution between the servers. Although the authoritative DNS can set the TTL value to limit the duration of each resolution in concept, it will not take effect without the cooperation of intermediate DNS servers and the client's local DNS server. In addition, a small value of TTL would lead to a possible bottleneck in the authoritative DNS.
- Low availability. Due to the existence of name resolutions in caches, client requests will continue to be sent to a server, whether the specific server is online or not, before the resolution expires. Consequently, some clients may perceive a service failure, even though there are other servers available for the service.

It is because of these drawbacks that DNS-based approaches became rarely used for load balancing in locally distributed Web clusters. However, they can be tailored to direct requests to geographically distributed servers for global load balancing. Related issues in the context of network-side load balancing will be reviewed in Network-Side Load Balancing.

Dispatcher-Based Load Balancing

In a Web cluster, a dispatcher serves as the cluster gate, as shown in Figure 3, which intercepts all incoming requests. It has a single public IP address representing the Web cluster. Since the Web cluster together with the dispatcher form a virtual server, the IP is often referred to as virtual IP (VIP) address of the virtual server. The dispatcher identifies each server in the cluster with a unique private address, either an IP or a media access address (MAC) address, depending on the architecture. In contrast to the DNS server, the dispatcher has complete control over the routing of all incoming requests and acts as a centralized scheduler. It selects the most appropriate server according to a load-balancing policy and forwards the requests to the selected server.

The choice of the routing mechanism has significant impact on the performance and efficiency of the strategies because different routing implementations will reveal different amounts of information about the client requests at different costs. Two major implementations are the network layer (layer 4) and the application layer (layer 7) with reference to the OSI network stack. Layer-4 implementations are oblivious to the client request content. They simply forward the client requests to a selected server regardless with the requests are connection requests or not. Layer-7 implementations distinguish connection establishment requests from those following up requests in a connection. On the receipt of a client connection request, the dispatcher establishes a TCP communication channel with the client. It then keeps receiving

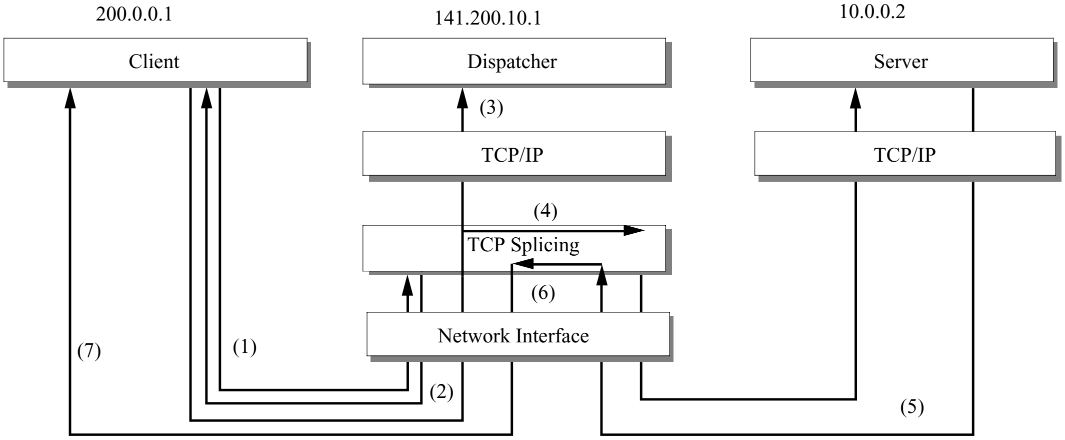


Figure 4: TCP splicing architecture.

requests from the channel and relaying them to the servers selected based on a load-balancing policy.

The dispatcher can be configured into two functional modes according to the way in which the response packets are routed: two-way (two-armed) and one-way (one-armed). The one-way configuration allows servers to return their responses directly to the clients, without going through the dispatcher. In contrast, the two-way configuration requires all the responses to pass through the dispatcher. Therefore, it allows different subnet configuration for the servers. There are many different ways to route packets at the network layer in each configuration.

Two-Way Dispatcher Architecture

Packet rewriting is a layer-4 implementation based on the IP network address translation (NAT) (Egevang & Francis, 1994). In a Web cluster, the dispatcher is set to a different IP address from the servers and configured as a default gateway. On the receipt of a client packet, the dispatcher selects a target server and then rewrites the packet's destination IP address accordingly. The server's response packets need to go through the dispatcher because it acts as the server's gateway. When the dispatcher receives the packets, it replaces their source IP addresses with the virtual IP address, making them look as if they were coming from the virtual server. For example, when a client packet arrives at the dispatcher, its destination address is the VIP, 141.200.10.1. The dispatcher decides the packet should be redirected to a server in the cluster, say Server 1. Then the dispatcher changes the packet's destination address to 10.0.0.2. Server 1's response packets contain 10.0.0.2 as the source addresses. When they reach the dispatcher, their source addresses are replaced with the VIP. Due to the modification of the IP head, the checksum needs to be calculated twice.

TCP splicing is a layer-7 implementation (Cohen, Rangarajan, & Slye, 1999). It inserts a special TCP splicing component in between the network and the MAC layers. When the dispatcher is contacted by a client to establish a TCP connection, it accepts the request instead of rerouting it immediately to a server. The dispatcher then intercepts the client's subsequent requests. After that, the dispatcher selects a destination server and establishes another TCP connection with the server. Once both connections are set up, the TCP splicing component transfers all the requests between the two connections by altering their headers accordingly (source and destination IP addresses, IP and TCP header checksums, and other fields). Therefore, the request and response packets look as if they were being transferred in a single TCP connection. Figure 4 presents the detail of the routing mechanism by assuming a client with IP address 200.0.0.1 wants to access http://www.xyz.com. The client sends a connection request to the Web cluster (step 1). The dispatcher gets this request and establishes a connection with the client (step 2). After getting the client's data request, the dispatcher chooses a server, say Server 1, to handle the request and then splices the two connections (step 3). On the receipt of a client request with a source address of 200.0.0.1 and a destination of 141.200.10.1, the TCP splicing component modifies its source and destination addresses to 200.0.0.1 and 10.0.0.2, respectively (step 4). After modification, the request is sent to the Server 1 directly (step 5). Similarly, when it receives a response from Server 1 with a source of 10.0.0.2 and a destination of 200.0.0.1, the splicing component changes its source and destination to 141.200.10.1 and 200.0.0.1, respectively (step 6). The response is forwarded to the client in a similar way without the intervention of the dispatcher (step 7). Consequently, TCP connection splicing avoids many protocol overheads at high-level network layers.

The dispatcher with two-way configuration is easy to implement by hardware and does not require modifications of existing environment. However, the dispatcher with a two-way configuration tends to be a performance bottleneck because it must process both client inbound requests and outbound responses (Schroeder, Goddard, & Ramamurthy, 2000). This problem becomes even more severe when the responses are bandwidth-demanding, particularly in streaming services.

One-Way Dispatcher Architecture

Packet forwarding is a layer-4 implementation. It assumes that the dispatcher is located in the same LAN as the

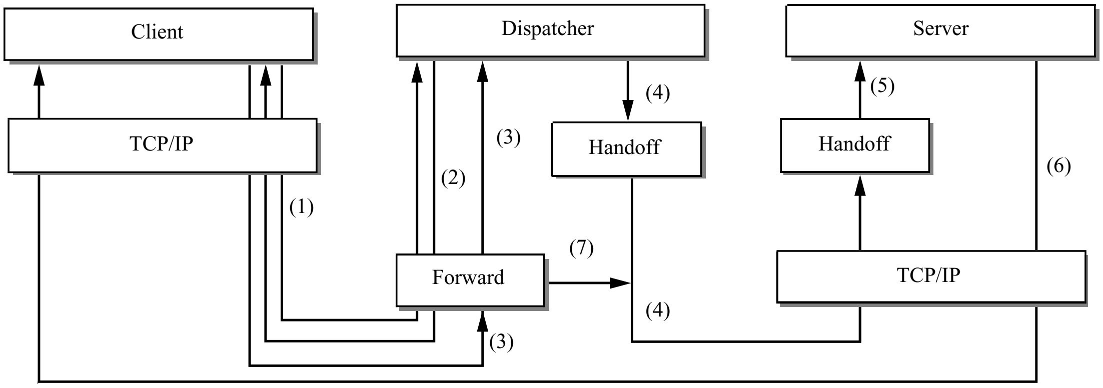


Figure 5: TCP handoff architecture.

servers. They all share one virtual IP address. The dispatcher configures the VIP as its primary address and the servers use the VIP as their secondary address. The servers disable their address resolution protocol (ARP) so that the incoming packets can reach the dispatcher without conflict. On the receipt of a client packet, the dispatcher forwards it to a selected server on the LAN by changing its destination MAC address to that of the server. The receiving server processes the client's packets and sends the response packets to the client directly, without the intervention of the dispatcher. For example, assume there is a client packet with a source address of 200.0.0.1 and a destination of 141.200.10.1 arrives and the dispatcher selects a server, say Server 1, to respond to the packet. The dispatcher changes its destination MAC to 12:34:56:78:90:bb. Since all response packets use the VIP as their source, the service transparency is retained.

Packet tunneling is another layer-4 implementation based on IP encapsulation on the dispatcher and the servers (Simpson & Daydreamer, 1995). IP encapsulation is to wrap each IP datagram within another IP datagram as payload, thus providing a method of redirecting the IP datagram to a destination different from the original one. In this approach, the IP addresses of the dispatcher and the servers should be configured in the same way as in the packet forwarding mechanism. After deciding a server for each incoming datagram, the dispatcher encapsulates the datagram via IP encapsulation; the source and destination of the outer header are the VIP and the selected server IP address, respectively. Then the new datagram is forwarded to the selected server. The target server unwraps the datagram and gets the original datagram. Because the server shares the VIP with the dispatcher, the original IP datagram is delivered to the server through its private (secondary) address. The packet tunneling approach requires IP encapsulation to be supported in the dispatcher and all the servers.

TCP connection hop is a software-based proprietary solution by Resonate (Resonate, 2001). It combines the advantages of IP encapsulation and TCP splicing. Its key component, resonate exchange protocol (RXP) operating between the network driver and TCP/IP stack, needs to be installed in the dispatcher and all the servers. When the RXP receives an IP datagram, it selects a server based on its load-balancing policy and encapsulates the original datagram in an RXP datagram for the target server. The RXP header of the RXP datagram is stripped off by the RXP component in the target server. The original datagram travels up the TCP/IP stack and is processed as if it came directly from the client. Since the server shares VIP with the dispatcher, it replies directly to the client.

TCP handoff is a layer-7 implementation (Pai et al., 1998). Figure 5 depicts the architecture of the TCP handoff approach. The dispatcher and the servers are modified to put the TCP handoff protocol on top of the TCP/IP stack. As an example, we assume that a client wants to access the Internet service. The client sends a connection request to the dispatcher; the request flows up the TCP/IP stack, and arrives at the dispatcher (step 1). The dispatcher accepts this request and establishes a connection with the client (step 2). The client sends a data request to the dispatcher (step 3). The dispatcher receives the data request; makes a server selection decision based on the request content; and hands the connection to the target server using the handoff protocol (step 4). The server takes over the established connection (step 5). The server's responses are returned to the client directly as if it came from the virtual server (step 6). For acknowledgment packets, a forwarding module located between the network interface driver and the TCP/IP stack is used (step 7). The module first checks whether an acknowledgment packet should be forwarded. If so, the packet is modified and sent to the appropriate server without traversing the TCP/IP stack.

Note that both the one-way and two-way implementations of a dispatcher define load-balancing mechanisms and leave load-balancing policies open. In the following, different policies will be reviewed.

State-Blind Load Balancing

As in the DNS-based approaches, both the round-robin and randomized assignment policies can be applied for load balancing in the dispatcher. Albeit simple, they overcome the limitation of the DNS-based implementations. The state-blind policies overlook the client request-specific information and server dynamic states. This makes these approaches not very appealing in practice.

Client State-Aware Load Balancing

State-blind load-balancing approaches treat client requests independently. However, the client's requests are dependent on many Internet services (i.e. e-commerce). That is, these requests must be forwarded to the same server to preserve the dependent information, such as the content of a client's shopping cart, between requests. This required coordination in the process of related requests is called *client affinity*. One approach to keeping client affinity is to use the SSL session ID to ensure it. SSL runs on top of TCP and uses the session ID to uniquely identify each session. Therefore, the dispatcher can achieve client affinity by forwarding the packets with the same session ID to the same server. Some other methods, such as cookie, can also be used to keep client affinity (Kopparapu, 2002).

Server State-Aware Load Balancing

Server-side state information, like dynamic workload in terms of active connections, response time, etc., can be considered to improve the performance of load balancing.

The *LocalDirector* (Cisco Systems, n.d.b) uses a least-connection policy to dispatch incoming client requests to the servers with the least active connections. Weighted least-connection policy takes the number of active connections and the servers' capacities into consideration. It is the same as that used in DNS-based approaches. The *fastest response* policy selects the server with the least response time.

Client and Server State-Aware Load Balancing

While the server state-aware load-balancing approaches consider the state information of servers, it is also necessary to take into account the client state information and combine them together. Except for the source IP address and TCP port number, the dispatcher can get more information from client requests; for example, it can get the URL information and the request content. Moreover, the client characteristic is also useful for load balancing.

Locality-aware request distribution (LARD) (Pai et al., 1998) makes server selection based on the client request content. It considers not only the load balancing but also the cache hit rate. For example, requests are categorized into content types, such as A, B, and C. For requests for content type A, they are dispatched to Server 1, while requests for contents of types B and C are dispatched to Server 2. Hence, the cache system of Server 1 is mainly used by content of type A while the cache system of Server 2 is occupied by contents of types B and C. Consequently, both servers' cache hit rates increase, which in turn improves the system performance in terms of response time. Meanwhile, the dispatcher limits the number of accepted requests. It hence has the flexibility to distribute the load among all servers. The requests for the same type of contents are dispatched to the same server when the server is nonoverloaded. Otherwise, the requests are dispatched to the server with the least active connections.

An extension of LARD that reduces the overhead of the dispatcher is *scalable content-aware request distribution* (Aron, Sanders, & Druschel, 2000). The client requests are redirected to different servers by a simple scheme, such as DNS-based round robin. Their request contents are inspected at individual servers and sent back to the dispatcher. The dispatcher uses the same policy as LARD to determine the most appropriate server.

The size interval task assignment with variable load (SITA-V) (Crovella, Harchol-Balter, & Murta, 1998) is based on that the distribution of the client request sizes is heavy-tailed with infinite variance. SITA-V assigns requests with small sizes to the servers with lighter workload. It improves the ratio of the requests' waiting times to their service times. This can improve the user-perceived performance.

Server Cooperative Load Balancing

In general, cooperative load-balancing approaches deploy a two-level request routing mechanism. Client requests are initially assigned by the authoritative DNS server to target servers. It is followed by a possible redirection from the target server to others for load balancing. That is, based on some load-balancing policies, all servers are allowed to participate in the process of request distribution. These servers can be locally distributed or geographically distributed. Since all servers can decide which is the most appropriate server, there is no centralized point of becoming a potential performance bottleneck. Note that these schemes assume that the client requests can reach different servers through broadcasting or DNS-based round robin. Also, different load-balancing policies can be applied to these implementations.

Request Redirection Mechanisms

The redirection mechanisms used in server cooperative load balancing include IP filtering, HTTP redirection, URL rewriting, packet rewriting, and IP tunneling. The last one is the same as IP tunneling in the dispatcher-based approaches. *IP filtering* only accepts wanted packets. The IP filtering component operates between the network interface driver and TCP/IP stack. An incoming packet is broadcasted to all servers. When it arrives at the IP filtering component, the component decides whether it should accept the packet based on the load-balancing policy. If so, the packet is forwarded up to the TCP/IP stack for further processing.

HTTP redirection makes use of the HTTP redirection provided in HTTP protocols (Fielding et al., 1999). A client sends a request to a server. Then the server can reply to it with another server's address according to its load-balancing policy and set the status code for redirection. By checking this status code, the client sends new requests to the selected server, if needed. For example, a client sends a request to Server 1. The server decides to let Server 2 process this request. It replies to the client with status code 301 ("Moved Permanently") or 302 ("Moved Temporarily") and sets Server 2's address as the new location. The client checks the status code and sends a new request to Server 2.

URL rewriting is based on dynamic URL generations. When a client requests a Web page, the URLs contained in this page are generated dynamically according to the server's load-balancing policy. These URLs point to different servers. Consequently, requests for these URLs are directed to different servers. For example, http://141.200.10.1/document.pdf is a link inside http://www.xyz.com/index.html. When a client accesses this index page, the server can replace it with http:// 10.0.0.2/document.pdf before responding to the client. Thus, if client accesses this link, the request is sent to the new location.

Distributed packet rewriting is similar to packet rewriting (Bestavros, Crovella, Liu, & Martin, 1998). When a packet reaches a server, the server uses a hash function to calculate the packet's destination by the use of the client's IP address and port number to determine to which server the packet should be redirected. Then the server rewrites the packet to redirect it to the selected server. However, for fragmented packets, since only the first fragment contains the required IP and port number, the rest of the packets could be misrouted.

Load-Balancing Approaches

Cooperative load-balancing approaches must have the knowledge about the servers. They are different in what kind of client state information is taken into account in server selection.

Network load balancing (Microsoft, 2000) makes use of IP filtering. When a packet arrives, an IP filtering component uses a randomization function to calculate the server priority based on the client's IP address, port number, and other state information. The server with highest priority accepts the packet and forwards the packet to the upper TCP/IP stack while other servers discard it. It also ensures consistent mapping—same client IP and port will always be mapped to the same server. However, the client's IP addresses and port numbers statistically depend on whether the client is residing behind proxy servers or other NAT devices. When client affinity is enabled, all clients from that domain are processed by one server and load balancing is defeated.

A distributed cooperative Web server (DCWS) uses the URL rewriting to implement its load-balancing policy (Baker & Moon, 1999). Each server maintains the information of all pages in a local document graph. The graph consists of a set of tuples that store the page size, hit rate, and other information. When the client requests a Web page, the hyperlinks inside the page are dynamically generated based on their loads (i.e., hit rates) and the server's load. The main disadvantage of this approach is the calculation overhead of dynamically generated hyperlinks and the reconstruction of related pages.

A scalable Web server (SWEB) uses the HTTP redirection to distribute load across servers (Andersen, Yang, Holmedahl, & Ibarra, 1995). On the receipt of a client request, SWEB selects a server with least response time to direct the request. The calculation of the response time is on the basis of several parameters: redirection latency, data reading time, CPU time, and network delay.

Workload-aware request distribution (WARD) uses a *ward-analysis* algorithm to determine the frequently requested contents, so-called base, that can be fitted into the cluster's memory (Cherkasova & Karlsson, 2001). These contents are further partitioned into two groups, core and part, according to their sizes and frequencies of requests. The core contents include the most frequently accessed objects and is replicated and placed on the cache of all servers. The part contents are divided among servers in a balanced manner using hit rates and object sizes. These partitions are called local and remote partition. On the receipt of a client request, a server checks the request content. If the requested content is not in the base partition, it is served locally. Otherwise, if the requested object belongs to core or local partition, it is served locally. In other cases, it is handed off to an appropriate server. WARD can reduce the request redirection number and improve the mean response time of client requests.

NETWORK-SIDE LOAD BALANCING

For scalability and availability, giant Internet services are often run on more than one geographical site on the Internet. Network-side load balancing is to distribute requests between different sites for balancing the workload of the sites and reducing the service transmission delay on the network. The geographically distributed servers can be organized in different ways. This section reviews two most important structures: caching proxy network and content delivery network , and their related load-balancing strategies.

Client State-Aware Load Balancing: Caching Proxy Network

Web caching takes advantage of the temporal locality in requests for reducing the network load and access latency. Web caching proxies are the intermediate servers between clients and servers that act as agents representing the server to the client and the client to the server. As shown in Figure 6, Web caching proxies, including client-side forward caching proxies, network-side forward and reverse caching proxies, and server-side caching proxies, form a network. Forward caching proxies forward a client's request to the Web cluster, cache the response, and deliver the response to the client. These proxies are normally controlled by the clients or clients' network providers. Reverse caching proxies return requested content either from their caches or after fetching the content from its origin servers. Furthermore, reverse caching proxies are under the same administrative control as the Web cluster. By serving clients' requests from the proxies, the proxy network reduces the client-perceived network transmission latency, saves network bandwidth, and offloads the traffic of origin servers.

The caching proxy network can be organized as a hierarchical or mesh structure. In the hierarchical structure, low-level caching proxies close to the clients forward requests to a high-level proxy closer to the origin server if they cannot satisfy the requests. The solid arrows in Figure 6 represent a chain of proxies, from the lowest client forward proxy to the highest server proxy. The hierarchical proxy structure is an extension of computer memory hierarchy. It is simple to implement, but costly in space because a highly demanded Web object may be stored in many low-level cache proxies. By contrast, the mesh (cooperative) structure of caching proxy servers allows a forward proxy to communicate with its predetermined peers using an intercache protocol like Internet cache protocol (Wessels & Claffy, 1997). If a caching proxy cannot find a valid response in its cache, it searches its peers. If they do,

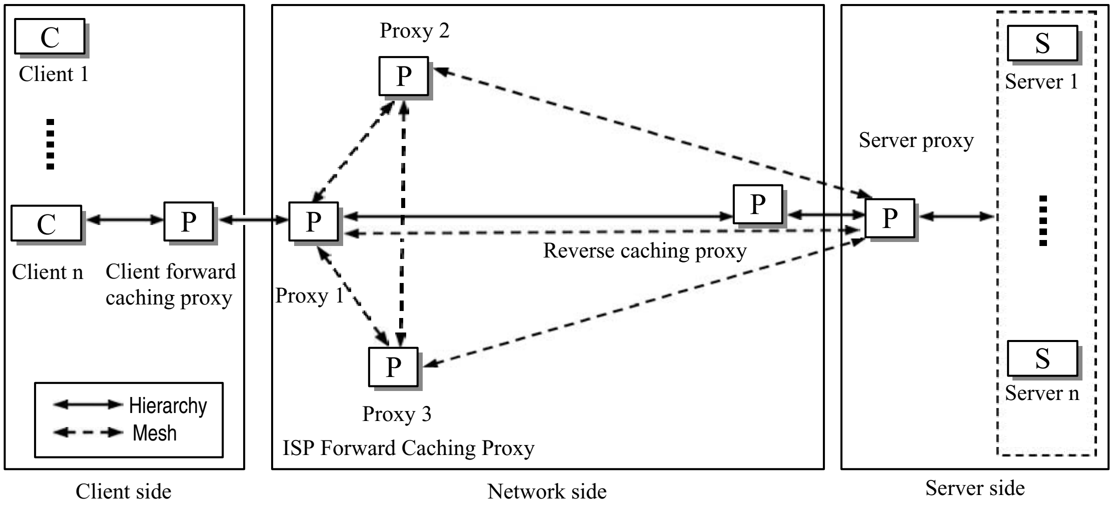


Figure 6: Architecture of a caching proxy network.

the request is directed to the first responding proxy. Otherwise, the proxy contacts the origin server for the response. The mesh structure reduces redundant copies in caching proxies at the cost of intercache communication. The hierarchical and mesh structures can be used in combination. For instance, in Figure 6, Proxy 1 can contact Proxies 2 and 3 when it is unable to satisfy an incoming client request. If neither Proxy 2 nor Proxy 3 has a valid response, the request is forwarded to the origin server through its reverse caching proxies and server-side proxy.

Client and Server State-Aware Load Balancing: Content Delivery Network

CDN is an infrastructure for delivering content from locations closer than original content provider sites to clients by replication. CDN improves client-perceived performance of Internet services by serving client requests from CDN servers. It works similarly to caching proxy servers. However, CDN has some advantages (Rabinovich & Spatscheck, 2001). First, CDN servers can be deployed around the network transparently, which is not always possible for caching proxies. The second is that service providers have full content control, such as content replication. For caching proxy servers, it is hard to control which content should be placed on which proxy servers. Another advantage is that CDN can improve performance for uncachable content, such as dynamic content and live streaming content. Finally, the content of Internet services can be placed on a CDN server before being accessed by clients.

There are two kinds of replication mode for CDN: full replication and partial replication. In full-replication mode, the origin sites are modified to use the authoritative DNS servers provided by CDN companies. The client requests can be delivered to CDN servers or be forwarded to the original sites. Partial replication needs modifications of the hyperlinks inside Web pages. For example, a hyperlink http://www.xyz.com/image.gif from XYZ company's homepage can be changed to http://www.cdn.com/www.xyz.com/image.gif. Therefore, when a client sends a request for this object, the host name needs to be resolved by CDN name servers. The products using such a DNS redirection mechanism include WebSphere from IBM (IBM, 2002) and EdgeSuite from Akamai (Akamai, 2002).

DNS-Based Request Redirection

For network-side load-balancing approaches, there are several mechanisms to redirect client requests to an appropriate geographically distributed server or site, such as DNS redirection, packet tunneling, and NAT peering. As forementioned, DNS cannot effectively control server selection since such resolution can be satisfied by caches or name servers other than authoritative DNS servers. Assigning a small value to TTL may cause the DNS servers to bottleneck and increase client-perceived latency. To alleviate this problem, two-level DNS redirection is proposed (Leighton & Lewin, 2000). On the top level, some DNS servers are used to resolve requests for higher-level domain names such as .com or .org. On the bottom level, the DNS database is replicated or partitioned among a number of geographically distributed DNS servers. When receiving a resolution request, a top-level DNS server responds with the address of a bottom-level DNS server that is close to the client. Therefore, it is possible to set a small TTL value without making the DNS servers overloaded. Another approach is to place fully replicated DNS servers on the edge of the Internet. Resolution requests are directed to the DNS server closest to the client.

Another issue with DNS-based request redirection is an originator problem. It is known that an authoritative DNS server receives a request from a client's local DNS server instead of the client directly. However, sometimes, the location of the client is quite different from the location of the client's local DNS server. For example, in AT&T WorldNet, most clients use only two DNS servers no matter where they are. Therefore, the closest server for a local DNS server sometimes is a distant one for the client compared to other servers.

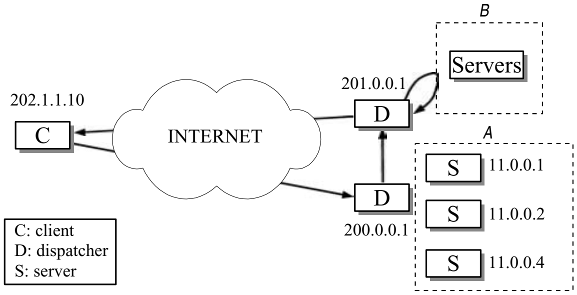


Figure 7: Network address translation and NAT peering.

NAT Peering

NAT peering is an implementation that can be used to redirect a client request from one site to another, avoiding the limitations of DNS-based redirection. The purpose of NAT is to provide a solution for IP address depletion. It provides a way to translate public IP addresses and TCP ports to private IP addresses and TCP ports. The main advantage of this approach is that it can be installed without modifications of hosts or routers.

Figure 7 presents an illustration of NAT peering for request redirection between different sites or servers. As an example, a client *c* issues a request to an Internet service on the public IP address 200.0.0.1. The NAT device in local area network *A* receives the request and selects an appropriate site or server based on its load-balancing policy. Assuming *B* is selected, the request's destination address is changed accordingly while its source IP address is preserved. Therefore, the response is sent directly back to client *c* while its source address is changed to *A*'s IP address 200.0.0.1. Consequently, all client requests are still sent to its original destination *A*. NAT peering is widely supported by such switches as Cisco CSS11000 (Cisco System, 2002).

Case Studies

Akamai CDN. By using more than 12,000 servers in over 1,000 networks, Akamai's distributed content delivery system fights service bottlenecks and shut-downs by delivering content from the Internet's edge (Dilley et al., 2002). To handle traffic bursts, Akamai's infrastructure directs client requests to appropriate servers as well as additional servers. When receiving a client's resolution request, the Akamai CDN name server determines the target server using both client- and server-side information. Specifically, to determine the target server, it takes many metrics into account, such as network topology, bandwidth, and server's load. Moreover, the target server should have the requested content. For example, a media stream request should not be directed to a server that handles only HTTP. To prevent a server from overloading, a CDN's load-balancing system must monitor the state of services, and their servers and networks. Akamai uses a threshold method for this purpose. All sites periodically report their service states and site load information to a monitoring application. When the load of a site exceeds a predefined threshold, more servers are allocated to process client requests. If the load exceeds another higher threshold, the site is excluded from the available site list until it is back to the threshold.

Enterprise CDN and Examples. Large-scale Internet services deploy and administer their own enterprise CDNs. They use multiple levels of load balancing to distribute client requests for a single giant Internet service. At the top level, the services distribute their servers geographically to offer higher availability and better performance. At this level, services can adopt various request redirection mechanisms to distribute client requests to an appropriate site. The site selection can take into account the site's load and network's traffic information. After receiving a client request, the front-end, which can be a dispatcher, selects an appropriate server from local servers. At this level, all server-side load-balancing approaches can be deployed.

The Web site of the 1998 Olympic Winter Games is a good example of enterprise CDNs that IBM developed (Iyengar, Challenger, Dias, & Dantzig. 2000). It supported nearly 635 million requests over the 16 days of the game and a peak rate of 114,414 hits in a single minute around the time of the women's freestyle figure skating. For scalability and availability, Web servers were placed in four different global wide sites. Client requests to http://www.nagano.olympic.org were routed from their local ISP into the IBM global services network (IGS), and IGS routers forwarded the requests to the server site geographically nearest to the client. Each site had an installation of one SP-2 machine configured with 10 R6000 uniprocessors and an 8-way symmetric multiprocessor and had 12 open IP addresses. Load balancing at each site was conducted in a hierarchical way. At the top level, two routers acted as the entry points of each site. One is the primary point and another one is a backup. Between the routers and the Web servers were four IBM network dispatchers (NDs) for load balancing. Each of these four NDs was the primary source of 3 of the 12 IP addresses and the secondary source for two other addresses. These NDs were assigned to different values based on whether they were the primary or secondary source for an IP address. Secondary addresses for a ND were assigned a higher value. Therefore, the routers delivered incoming requests to the ND with the lowest value. On the receipt of the request, the ND distributed it to a server based on the servers' load.

Three more examples of large-scale representative Internet services on their own CDNs were recently surveyed (Oppenheimer & Patterson, 2002). To keep the services' identities confidential, the authors used the names of "Online," "Content," and "ReadMostly" to refer to an online service/Internet portal, a global content-hosting service, and a high- traffic Internet service with a very high read-to-write ratio, respectively. At the survey time, both the Online and ReadMostly services supported around a hundred million hits per day and the Content service supported around 7 million hits per day. The Online service used around 500 servers at two data centers, the Content 500 servers at four data centers, and the ReadMostly more than 2000 servers at four data centers. They generally use policies that take into account servers' load and availability to distribute client requests. The Online service provided its clients with an up-to-date list of appropriate servers for their selection. In the Content service, each two of its four sites worked in redundant pairs. Its clients need to install a software package that points to one primary and one backup site of the content service. The Read-Mostly service used DNS-based load-balancing strategies based on its sites' load and health information. Among locally distributed servers, round-robin DNS servers or layer-4 dispatchers were used to direct client requests to the least loaded server.

Finally, we note that CDN was originally proposed for fast delivery of media-rich contents. For CDN to deliver live streaming media, there are some challenges because the media cannot be stored in advance on CDN servers. IP multicast or application-level multicast are used to distribute streaming media to CDN servers. For IP multicast, a multicast group is created by the origin server before the stream is distributed. A CDN server joins or leaves the group when it receives the first client request or finishes the last client request. Application-level multicast is based on a distribution tree of intermediate servers that makes use of IP unicast to deliver stream from the root (origin server) to the leaves (CDN servers).

CLIENT-SIDE LOAD BALANCING

Client-side load balancing mainly refers to load distribution between clients and servers. In addition, some services running on several servers are not transparent to their clients and the server selection is left to clients. We classify this type of client-guided load distribution as client-side load-balancing strategies, as well. In concept, client-side approaches can get as much client state information as they want. Thus, we classify these approaches in terms of whether they take server state information into account.

Server State-Blind Load Balancing

Client-side caches, like browsers' caches, can be constructed as an efficient and scalable peer-to-peer Web cache system for reducing servers' load by delivering requested content from this system (Xiao, Zhang, & Xu. 2002). This approach assumes that the communication within a local area network is much faster than that between the clients and origin servers. These caches are managed by middleware components, which function as proxies for client requests, installed in all clients. When a client wants to access an Internet service, it first issues a request to the component within the same host. If the requested content is not cacheable, the component forwards the request to the origin servers. Otherwise, the component communicates with components running on other hosts to search the requested content. If a valid content is not present there, the request is sent to the external servers and the content is stored in the clients' caches for future use.

Server State-Aware Load Balancing

Some software distribution sites use another approach to balance the load of their servers. Before letting a client download their software, they provide several available servers with some location information. Meanwhile, they also provide the servers' URLs, geographical locations, and domain information. It is up to the clients to decide where to download the software. As an example, from the homepage of Free Software Foundation, we can find a list of available mirror servers that includes their location information. A client can select the geographically closest server to reduce the network latency. An extension of this approach is *dynamic server selection* (Crovella & Carter, 1995). In this approach, the round-trip time is measured periodically. Based on the mean of four round-trip measurements, the server with the least average round-trip time is selected to send the client's requests.

Smart client determines an appropriate server on the basis of server-side state (Yoshikawa et al., 1997). When a client visits a site, it downloads an applet from the site. The applet retrieves some server state information, such as its workload, response time, network connectivity, and configurations. As a result, the applet selects a server based on the state information. One disadvantage of this approach is that the information exchanging between the servers and the applet increases the network traffic.

CONCLUSION

Load balancing on the Internet is to distribute the workload generated by Internet services between the server, client, and network aspects for the objective of providing high-level quality of service. This chapter presents a classification of load-balancing strategies in three dimensions: where they are realized (client side, server side, or network side), what knowledge the strategies take into account in making decisions, and on which network layer they are implemented. Under the framework, state-of-the-art load-balancing strategies in each category are reviewed, focusing on their mechanisms, polices, and interplay with existing network routing and service delivery infrastructure.

ACKNOWLEDGMENT

This material is based on work supported in part by the National Science Foundation under Grants CCR-9988266 and ACI-0203592. Any opinions, findings, and conclusions or recommendation expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

GLOSSARY

Caching proxy network A group of caching proxies that collaborate to reduce the network traffic and the origin server load by exploiting the temporal locality of client requests.

Client load balancing A kind of technology performed with the cooperation of clients and servers to distribute client request load of an Internet service between clients and servers.

Client-aware load balancing The distribution of client requests to the servers of a Web cluster based on information about the clients.

Content delivery network A network of geographically distributed servers delegated to deliver content for origin servers.

Content-aware load balancing The distribution of client requests to the servers of a Web cluster based on the requested content (see content-blind load balancing).

Content-blind load balancing The distribution of client requests to the servers of a Web cluster without considering any information about the requested content (see content-aware load balancing).

DNS-based load balancing The reliance on DNS servers to distribute client requests between locally and globally distributed servers during the process of domain name resolutions of requests.

Layer-4 dispatcher A device that distributes the client requests at the TCP layer when it is contacted for establishing a TCP connection.

Layer-7 dispatcher A device that distributes the requests at the application layer by establishing a TCP connection with a client per request and delaying the request distribution until the request content is examined.

Least loaded distribution A load-balancing policy that assigns client requests to the server in a Web cluster with least load, which can be measured as a function of active connections, CPU utilization, etc.

Load balancing dispatcher A switch in front of a Web cluster that receives the client requests and directs them to appropriate servers; also referred to as load switch or load balancer.

Network load balancing The distribution of client requests of an Internet service between servers across wide area networks to reduce the service access latency and increase the service availability.

One-way dispatcher A device that intercepts all client requests of a Web cluster and allows the cluster servers to respond to the requests directly.

Round-robin load distribution The assignment of client requests to the servers of a Web cluster in a cyclic manner whereby each server appears one or more times in a round.

Server cooperative load balancing A two-level request routing mechanism that allows an assigned client request to be redirected to another server.

Server load balancing The distribution of client requests of a Web cluster between the cluster servers so as to reduce access latency and increase service scalability and availability.

Two-way dispatcher A device that intercepts both client requests and server responses.

Virtual IP The only IP address of a Web cluster visible to clients; in practice, often the public IP address of the load-balancing dispatcher.

Web cluster A group of servers that work collectively to provide the same Internet services at a single location.

CROSS REFERENCES

See *Web Quality of Service; Web Services.*

REFERENCES

Akamai (2002). *EdgeSuite*. Retrieved October 5, 2002, from http://www.akamai.com/en/html/services/edgesuite.html

Andersen, D., Yang, T., Holmedahl, V., & Ibarra, O. H. (1995). SWEB: Towards a scalable World Wide Web server on multicomputers. In *Proceedings of 10th IEEE International Parallel Processing Symposium* (pp. 850–856). Los Alamitos, CA: IEEE Computer Society Press.

Arlitt, M. F., Krishnamurthy, D., & Rolia, J. (2001). Characterizing the scalability of a large Web-based shopping system. *ACM Transaction on Internet Technology, 1*(1), 44–69.

Arlitt, M. F., & Williamson, C. L. (1997). Internet web servers: Workload characterization and performance implications. *IEEE/ACM Transaction on Networking, 5*(5), 631–645.

Aron, M., Sanders, D., & Druschel, P. (2000). Scalable content-aware request distribution in cluster-based network servers. In *Proceedings of 2000 USENIX Annual Technical Conference* (pp. 323–336). Berkeley, CA: USENIX Association.

Baker, S. M., & Moon, B. (1999). Distributed cooperative Web servers. In *Proceedings of the 8th International World Wide Web Conference* (pp. 1215–1229). New York: Elsevier Science.

Bestavros, A., Crovella, M., Liu, J., & Martin, D. (1998). Distributed packet rewriting and its application to scalable server architectures. In *Proceedings of 6th International Conference on Network Protocols.* Los Alamitos, CA: IEEE Computer Society Press.

Cherkasova, L., & Karlsson, M. (2001). Scalable web server cluster design with workload-aware equest distribution strategy WARD. In *Proceedings of the 3rd International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems* (pp. 212–221). Los Alamitos, CA: IEEE Computer Society Press.

Chesire, M., Wolman, A., Voelker, G. M., & Levy, H. M. (2001). Measurement and analysis of a streaming-media workload. In *Proceedings of the 3rd USENIX Symposium on Internet Technologies and Systems (USITS)* (pp. 1–12). Berkeley, CA: USENIX Association.

Cisco Systems. (2002). *Cisco CSS11000.* Retrieved October 5, 2002, from http://www.cisco.com/warp/public/cc/pd/si/11000/

Cisco System. (n.d.b). *Cisco LocalDirector 400 Series.* Retrieved October 5, 2002, from http://www.cisco.com/warp/public/cc/pd/cxsr/400/index.shtml

Cohen, A., Rangarajan, S., & Slye, H. (1999). On the performance of TCP-splicing for URL-aware redirction. In *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems (USITS).* Berkeley, CA: USENIX Association.

Colajanni, M., Yu, P. S., & Cardellini, V. (1998). Dynamic load balancing in geographically distributed heterogeneous web servers. In *Proceedings of the 18th IEEE International Conference on Distributed Computing Systems* (pp. 295–302). Los Alamitos, CA: IEEE Computer Society Press.

Crovella, M. E., Harchol-Balter, M., & Murta, C. D. (1998). Task assignment in a distributed system: Improving performance by unbalancing load. In *Proceedings of ACM Sigmetrics Conference* (pp. 268–269). New York: ACM Press.

Crovella, M. E., & Carter, R. L. (1995). Dynamic server selection in the Internet. In *Proceedings of the 3rd IEEE workshop on the Architecture and Implementation of High Performance Communication Subsystems (HPCS'95)* (pp. 158–162). Los Alamitos, CA: IEEE Computer Society Press.

Dilley, J., Maggs, B., Parikh, J., Prokop, H., Sitaraman, R., & Weihl, B. (2002). Globally distributed content delivery. *IEEE Internet Computing, 6*(5), 50–58.

Egevang, K., & Francis, P. (1994). *The IP network address translator (NAT)* (RFC 1631). Network Working Group. Reston, VA: The Internet Society.

F5 Networks (n.d.). *3-DNS Controller: Make the most of your network*. Retrieved October 5, 2002, from http://www.f5.com/f5products/3dns/

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., et al. (1999). *Hypertext transfer protocol—HTTP/1.1* (RFC 2616). Network Working Group. Reston, VA: The Internet Society.

Harchol-Balter, M., Crovella, M. E., & Murta, C. D. (1999). On choosing a task assignment policy for a distributed server system. *International Journal of Parallel and Distributed Computing, 5*(2), 204–228.

Hennessy, J. L., & Patterson, D. A. (2003). *Computer architecture: A quantitative approach* (3rd ed.). San Mateo, CA: Morgan Kaufmann.

IBM (2002). *IBM WebSphere software platform*. Retrieved October 5, 2002, from http://www.ibm.com/websphere

Iyengar, A., Challenger, J., Dias, D., & Dantzig, P. (2000). High-performance Web site design techniques. *IEEE Internet Computing, 4*(2), 17–26.

Kopparapu, C. (2002). *Load balancing servers, firewalls, and caches*. New York: Wiley.

Kwan, T. T., McGrath, R. E., & Reed, D. A. (1995). NCSA's World Wide Web server: Design and performance. *IEEE Computer, 28*(11), 68–74.

Leighton, T., & Lewin, D. (2000). *Global document hosting system utilizing embedded content distributed ghost servers* (International Publication WO 00/04458). World Intellectual Property Organization.

Menasce, D. A., Almeida, V. A. F., Fonseca, R., & Mendes, M. A. (1999). A methodology for workload characterization of e-commerce sites. In *Proceedings of the ACM Conference on Electronic Commerce* (pp. 119–128). New York: ACM Press.

Microsoft (2000). *Network load balancing technical overview* (white paper). Retrieved October 5, 2002, from http://www.microsoft.com/windows2000/techinfo/howitworks/cluster/nlb.asp

Oppenheimer, D., & Patterson, D. A. (2002). Architecture and dependability of large-scale Internet services. *IEEE Internet Computing, 6*(5), 41–49.

Padmanabhan, V. N., & Qui, L. (2000). The content and access dynamics of a busy Web site: Findings and implications. In *Proceedings of the ACM SIGCOMM* (pp. 111–123). New York: ACM Press.

Pai, V. S., Aron, M., Banga, G., Svendsen, M., Druschel, P., Zwaenepoel, W., et al. (1998). Locality-aware request distribution in cluster-based network servers. In *Proceedings of the 8th ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 205–216). New York: ACM Press.

Rabinovich, M., & Spatscheck, O. (2001). *Web caching and replication*. Reading, MA: Addison-Wesley.

Resonate. (2001, April). *TCP connection hop* (white paper). Retrieved October 5, 2002, from http://www.resonate.com/solutions/literature/iwp_cd_tcp_connect_hop.php

Schroeder, T., Goddard, S., & Ramamurthy, B. (2000). Scalable Web server clustering technologies. *IEEE Network, 14*(3), 38–45.

Simpson, W., & Daydreamer. (1995). *IP in IP tunneling* (RFC 1853). Network Working Group. Reston, VA: The Internet Society.

Wessels, D., & Claffy, K. (1997). *Internet cache protocol (ICP), version 2* (RFC 2186). Network Working Group. Reston, VA: The Internet Society.

Wolf, J. L., & Yu, P. S. (2001). On balancing the load in clustered Web farm. *ACM Transaction on Internet Technology, 1*(2), 231–261.

Xiao, L., Zhang, X., & Xu, Z. (2002). On reliable and scalable peer-to-peer Web document sharing. In *Proceedings of 2002 IEEE International Parallel and Distributed Processing Symposium*. Los Alamitos, CA: IEEE Computer Society Press.

Xu, C.-Z., & Lau, F. (1997). *Load balancing in parallel computers: Theory and practice*. Dordrecht: Kluwer Academic.

Yoshikawa, C., Chun, B., Eastham, P., Vahdat, A., Anderson, T., & Culler, D. (1997). Using smart clients to build scalable services. In *Proceedings of the USENIX Annual Technical Conference* (pp. 105–117). Berkeley, CA: USENIX Association.

Local Area Networks

Wayne C. Summers, *Columbus State University*

INTRODUCTION TO LOCAL AREA NETWORKS

A network is a collection of two or more devices linked together. Typically the connection is a physical connection using wires or cables, although wireless connections for networks are also possible. In addition to the hardware required for the connection, communication software is necessary to allow the communications to occur. Networks facilitate the sharing of resources including hardware, software, and data as well as enhancing communications between computers and users of computers.

Networks can be classified as local area networks (LANs) and wide area networks (WANs). The main distinction between these classifications of networks is the radius of the network. A local area network is a network where the computers are physically close together. This may mean that the computers are in the same room, in the same building, or even at the same site. Computers in a wide area network are often distributed beyond metropolitan areas. The Internet is an example of a WAN.

Why Do We Want to Network Computers?

In the early days of computing, there were a small number of computers, which could be used by only one person at a time. With the emergence of time-sharing in the 1960s, individual computers could be used by more than one user simultaneously. This significantly expanded the functionality of computers, but had several limitations. Chief among the limitations was that as more users connected to the shared computer, the amount of resources available for each user's transaction became less. In the late 1970s and early 1980s, the personal computer (PC) resulted in the return of one computer—one user (Figure 1). Finally, in the 1990s, hardware and software became available to network multiple PCs (Figure 2). Before LANs, copies of data needed to be kept by every user of the data on each computer; copies of software or application programs used by each user had to be installed on each computer; every computer needed its own printer. Networking computers removes this need for redundancy.

Data in a networked environment can be shared. Each user can access the data from other computers via the network. This feature of networks helped speed the transition from mainframe computing to networked computing. Networked computers allow important information to be shared among different computer users. Rather than copies of data being kept on each computer, one copy of the data can be kept on a server and accessed remotely via the network. Changes to the data can be made once and then accessed by all.

Rather than software being installed on every computer, it can be shared in a network environment. Application programs can be stored on one computer and run remotely from another computer. In an office configured with multiple non-networked computers, each computer must have a copy installed of each application program that is used. In addition to the need to purchase copies of the software for each computer, the software must be installed and maintained on each computer. Network versions of many application programs can be purchased. A networked version of software is typically much cheaper than purchasing large numbers of copies of a particular piece of software. Network software only needs to be installed once on a server, allowing users on the other computers to access the software. When it is time to upgrade the software, it only needs to be done once on the server, instead of on all of the computers. Installing software on multiple computers simultaneously can be facilitated using a computer network.

Networks facilitate the sharing of hardware. Hardware can be installed in one location and accessed

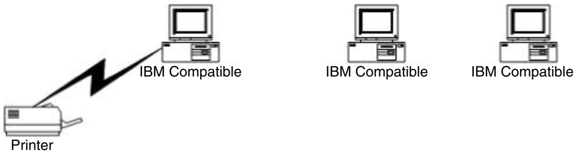


Figure 1: Before networks.

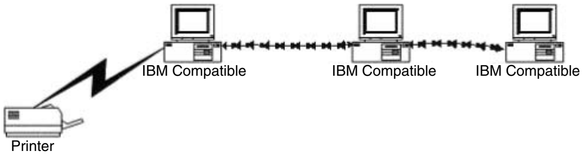


Figure 3: Print server.

over the network from other computers. Printers can be networked so that multiple users can share the same printer or other hardware device. Other peripheral devices including modems, CD and DVD ROMs, and networking devices such as routers can all be shared using a LAN.

Before LANs, computer users who needed to communicate with others had to use traditional methods such as physically visiting another user, calling on the telephone, or having a letter delivered to the other person. Communication has been enhanced tremendously with e-mail and instant messaging. Both of these methods require the use of a LAN.

TYPES OF LANs

Computers that can access a network are often referred to as computer workstations. Any device (workstation, printer, modem, etc.) that connects to a network is called a node.

Many of the early networks allowed simply the sharing of resources between PCs. These types of networks are called peer-to-peer networks. Each computer has the same potential for sharing files and hardware devices. A peer-to-peer network is easy to design and maintain, but is limited in its capabilities.

Most networks today are classified as client/server networks. In a client/server network, one or more of the computers function as servers, while the remainder of the computers functions as clients. A server is a computer that provides a service while a client computer makes use of the service provided. Examples of servers include print servers, file servers, mail servers, and web servers. A print server (Figure 3) is a computer that provides access to one or more printers across the network. Print servers were among the earliest types of servers. A file server provides a repository for files that can be accessed by other computers over the network. Mail or communication servers manage the flow of incoming and outgoing electronic mail for users accessing the server from client workstations. Web servers are computers running software that provides access to World Wide Web documents.

Computers on a network can run more than one type of server software and can function as multiple types of servers. For example, a computer can have both Web server and e-mail server software installed and function as both a Web server and a communications server. A workstation can also be one type of server and a client for another type of server. For example, a computer can be running Web server software but print through the network using another computer that functions as a print server.

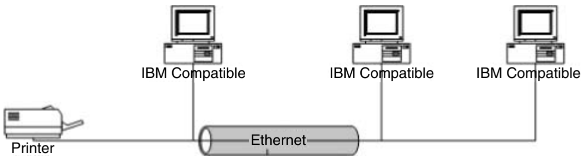


Figure 2: After networks.

Difference Between LANs and WANs

As mentioned earlier, the main distinction between LANs and WANs is the radius of the network. A LAN is a network where the nodes are physically close together. In most cases, the nodes are in the same room, although they may be in the same building or in nearby buildings. Typically, networks with a radius greater than a kilometer or two are classified as wide area networks. Other ways of distinguishing between LANs and WANs include transmission speed and ownership. LANs are typically faster networks with speeds of at least 10 Mbps. WANs are generally significantly slower, with most connections to WANs around 1.5 Mbps. LANs are owned by the organization where the network is used. WANs generally use hardware that is owned by a network provider. A final distinction is with the difference in protocols used by LANs and WANs. The next section describes two of the protocols (Ethernet and token ring) used by LANs. WANs typically use different protocols, including frame relay and X.25.

LAN Topology

LANs can be organized in a variety of ways. One way to classify networks is by their electrical configuration or *logical topology*. This is basically the way that the data are transmitted between nodes. The main two logical topologies are *bus* and *ring*.

In a bus network, the data are broadcast from one node to all other nodes in the LAN even though the data may be intended for only one node. Each of the nodes receives the data, but they are only "read" by the node that the data are intended for. The data include an address for the destination node or nodes. Ethernet is the primary protocol that supports the bus logical topology.

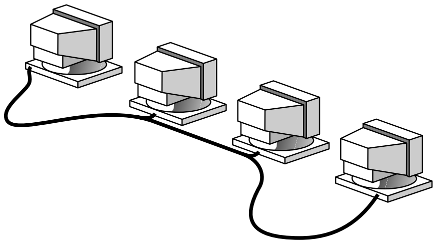


Figure 4: Bus network.

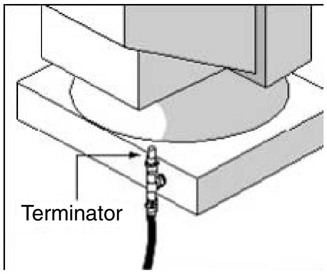


Figure 5: Terminator and BNC T-connector.

In a ring network, the data are sent from one node to the next in sequential order in a circular fashion. Each node inspects the destination address of the data packet to determine if the data are meant for it. If the data are not meant for the node, the data packet is passed along to the next node in the logical ring.

LANs can also be classified by the physical layout of the network. The physical layout is known as the *physical topology*. The physical topology of the network can have a significant influence on a LAN's performance and reliability. The three main physical topologies are *bus, ring,* and *star*. There are also hybrid networks including star–bus and star–ring, which incorporate parts of both types of networks.

In a bus topology (Figure 4), the nodes are arranged in a linear fashion, with terminators (Figure 5) on each end. The nodes are connected to the "bus" with connectors. Bus networks are easy to install but not very reliable. Any break in the connection, or a loose connection, will bring down the entire network.

In a ring topology (Figure 6), each connected node is an active participant in the ring network. Each data packet is received by a node, and, if it is not intended for the node, it is passed along the ring to the next node. If one of the nodes or its network card malfunctions, the network stops functioning.

In a star network (Figure 7) each connected node is attached to a central device. Typically this device is a hub or a switched hub, but it could also be other devices, including a multistation access unit (MAU). Star networks are more reliable and easier to troubleshoot. Star networks do require an additional hardware device such as a hub or switch and additional cable. Because each node is independently connected to the central device, a failure only affects the single node. Of course if the central device fails, the entire network fails.

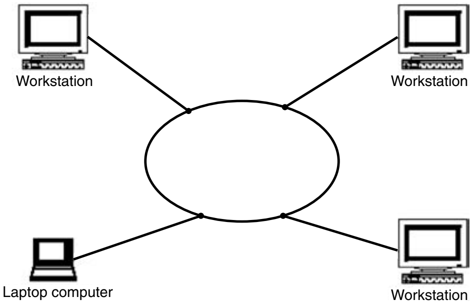


Figure 6: Ring network.

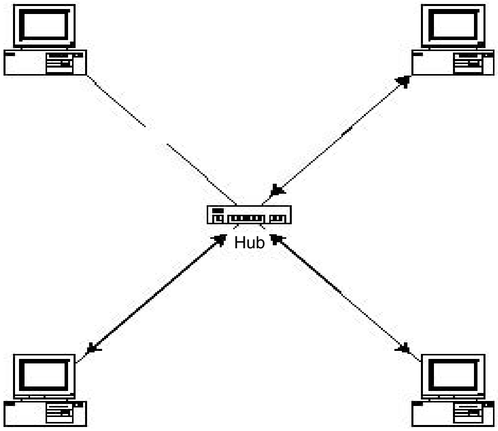


Figure 7: Star physical network.

LAN Architecture

The most popular network architecture for LANs today is Ethernet. Ethernet was invented by Robert Metcalfe and others at the Palo Alto Research Center (PARC) around 1973. Ethernet uses the carrier sense multiple access with collision detection (CSMA/CD) access method. Carrier sense refers to each node being able to "listen" for other users using the network, only attempting to use the network if it is not being used. Multiple access means that any node on the network may use the network without requiring further permission. Collision detection lets the node know if a message was not delivered and controls the mechanism for retransmitting the data packet. CSMA/CD is most efficient when there are a limited number of nodes requesting access to the network.

In 1981, the first Ethernet standard was developed by a consortium comprising Digital, Intel, and Xerox. This was followed by a second Ethernet standard in 1982, called Ethernet II. Ethernet II (Figure 8) had the following characteristics:

Bus topology,
Coaxial cable using baseband signaling,
Data rate10 mbit/sec,
Maximum station separation of 2.8 kilometers, and
Maximum number of stations 1024.

In addition, the IEEE has developed a standard, also often referred to as Ethernet, called the IEEE 802.3 standard (2002; see Figure 9). The two standards are very similar and have similar frame layouts.

Ethernet can run over a variety of media types including several types of coaxial cable, twisted pair cable,

Preamble	Destination Address	Source Address	Type	Data Unit 46- 1500 bytes	Frame Check Sequence
8 octets	6 octets	6 octets	2 octets		4 octets

Figure 8: Ethernet II frame layout.

Preamble	Start Frame Delimiter	Destination Address	Source Address	Length	Logical Link Control IEEE 802.2 Data	Frame Check Sequence
8 octets	1 octet	6 octets	6 octets	2 octets	46- 1500 bytes	4 octets

Figure 9: IEEE 802.3 frame layout.

and fiber optic cable, as well as wireless formats, including radio signals and infrared. Table 1 lists several of these media types. The first number indicates the speed in megabits, the “Base” refers to *baseband transmission,* meaning that the entire bandwidth is dedicated to just one data channel, and the last number or letter indicates the approximate maximum segment length or the media type.

A second network architecture, token ring (Figure 10), was developed in the early 1970s by IBM. Token ring uses the token passing access method. Only the computer that has the 24-bit packet of data called the token may use the network. This token is generated by a designated computer called the active monitor and passed around the ring until one of the computers wishes to use the network. When a computer wants to use the network, it seizes the token, changes the status of the token to busy, inserts its data frame onto the network, and only releases the token when it receives a confirmation that the data packet has been received. A token ring network uses a sequential logical topology, which was traditionally a ring physical topology but now is typically a star topology. IBM specified two architectures that operated at 4 and 16 Mbps.

LAN HARDWARE AND MEDIA

There are a variety of media choices for connecting computers to a local area network. Early networks used copper wires, either coaxial or twisted-pair.

Copper Wire

Coaxial cable consists of a center wire surrounded by insulation and then a grounded shield of braided wire. The shield minimizes electrical and radio-frequency interference. Coaxial cable was typically either *thinnet* or *thicknet*. Thicknet (Figure 11) was the original standard for Ethernet, as defined by the IEEE 10Base-5 standard, and uses 50-Ω coaxial cable (RG-8 or RG-11 A/U) with maximum length 500 m. Thinnet (Figure 12) is defined by the IEEE 10Base-2 standard and uses 50-Ω coaxial cable (RG-58 A/U) with maximum length 185 m. RG-58 is similar to the coaxial cable used with cable TVs. Cables in the 10Base-2 system connect to other devices with BNC connectors (Figure 13).

Twisted-pair networking cable also has two different forms—*UTP* (unshielded twisted-pair) and *STP* (shielded twisted-pair). Both types of cable consist of either two or four pairs of wires. Each pair are twisted together. Shielded twisted-pair cable has an additional layer of conducting material surrounding the twisted pairs of wires. Unshielded twisted-pair cable does not have the additional layer. Telephone companies use UTP cable with two twisted pairs of wires. UTP is the most common and least expensive method for networking computers (Figure 14). There are six categories of unshielded twisted-pair cabling ranging from *Category 1* (CAT1), which is ordinary telephone cable used to carry voice, to *Category 6* (CAT6) (Figure 15), which is designed for high-speed networks. CAT6 uses 23 AWG copper as opposed to the 24 AWG used in Cat5 and lower; therefore the signal attenuates less with speed and distance. CAT6 also uses a tighter twist ratio that cuts down on internal crosstalk. UTP uses RJ45 connectors (Figure 16) to plug into the different networking devices.

Table 1 Types of Network Media

Standard	Popular name	Speed	Media	Maximum Segment Length
10Base2	Thinnet cheapnet	10Mbps	Thin coaxial cable RG-58	185 meters
10Base5	Thicknet yellow hose	10Mbps	Thick coaxial cable RG-8 or RG-11	500 meters
10BaseT	10BaseT twisted pair Ethernet UTP	10Mbps	Unshielded twisted pair CAT3, CAT5	100 meters
100BaseT4	Fast Ethernet	100Mbps	4 pair telephone grade cable	100 meters
100BaseTX	Fast Ethernet	100Mbps	2 pair data grade cable	100 meters
10BaseFL	Fiber Ethernet FOIRL	10Mbps	Multimode fiber optic cable	1000 meters
100BaseFX	Fast Ethernet	100Mbps	2 strands fiber cable	412 meters
1000BaseT	Gigabit Ethernet	1000Mbps	Cat5e	100 meters
10GBase	10 Gigabit Ethernet	10000Mbps	Fiber	300 meters

Starting delimiter 1 octet	Access Control 1 octet	Frame control 1 octet	Destination address 6 octets	Source address 6 octets	Optional Routing Information Field upto 18 octets	Optional LLC Fields 3 or 4 octets	DATA Unlimited size	Frame Check Sequence 4 octets	Ending delimiter 1 octet	Frame status 1 octet

Figure 10: IEEE 802.5 token frame layout.

Fiber Wire

Fiber-optic cable (Figure 17) is becoming more common as demand for higher transmission speeds increases. Fiber-optic cable transmits data using pulsating laser light instead of electricity. Fiber-optic cable consists of a thin glass or plastic filament protected by thick plastic padding and an external plastic sheath. A light signal travels faster, farther, more reliably, and more securely than electricity. Fiber cable can send reliable signals at speeds of 100 GB/s as far as 10 kilometers. Unfortunately, fiber-optic cable is expensive to buy, install, and maintain.

Wireless

Wireless LANs are rapidly becoming commonplace in businesses and homes. Early wireless LANs were light-based, using infrared light to transmit the data. Wireless LANs that are light-based require line of sight for all devices on the network. Because of this limitation and the slow data speed, there are few light-based infrared wireless LANs.

Most wireless LANs use radio waves to transmit the data. Each device in a wireless network requires an antenna to receive and transmit the radio signals. Wireless LANs can be peer-to-peer (Figure 18), which only requires that each device be equipped with a wireless network card that contains the antenna, or a more complete network that requires an *access point* (Figure 19). The access point contains an antenna, a radio transmitter, and a wired network interface, typically an RJ45 port. The access point acts as a base station (similar to a hub) for the wireless network and also as a bridge between the wireless and wired networks.

Regardless of whether the network is wired or wireless, every device on the network must be connected to the network with a network adapter or network interface card (NIC) (Figures 20–22). The card must be physically connected to the device, either installed directly into a slot in the computer or connected via a port like a USB port. The network card provides the interface between the node on the network and the network media.

Hardware Devices

Several factors limit the radius of a local area network. The farther a signal travels along a twisted pair or coaxial cable, the more likely it is to be degraded by noise. As the signal travels, it loses energy and becomes weaker, thus becoming difficult to read. As the network cable becomes longer, it becomes more likely that two or more machines will transmit at the same time, causing a "collision." It will take longer for the machines to detect the collision. There are several ways to increase the radius of a network.

The simplest device to use is a repeater. A repeater (Figure 23) is an electronic device used to extend the distance of a network by amplifying the signal and reducing the electrical interference within the network. The repeater relays the data from one segment of the network to another without inspecting or modifying the data. Repeaters can also be used to connect segments of the network that use different cable media. Repeaters operate at the physical layer of the network.

A hub (Figure 24) is a multiport repeater that allows the distance of the network to be extended as well as allowing multiple devices to connect to the LAN. A hub is a device that brings all of the connections together (Figure 25). Like a repeater, the hub does not inspect or modify the data.

Repeaters and hubs boost the signals on a network, but do not solve problems involving collisions. Other devices, however, alleviate this problem by limiting traffic on the network. A bridge (Figure 26) is an electronic device that connects two or more networks running the same network protocols. It allows the isolation of the different networks. The different networks may have different topologies. Bridges are used to increase the efficiency of a network by limiting the amount of broadcast traffic to "one side" of the bridge. By doing this, networks can be expanded without a significant increase of traffic. Bridges operate at the physical and data-link layers and make use of the physical addresses associated with the network cards.

A switch (Figure 27) is a multiport bridge. Switches are often referred to as switching hubs. A switch combines the appearance and functionality of a hub with the additional functionality of a bridge. Switches have all but replaced hubs in local area network installations.

Although typically not part of a LAN, routers provide the interface between a LAN and WAN or between different LANs. Routers route the traffic between different

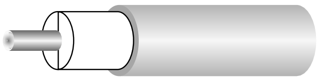


Figure 11: RG-58 coaxial cable.

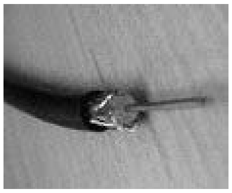


Figure 12: RG-8 coaxial cable.

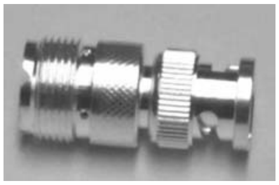


Figure 13: BNC connector.

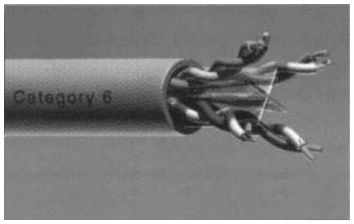


Figure 15: CAT6 twisted pairs of wires.

networks by maintaining tables of networks and their addresses. Routers are more sophisticated than switches and bridges and work at a higher level of the network.

LAN SOFTWARE

Local area network software can be classified into three categories: network operating systems, network utilities, and network applications software. Not too many years ago, operating systems and network operating systems were distinct. Today almost all operating systems have network operating system functionality built in. In other words, the user does not need to add any additional software to the operating system to get the computer to function on a network. All of today's popular operating systems, including all recent versions of Microsoft Windows, all versions of Linux and Unix, and all versions of the Macintosh OS, support the connection of a computer to a network right out of the box. In addition to the usual computing tasks performed by an operating system, a network operating system also

- Manages the network connection,
- Directs data traffic onto and off of the network,
- Manages the flow of data between the different devices,
- Manages communication and messages between the network users, and
- Provides for security of the data and other resources available on the network.

The network operating system provides the interface between the LAN hardware and the applications running on the host.

Included with most of the network operating systems are specific network utilities such as ping, arp, and traceroute, which provides network functions. Ping sends ICMP ECHO_REQUEST and ECHO_REPLY packets of data to another device to indicate the time it takes to reach a target machine. Ping is a useful utility for troubleshooting networking problems. Arp is used to map the physical address of a networked device to the corresponding network IP address that has been assigned to the device. Traceroute uses the ping utility to map the route packets of data take from a source to a destination machine. Network operating systems typically include drivers for most network adapters, so that the adapter can be plugged into the computer and the computer can function on the network without too much additional configuration.

Network application software includes client front end software that is specific for use by client computers. This would include Web browsers and e-mail software clients that would be run when needed. Hosts functioning as servers would have server software that would be constantly running, waiting for connections from clients. Web servers and e-mail servers would be examples of this. Other types of network application software would include database client and server software as well as groupware software.

ROLE AND APPLICATIONS OF LANs IN THE INTERNET, INTRANET, EXTRANET, AND E-COMMERCE WORLDS

One of the major uses of local area networks is to facilitate connections by users to the Internet. This requires the connection of the LAN to the Internet via either a dial-up telephone connection or a leased line. A dial-up connection requires a modem that converts the network's serial digital signal to the phone line's analog signal. A leased line connection requires a router that is then connected to another hardware device called a CSU/DSU (channel service unit/data service unit). The CSU/DSU in turn

Figure 14: CAT5 patch cable.

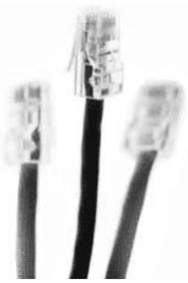


Figure 16: RJ45 connector.

Figure 17: Fiber-optic cable.

connects the network's router to the end of the leased line and converts the network's serial data signal to and from the leased line's digital signal. The leased line provides a high-speed Internet connection for the organization owning the LAN. The leased line is typically leased from an Internet service provider (ISP). The ISP maintains the actual connection to the Internet using its own router.

Connecting a LAN to the Internet requires that the devices on the LAN support the TCP/IP suite of protocols that provide the foundation of the Internet. These protocols are necessary for computers on the LAN to be able to communicate with devices in other parts of the Internet. The TCP/IP suite of protocols include

- Transmission control protocol (TCP)—establishes and maintains the Internet connection,
- Internet protocol (IP)—handles the routing of packets of data across the Internet,
- Simple mail transfer protocol (SMTP)—receives and delivers e-mail messages,
- Hypertext transfer protocol (HTTP)—facilitates the delivery of Web documents,
- File transfer protocol (FTP)—transfers files, and
- Telnet—allows users to remotely connect to other computers over the Internet.

Many organizations have set up their own internal Internets called intranets. An intranet is a private network that uses many of the same protocols as the Internet. An intranet appears to the user like the Internet, but it is typically not open to anyone outside the organization. An intranet is not a replacement for a LAN but rather runs within a LAN and supports many of the same applications as the Internet, typically Web servers and browsers and e-mail servers and clients, as well as additional groupware software. The core of most intranets is the Web site, which typically contains most of the internal documents that need to be disseminated among the organization's members. Setting up an intranet site on an organization's LAN requires a lot of organization and planning in selecting the hardware, software, and data needed to create a functional intranet.

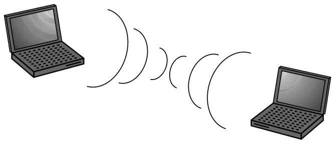


Figure 18: Peer-to-peer wireless network.

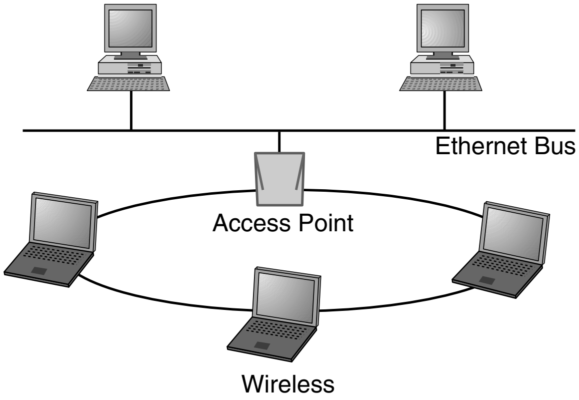


Figure 19: Wireless network with access point.

Some organizations have taken the intranet concept one step further and linked distributed parts of the organization together through the Internet via an extranet. An extranet is basically an intranet that may include access by customers, suppliers, and trusted partners.

WIRELESS LOCAL AREA NETWORKS

Wireless local area networks (WLANs) provide the freedom to access data without being tethered to the network with wires. WLANs enable users to take laptops and handheld computers anywhere, any time, and still access the network. This is becoming more important in today's world of information.

Wireless networks are easier and less expensive to install. Wireless networks do not require spending money on cable and then spending additional money and time installing the cable. Wireless access points and NICs have become much cheaper in the past year. Installing a wireless network involves turning on the access points, installing the software for the access points and NICs, and identifying the access points for the NICs to connect to. Modifying a wireless network is easy. There is no need to remove and/or relocate cable. There is no longer a concern for network cable failure.

Companies are using WLANs to keep track of inventory in warehouses. Workers who need to be constantly in contact with their networks are more frequently using

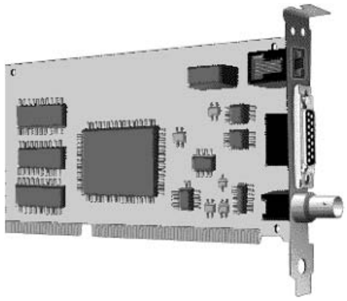


Figure 20: Network Interface Card with RJ45, BNC, and AUI connections.

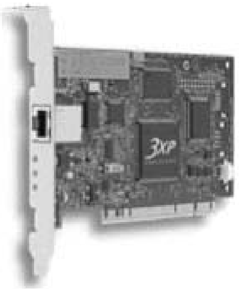


Figure 21: Network Interface Card with RJ45 connection.

wireless LANs. WLAN devices have become commonplace among workers in the health care industry. One area where WLANs are beginning to have a great impact is in education. Students and faculty no longer need to find wired computer labs to communicate and work. With wireless devices, students and faculty can access the networks in any buildings where wireless access points have been installed.

Wireless networks still have some drawbacks. Chief among these drawbacks are the limitations on distance and bandwidth. The radius of a WLAN can be extended by adding additional points of access to the network. Adding additional access points can also help increase the bandwidth of the WLAN. There are also new standards and associated wireless devices emerging that support increased bandwidth. The other major drawback for WLANs is security. A WLAN transmits radio signals over a broad area. This allows an intruder to lurk anywhere and intercept the signals from the wireless network. One way of inhibiting access to wireless networks is to turn on Wired Equivalent Privacy (WEP). WEP relies on a secret key that is shared between a mobile station and the access point. The secret key is used to encrypt packets before they are transmitted, and an integrity check is used to ensure that packets are not modified in transit. Unfortunately WEP is easily compromised. Additional levels of security are now being developed and implemented to protect the transmission of data.

Figure 22: PCMCIA wireless NIC.

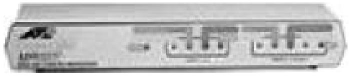


Figure 23: Repeater.

LAN INSTALLATION

Before a LAN is installed, a lot of planning needs to take place. The process can typically be broken down into seven steps:

- Needs analysis,
- Site analysis,
- Equipment selection,
- Site design,
- Server configuration,
- Installation schedule, and
- Installation.

Needs Analysis

The first aspect of installing a local area network is determining the needs of the organization and the users. Is a local area network needed? What aspects of the network are needed? Who will be using the network? What will they be using the network for? Will a local area network help the bottom line of the organization?

Reasons for installing a local area network might include

- Need for improved communication,
- Need for centralizing data,
- Need for sharing hardware,
- Need for automating work flow, and/or
- Need for enhanced security of data.

Site Analysis

Once a need has been established, it is necessary to determine where the LAN will be installed. What parts of the organization's site will be networked? Where will the servers be located? A site plan will need to be drawn. If a fire escape plan is available, it can be used as a template for the building and the locations of rooms and doors. It is best if the architectural plans can be found. The site plan (Figure 28) is a map of the location where the network is installed and should include

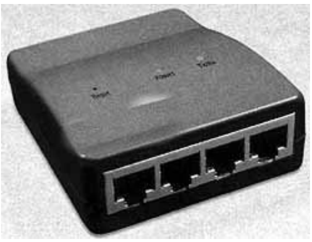


Figure 24: 4-port hub.

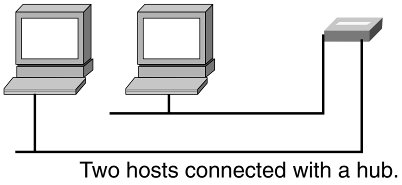


Figure 25: Hub connecting two hosts.

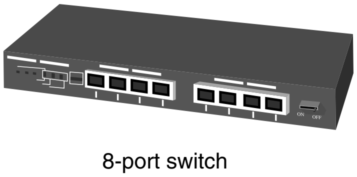


Figure 27: Switches.

- The dimensions of the site including the location of each employee,
- The location of all immovable objects including doors and windows,
- The current location of all moveable objects,
- The location of heating, ventilation and air conditioning systems and ducts,
- The location of electrical outlets and the current wiring scheme, and
- The current location of all computer equipment and the planned location for any additional devices.

Equipment Selection

As the site plan is developed, an inventory of equipment on hand needs to be conducted. An inventory of equipment will identify the capabilities of the equipment incorporated into the proposed network. This will identify which equipment will be obsolete once the network is installed and which equipment will require modification. Older workstations may still be useful as print servers. Table 2 shows the features to be noted in the equipment survey.

Once the current equipment has been inventoried, it is now time to identify new equipment that will need to be purchased. This list should be correlated with the user needs identified earlier. Once this list is prepared, vendors can be contacted, soliciting their recommendations for meeting the hardware and software needs. Be sure to consider any infrastructure constraints including electrical power and distances. Table 3 is an example of a form that could be used to compare different vendor proposals.

Site Design

Once the site plan from the site analysis has been completed and the equipment lists have been completed, it is time to create a working site design (Figure 29). This design will include details of where all devices, including networking devices, will be located. The locations for all network connections must be indicated. The locations of all network cable must be delineated.

Server Configuration

Once the computers that will be installed as servers arrive, they need to be configured. Server software needs to be installed and the directory structure of the server needs to be organized. The directory structure begins with the root directory, with all other directories within it and the files and subdirectories within those. Typically, there will be directories for the network operating system, separate directories for each server application, and directories for the clients who will be connecting to the server.

Installation Schedule

Networks take a considerable amount of time to install. It is important to have an installation schedule. There will be times when employees' computers will need to be turned off and possibly moved. Disruption needs to be minimized wherever possible. Be sure to include in the installation schedule the possibility for shipping delays on the equipment that has been ordered. Be sure to read the manuals before the installation begins so that there will not be any surprises once the installation starts. Also prepare the site before the installation begins. This may involve moving furniture, installing new outlets, and removing equipment that will no longer be used. Don't forget to back up any data that are on systems that will be affected by the move.

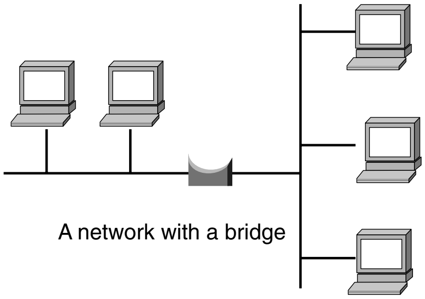


Figure 26: Bridges.

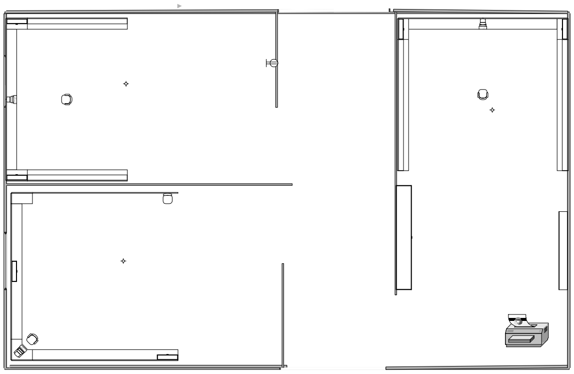


Figure 28: Sample site plan.

Table 2 Equipment Inventory Form

Serial Number	xxxxxxxxxxx
Processor	Intel Pentium 4–1.8 GHz
RAM size and configuration	256 MB; 2 × 128 MB DIMM
Hard disk	20 Connor GB
Other drives	One 3.5,″ 1.44 MB
CD-ROM	Toshiba 24×
Monitor	17″ Toshiba XVGA
Warranty information	Expires Jan. 2003

Installation

Before the installation begins, it is best to discuss everything with someone who has been through a LAN installation before. Have this person look over the designs, schedules, and forms to ensure that nothing was forgotten. Depending on the size of the installation, it may take anywhere from a couple hours to a couple of days. Be prepared for delays. If this is a wired network, the cable media will need to be pulled first. This is the part that typically takes the longest time. While the cable media are being pulled or shortly thereafter, the remaining components will need to be installed. This would typically include the patch panels in the communications rooms and the wall outlet for each device that will be networked. Once this is completed, the electronic devices (hubs, switches, etc.) will need to be installed. If network cards need to be installed in the older computers, this will be done next. Finally, the software for the network cards will need to be installed on each workstation and server if this has not already been done. Be sure to install the appropriate network software.

The final two stages of the installation are testing and training. Every device will need to be tested to ensure that its network connection works. This should be done before any of the users access the network. Be sure to test all of the network devices and the networked printers. Users will need to be trained in the procedures they will need to follow to use the network.

LAN ADMINISTRATION

Once the LAN has been installed, there are several key aspects of administering the network.

Configuration List

Using the equipment inventory forms developed during the installation, along with similar documentation for the new devices, a set of configuration lists needs to be developed. These would include a list of all computers, similar to the equipment inventory form, directory lists for each server that is installed, a list of all server users, and a list of all printers and other shared devices. It is also very important to keep copies of the network plans developed for the installation of the network.

System Log

The system log is documentation for the network. It provides a detailed history of the network's hardware, software, and configuration features. As changes are made to the network, these changes need to be documented in the system log. As problems arise with the network, these also need to be documented in the system log. This log needs to be maintained from the beginning. The system log should include all hardware and software warranties, hardware and software information, current setup structure, backup and recovery plan, backup logs, and error/downtime logs.

Training

Training does not end after the network has been installed. The network is a dynamic part of an organization's computing system. As changes are made, employees will need additional training.

Table 3 Vendor worksheet

	Vendor 1			**Vendor 2**			**Vendor 3**		
	Model	**Quan.**	**Price**	**Model**	**Quan.**	**Price**	**Model**	**Quan.**	**Price**
Server									
H'ware									
S'ware									
Workstations									
H'ware									
S'ware									
NIC									
Switches/hubs									
bridges									
Access points									
Routers									
Cabling									
Other									
TOTAL COST:									

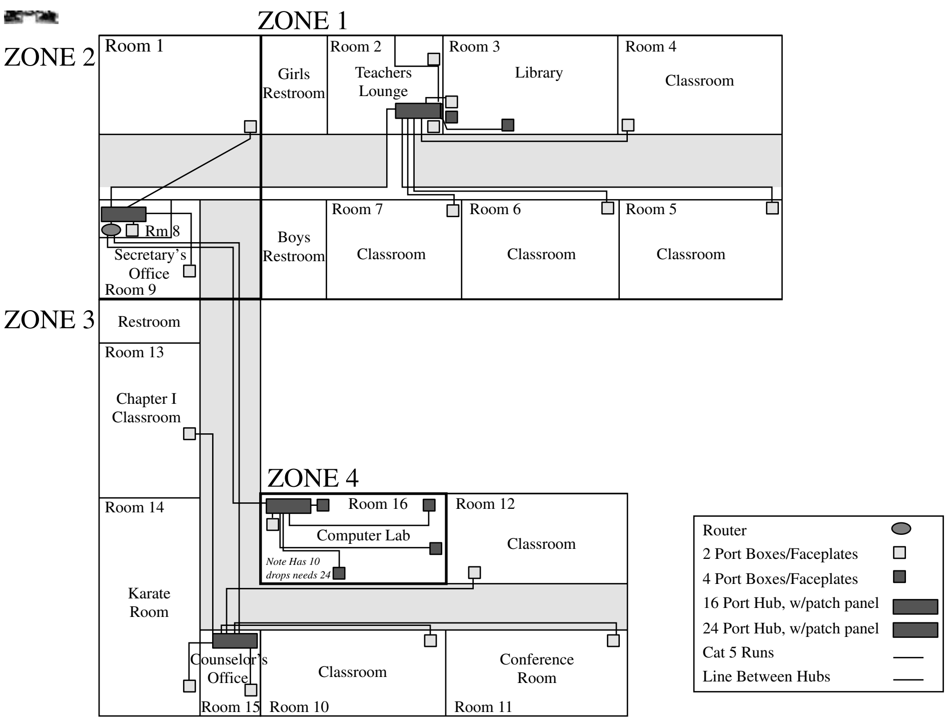


Figure 29: Sample site design.

Backup

As more and more data are kept on networked servers and shared by more than one user, it becomes critical that a procedure be established for backing up critical data. This may be the most important task for a network administrator. Hardware and software can be replaced. New employees can be hired. But it is difficult to recreate large amounts of data. The network administrator must establish a schedule for backing up all critical data.

Security

Once computers become accessible to other individuals, security issues need to be considered. In a networked environment, users will need user IDs and passwords. If security is extremely important, then encryption of data might also need to be implemented. If the local area network is attached to another network, then a firewall may be necessary to control access to critical data between the two networks.

CONCLUSION

Local area networks have played a very important role in providing computer access to large numbers of users. They have allowed users to share hardware, software, and most importantly data. They now also provide access to the Internet as well as organizations intranets and extranets. With the emergence of wireless local area networks, the applications of local area networks will continue to expand.

GLOSSARY

Access point A hardware device consisting of an antenna, a radio transmitter, and a wired network interface. Used for wireless networks.

Bridge A network device for segmenting a local area network into two or more segments.

Bus network A network where the nodes are arranged in a linear fashion.

Ethernet The most widely used local area network architecture standard.

Extranet A private portion of the Internet accessible only to individuals within an organization.

Hub A network device for connecting multiple nodes to a network.

Internet A vast collection of networks linked together, built around the TCP/IP Suite of protocols.

Intranet An organization's internal network, built around the TCP/IP Suite of protocols.

Local area network (LAN) A network where the nodes are typically physically close together.

Logical topology Based on the electrical configuration of the network.

Network A collection of two or more devices.

Network interface card (NIC) An adapter for connecting a computer to the network media.

Physical topology Based on the physical configuration of the nodes in the network.

Repeater A network device for extending the radius of a network.

Ring network A network where the nodes are arranged in a circular fashion.

Router A network device for connecting a LAN to another network, typically a WAN.

Star network A network that has the nodes connected to a central device, typically a hub or a switch.

Switch A network device for connecting multiple nodes to a network, allowing two devices to communicate only with each other at a given moment.

Wide area network (WAN) A network where the nodes are physically far apart.

CROSS REFERENCES

See *Conducted Communications Media; Extranets; GroupWare; Intranets; Wide Area and Metropolitan Area Networks; Wireless Communications Applications.*

FURTHER READING

Charles Spurgeon's Ethernet Web Site (2002). Retrieved October 7, 2002, from http://www.ethermanage.com/ethernet/ethernet.html

Ciama, Mark (2001). *Guide to designing and implementing wireless LANs.* Boston, MA: Course Technology.

Comer, D. E. (2001). *Computer networks and internets with Internet applications* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Ethernet Codes master page (2002). Retrieved October 7, 2002, from http://map-ne.com/Ethernet/

Goldman, J. E. (1997). *Local area networks: A client/server approach.* New York: Wiley.

Goldman, J. E. (1998). *Applied data communications, A business-oriented approach* (2nd ed.). New York: Wiley.

IEEE 802.3 CSMA/CD (ETHERNET) (2002). Retrieved October 7, 2002, from http://grouper.ieee.org/groups/802/3/index.html

Link-layer technologies (2002). Retrieved October 7, 2002, from http://www.cs.columbia.edu/~hgs/internet/ethernet.html

Official Internet protocol standards (2002). Retrieved October 7, 2002, from http://www.rfc-editor.org/rfcxx00.html

Panko, R. (2001). *Business data communications and networking* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Stallings, W. (2000). *Data and computer communications* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Subramanian, M. (2000). *Network management: Principles and practice.* Reading, MA: Addison Wesley.

Taylor, E. (2000). *Networking handbook.* New York: McGraw Hill.

Thomas, R. M. (1989). *Introduction to local area networks* (2nd ed.). Sybex Network Press.

Webopedia (2002). Retrieved July 5, 2002, from http://www.pcwebopaedia.com/TERM/l/local_area_network_LAN.html

M

Machine Learning and Data Mining on the Web

Qiang Yang, *Hong Kong University of Science and Technology, China*

INTRODUCTION

In recent years, machine learning and data mining have made great impact on the Web, both in research and applications. This impact is, to a large part, due to the availability of large quantities of data in many information sources on the Web. A major source of data is the Web itself, which is ever increasing at a rapid rate. The Web contains millions of Web pages that are interconnected to each other in complex ways. The content of many Web pages is also dynamically changing. Visitors to the millions of Web servers leave large quantities of Web log data. It is natural for one to wonder if any significant, interesting, and even unexpected patterns can be discovered from the Web using machine learning and data mining techniques.

The possibility of learning useful information from the Web has been recognized early by researchers such as Etzioni (1996) and has been tackled by researchers from many different angles including machine learning, agent technologies, data mining and knowledge discovery, information retrieval, and computer networks. In this chapter, we give a focused survey of some significant techniques and applications of machine learning and data mining on the Web. In this survey, we follow the categorization of Zaïane, Xin, and Han (1998) to classify the Web mining and machine learning issues into Web content mining, Web structure mining, and Web usage mining. To this categorization we also add Web-based recommendations using collaborative filtering, which is often used in electronic commerce applications.

For introductory tutorials on machine learning, two of the popular books are *Machine Learning* by Mitchell (1997) and *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* by Witten and Frank (1999). For an introduction to data mining, the recommended readings are *Data Mining Concepts and Techniques* by Han and Kamber (2001) and *Principles of Data Mining* by Hand, Mannila, and Smyth (2001).

WEB CONTENT MINING

Wrapper Induction

In its raw form, Web pages are coded in the HTML format. This unstructured format makes it very difficult for computer programs to make sense of its content and manipulate it. To first convert the Web page into structured formats such as relational table formats or hierarchical formats (XML), users often extract useful content information from the Web pages by following a set of rules. When made explicit, the set of rules is called a wrapper. The process of building the rules from experience and handcrafted examples is called wrapper induction. The term *wrapper induction* was coined by Kushmerick, Weld, and Doorenbos (1997). We illustrate the key idea behind wrapper induction using the Stalker system of Knoblock, Lerman, Minton, and Muslea (2000).

Consider the example fragments from a university Web site in Figure 1, where Ex1 through Ex4 are portions of the faculty home pages that contain address information.

From these examples, we wish to build a relational database table such as that shown in Table 1.

The task of wrapper induction is of great importance to practitioners. With a cleaned table such as Table 1 it is possible to offer the user with more advanced services, such as structured search by attributes, integration of multiple

```
Ex1: ...Faculty:<i>John</i><p>Address:<i> 3435 CS Building </i><p>Phone:<i>...

Ex2: ... Faculty:<i>Smith </i><p>Address:<i>233 Engineering </i><p>Phone:<i>...

Ex3: ... Faculty:<i>Mary</i><p>Address:<i> 3225 CS Building </i><p>Phone:<i>...

Ex4: ... Faculty:<i>Steve</i><p>Address:<b> 3112 Admin </b><p>Phone:<i>...
```

Figure 1: Training examples for the wrapper learning algorithm.

databases that are distributed on the Web, or accomplishment of even complex tasks such as comparison shopping and travel planning (Knoblock et al., 2000).

Inspecting these examples, one possible extraction rule can be constructed as

```
R1= SkipTo(Address)SkipTo(<i>),
```

which has the following meaning: start from the beginning of the document and skip every token until you find a landmark consisting of the word *Address,* and then, again, ignore everything until you find the landmark <i>. **R1** is called a *start rule* because it identifies the beginning of the address. One can write a similar *end rule* that finds the end of the address; for sake of simplicity, we restrict our discussion here to start rules.

Other rules can also be used to identify the address. For example,

```
R2=(Address:<i>)
R3=(Faculty:<i>_Capitalized_</i>
 <p>Address:<i>)
```

can also be used as start rules. **R2** uses the three-token landmark that immediately precedes the beginning of the address in examples E1, E2, and E3, while **R3** is defined based on a nine-token landmark that uses the wildcard _*Capitalized*_, which is a placeholder for any capitalized alphabetic string.

To deal with variations in the format of the documents, one can introduce *disjunctions* into the extraction rules. For example, let us assume that the addresses that are within one mile from your location appear in bold (see example Ex4 in Figure 1), while the other ones are displayed as italic (e.g., Ex1, Ex2, and Ex3). We can extract all the names based on the disjunctive start rule

```
either(Address:<b>)
orSkipTo(Address)SkipTo(<i>)
```

Wrapper induction algorithms are based on supervised learning methods, by first constructing a training data set with the help from human users. From the Web pages, for example, a human expert can build the following table of potential indicative tokens. The inductive learning algorithms can then use these examples to train different classification models for learning how to recognize the given class labels such as "Address" and so on.

Table 1 Cleaned Training Examples

Faculty Names	Address	Phone
John	3435 CS Building	—
Smith	233 Engineering	—
Mary	3225 CS Building	—
Steve	3112 Admin	—

Table 2 Data Prepared for Training the Covering Algorithm

Example ID	Token 1	Token 2	Class
1	Address	<i>	*Address*
2	Address	<i>	Address
3	Faculty	:<i>	Faculty
4	Phone	:<b>	Phone

The examples can be further extracted to form the training examples as shown in Table 2. Using these training examples, various classification models can be learned. Kushmerick (2000) studied different classes of inductive learning algorithms that are grouped under the class PAC learning. The idea is to feed positive and negative examples to a learning algorithm, which then constructs a disjunctive normal form to capture the commonalities of the examples. An issue is how many examples must the learner see before it can confidently reduce its error to a minimum. Kushmerick explored this so-called sample complexity. Knoblock et al. (2000) used a rule-learning algorithm known as a *covering algorithm* by performing a local search in the space of extraction rules. Using this algorithm, a learner iteratively builds a set of rules based on the examples it has seen so far. Each time a new rule is selected, a subset of remaining examples are covered by the new rule. The process continues until all examples are covered. Recognizing the sequential nature of the problem, Hsu and Dung (1998) extended this idea by designing a finite-state automaton to learn the states and their transitions. These states and transitions are essentially the inductive rules for extracting the contents. Hsu and Dung demonstrated that this method is superior to many other algorithms in accuracy.

Learning Web Ontology

As the XML is moving into the forefront of Web-based computing and database management arenas, the importance of ontology learning becomes apparent. Ontology refers to shared domain theories about object names and meaning that help not only humans, but also machines (Berners-Lee, Hendler, & Lassila, 2001). For example, a database may use "phone" to mean agent phone for one database schema in a real-estate Web site, while "agent-phone" in another schema. The mapping function between the two schemas or XML DTDs can be learned from examples. Doan, Domingos, and Levy (2001) described a multistrategy learning algorithm to learn a variety of mapping functions between the two schemas, including a naïve Bayesian classifier and a domain-dependent name matcher.

Web Page Classification

Another active area of Web content mining is Web page classification. The idea is to classify the Web pages into a known class using supervised learning techniques. For example, based on the content of the page, a Web page classifier may learn how to categorize Web pages into one of financial news or political news or some

Table 3 Classification Example

Key words				
Stock	Movie	Show	Banks	Class
1	1	0	1	financial
0	1	1	0	entertainment
1	1	1	0	financial
—	—	—	—	—

other categories. Using this idea, Dumais, Cutrell, and Chen (2001) experimented with building a user-friendlier search engine. Liu, Ma, and Yu (2001) applied Web page classification to discover unexpected information from competitors' Web sites.

We now illustrate Web page classification using the naïve Bayesian models. The naïve Bayesian method assumes that the attributes of the data are independent and use the joint distribution of the data to make classifications; Witten and Frank (1999) provided a detailed description.

Given a set of Web pages, each page can be considered a collection of words; this collection is also known as a bag of words. The words in each document form a vector, and together we have a vector space model. Essentially, we can consider the vector space model as a relational training table. In a supervised learning scheme, each vector has a class label such as "entertainment" or "financial." For example, suppose we choose the words "stock," "movie," "show," "banks" as the key words to classify the Web documents into entertainment and financial classes. The training data can then be formed as shown in Table 3.

Suppose that we choose to use a naïve Bayesian method for classifying this collection of documents. Let C be a class label such as "financial," K_i be the binary presence of a key word such as "stock." Let $Pr[C \mid K_i]$ be the probability that a document belong to the class C given that the key word K_i appears in the document. Then the probability that a particular document D belong to a class C can be calculated using the naïve Bayesian formula (1) with the labeled documents as training data. The class with the largest probability value is taken as the predicted document.

$$\Pr[C \mid D] = -\frac{\Pr[K_1 \mid C]^* \Pr[K_2 \mid C]^* \ldots \Pr[K_n \mid C]^* \Pr[C]}{\Pr[D]}. \tag{1}$$

Besides machine learning, the information retrieval community is strongly interested in Web page classification and full-text search on the Web. This community has a long history of research in text retrieval, where the text can be any text-based document. Both supervised and unsupervised learning methods have been applied to document retrieval. McCallum and Nigan (1998) and Yang (1999) provided systematic studies of statistical methods for document retrieval.

WEB STRUCTURE MINING

Page Rank Algorithm

We now turn our attention to the discovery of important structural relations on the Web. The hyperlink structure of the Web presents a unique opportunity for researchers to learn a great deal about the Web and its users. Conceptually, we can consider each Web page as a node in a graph structure and the hyperlinks that connect the Web pages as edges between the nodes. Then the Web can be modeled as a giant directed graph. This graph encodes important information about the intention of the Web page authors and relation between Web contents. For example, by counting how many pages are pointing at a particular page, we can estimate how important the authors of the Web pages would consider that page. This important page ranking score can be constantly updated using a crawler and used to enhance a search engine. In fact, this is the key idea behind powerful search engines such as Google (http://www.google.com).

Before the arrival of the Google search engine, the hyperlink structure had largely gone unused in Web search engines. In developing Google, Brin, and Page (1998) had an intuitive idea. By crawling millions of Web pages with hundreds of servers, a Web page's "PageRank" can be computed that corresponds to an objective measure of its citation importance that corresponds well with people's subjective idea of importance. At a high level, PageRank captures the intuition that a page is important if it is pointed at by many other pages, or by other important pages. This knowledge can be learned by counting the so-called *back links* to each page; that is, the hyperlinks that lead to the page. Practice has shown that PageRank is an excellent way to prioritize the results of key word based searches on the Web.

PageRank is defined as follows. Suppose that a Web page A has pages $T_1 \ldots T_n$ that point to it. The parameter d is a damping factor, which can be set between 0 and 1. Let $C(A)$ be defined as the number of links going out of page A. The PageRank of a page A is given as

$$PR(A) = (1 - d) + d(PR(T_1)/C(T_1) + \cdots + PR(T_n)/C(T_n)). \tag{2}$$

PageRank or $PR(A)$ can be calculated using a simple iterative algorithm, and has been found to correspond to the principal eigenvector of the normalized link matrix of the Web.

Reinforcement Learning

The PageRank algorithm is an instance of the more general reinforcement learning algorithm in machine learning for solving sequential planning and learning problems (Kaelbling, Littman, & Moore, 1996). Reinforcement learning builds on a collection of states that are interconnected to each other in a network, similar to how hyperlinks are connecting the Web pages. The value of certain states in this network is considered high if they are worthwhile to reach from one or several starting states. The system learns, among the many possible edges in the network, which are the good intermediate states to reach in order to find the final goal states. This is accomplished by letting the important states reinforce each other. A state is important if it is connected to other important states. Furthermore, the closer a state is to an important state, the more important it is as well. This is similar to the damping factor d used in the PageRank algorithm.

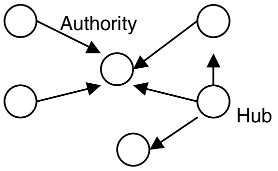


Figure 2: Illustrating hubs and authorities in Kleinburg (1998).

Hubs and Authorities

Google's idea of page ranking can be further extended to include a classification system for Web pages that are important in two different ways. The *authority* Web pages are those that are pointed to by many other Web pages. The *hub* pages are those that contain many pages that point to other pages. These hubs and authorities relationship reinforce each other similar to the reinforcement-learning algorithm. These pages can be used when a search engine decides which page should get higher priority in search results ranking. This is the essential idea behind Kleinberg's *Hubs & Authorities* (1999) algorithm for searching Web pages.

With each page p, we associate a nonnegative *authority weight* $x(p)$ and a nonnegative *hub weight* $y(p)$. We maintain the invariant that the weights of each type are normalized so their squares sum to one. We view the pages with larger x- and y-values as being better authorities and hubs, respectively. The scores for all pages can be computed using the reinforcement idea. If p points to many pages with large x-values, then it should receive a large y-value; and if $p$is pointed to by many pages with large y-values, then it should receive a large x-value. Given weights $x(p)$ and $y(p)$, and the set of hyperlinks E that contains all links of the form (q, p) or (p, q), the weights can be computed from the formulas (3) and (4). Then

$$x(p) \leftarrow \sum_{q:(q,p)\in E} y(q) \tag{3}$$

$$y(p) \leftarrow \sum_{q:(p,q)\in E} x(q). \tag{4}$$

With these definitions and algorithms, it is possible to greatly improve the basic key word based search on the Internet. For any key word based search, the resulting pages can be further ranked by their hubs and authorities scores. The pages that are ranked high in the list are returned as the search result. In experiments on real Web pages, this method has shown dramatic improvement over text-based search engines.

WEB USAGE MINING

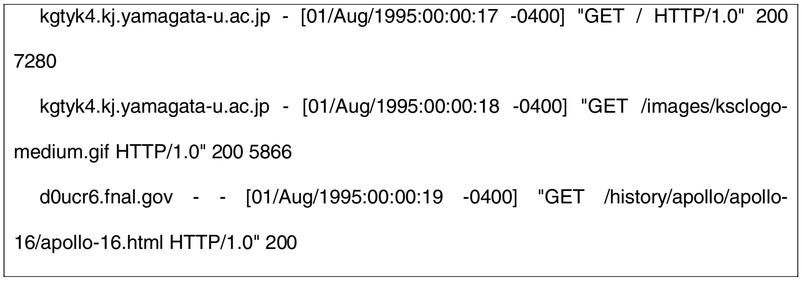
Web servers record the visiting behavior of millions of Web users. By analyzing the Web logs, it is possible to obtain important insights into user behavior on Web sites and use the information to guide other applications. For example, it is possible to predict a user's future needs based on their past visits to the Web servers. With the learned information on user preferences and behavior, it is also possible to design better Web caching systems and prefetching systems for retrieving new Web pages. It is even possible to dynamically reconfigure the user interface for a particular user to cater to the user's individual interests (Srivastava, Cooley, Deshpande, & Tan, 2000).

Obtaining Web Logs

Much of the machine learning and data mining research in Web usage mining starts from the log data accumulated at Web servers or proxy servers. Figure 3 shows an example Web log data from a NASA Web site; many of such logs are available for experimentation (http://www.Webcaching.com). Each record in the log file contains the IP address of the visitor, the date and time of the visit, the HTTP protocol used in the transaction, the error code, and the file size. Any popular Web server can easily accumulate gigabytes of such data in a short amount of time.

Extracting User Sessions

Based on the Web logs, researchers have built various machine learning algorithms to extract the Web usage patterns of user groups as well as individuals. The first task is usually data cleaning, by breaking apart the long sequence of visits by different users into user sessions. Here a user can be identified by an individual IP address. Furthermore, the sequence of visits may be made by the same user at different times. Thus, there is a further need to separate the visiting sequence of pages into visiting sessions. Most work in Web usage mining employs a predefined time interval to find the visiting sessions. For example, one can use a two-hour time limit as the separating time interval between two consecutive visiting sessions. However, by



```
   kgtyk4.kj.yamagata-u.ac.jp - [01/Aug/1995:00:00:17 -0400] "GET / HTTP/1.0" 200
7280
   kgtyk4.kj.yamagata-u.ac.jp - [01/Aug/1995:00:00:18 -0400] "GET /images/ksclogo-
medium.gif HTTP/1.0" 200 5866
   d0ucr6.fnal.gov - - [01/Aug/1995:00:00:19 -0400] "GET /history/apollo/apollo-
16/apollo-16.html HTTP/1.0" 200
```

Figure 3: An example Web log.

studying the Web logs carefully, one can find more meaningful session separators. The work by Lou, Liu, Lu, and Yang (2002) designed an algorithm using clustering analysis to find the group of related Web servers visited by users from a Web proxy server. If a user jumps from one group of related servers to another, then it is highly likely that the user ends one session and starts another.

Learning Path Profiles

Once a Web log is organized into separate visiting sessions, it is possible to learn user profiles from the data. Schechter, Krishnan, and Smith (1998) developed a system to learn users' path profiles from the Web log data. A user visiting a sequence of Web pages often leaves a trail of the pages URL's in a Web log. A page's *successor* is the page requested immediately after that page in a URL sequence. A *point profile* contains, for any given page, the set of that page's successors in all URL sequences and the frequency with which that successor occurred. A *path profile* considers frequent subsequences from the frequently occurring paths. They can be used to predict the next pages that are most likely to occur by consulting the set of path profiles whose prefixes match the observed sequence. The idea was similarly exploited by Pitkow and Pirolli (1999) in their work on finding the longest repeating subsequences from a Web log and using these sequences to predict users' next likely requests. Similar to the path–profile-based prediction, Albrecht, Zukerman, and Nicholson (1999) designed a system to learn an *N*th-order Markov model based on not only the immediately preceding Web page, but also pages that precede the last pages as well as the time between successive Web page retrieval activities. Together these factors should give more information than the path profiles. The objective is still the same: to predict which is the next page to be most likely requested by a user.

One particular fact about the Web is that the user sessions follow a Zipf distribution. By this distribution, most visitors view only few pages in each session, while few visitors visit many pages in a sequence. However, experiments show that the longer paths also provide more accurate predictions. To exploit these facts, Su, Yang, and Zhang (2000) developed an algorithm to combine path profiles with different lengths. For any given observed sequence of visits, their "cascading" model first selects the best prediction by looking for path profiles with long lengths. When such paths do not exist in the model, the system retreats to shorter length paths and makes predictions based on these paths. Experiments show that this algorithm can provide a good balance between prediction accuracy and coverage.

Web Object Prediction

It is possible to design systems that learn from Web logs and make predictions. However, how are these predictions actually applied on the Web? One way to apply the prediction is to integrate traditional Web caching and model-based Web-object prefetching together. Web caching is already a well-studied area in the computer network area, where mature technology exists for ranking objects that are in the cache when new objects arrive.

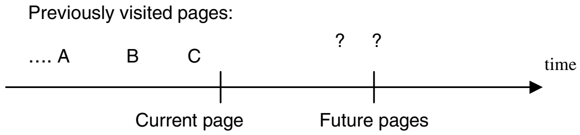


Figure 4: Illustrating Web page prediction using path profiles.

These algorithms are called cache-replacement policies (Cao & Irani, 1997). To make the integration possible, Yang, Zhang, and Li (2001) computed the probability that an object will be requested in the near future using an association-rule-based model and then used the probability to estimate the future access frequencies. The association rules are similar to the path profiles mentioned above, but with a more sophisticated rule selection and pruning method. Each association rule comes with a support and confidence measurement as in association rule mining (Argawal & Srikant, 1994):

$$P_1, P_2, \ldots, P_n \rightarrow O_i$$
$$\text{support} = \text{count}\,(\{\mathrm{P}_1, \mathrm{P}_2, \ldots, \mathrm{P}_n, \mathrm{O}_i\})$$
$$\text{conf} = \frac{\text{count}\,(\{P_1, P_2, \ldots, P_n, O_i\})}{\text{count}\,(\{P_1, P_2, \ldots, P_n\})}.$$

In this rule, P_1 through P_n are Web pages that occur in a row up to the prediction window, and O_i is the object that is predicted to occur. An object can be an HTML page itself or it can be an image or a sound object that is embedded in an HTML page. If the number of items $\{P_1, P_2, \ldots, P_n, \mathrm{O}_i\}$ is over a minimum support threshold, and the confidence value is above a predefined confidence threshold, then the rule is stored in a prediction module.

Learning to Prefetch

When a user requests a sequence of pages that matches the left-hand side of a rule, the right-hand side can be predicted to occur with the associated confidence value as the probability of occurrence of O_i. The Web prefetching system can then add all such predicted probabilities together from all rules in the prediction model as a measure of the potential future frequency estimation for O_i. These scores are compared, and the top percentages of the objects are selected to be prefetched into the cache if they are not already there. This prefetching method runs side by side with an existing caching system. Empirical tests on several real Web logs show the integrated system can indeed outperform Web caching alone by a large margin in terms of higher hitting rates, byte-hit rates, and reduction in latency. It should be noted that the integrated system demonstrated an increase in network bandwidth as well. Thus, a balance should be reached in the number of pages that are prefetched versus the number of pages that are cached using the traditional caching algorithms.

Web Agents

Web users have been burdened not only by the delayed response from Web servers, but also by the sometimes confusing Web interfaces. The problem is caused by the complex structures of Web sites and of the contents of the pages themselves. For example, a user looking for a particular faculty member in a university Web site may find himself lost in the wrong path. To this end, machine learning has also been applied to help alleviate the problem. WebWatcher (Joachims, Freitag, & Mitchell, 1997), an agent tour guide system, learned users' Web browsing preferences through an adaptive user interface. It does this by integrating both content and usage information in order to prioritize the links on each page a user visits. To use the system, a user types in a text description of the final objective of the browsing session; for example, a user may first specify that he or she wants to find the publications of a particular faculty member working on machine learning at a university Web site by providing the system with the key word "machine learning faculty." This description is then associated with each link the user traverses. This information is used to assess the similarity to other users browsing sessions for the same goal. The WebWatcher agent can then determine the links that are most relevant to the targeted page and highlight such links for the user. Another example is the AVANTI Project (FN96), which attempts to predict the user's eventual goals. Similar to WebWatcher, AVANTI presents prominently the links leading to pages a user will want to see.

Web Page Clustering

The PageGather system (Perkowitz & Etzioni, 2001) provides users with shortcuts to the users' intended pages and displays them on a page. It takes as input a Web server's access log, where the log data record the pages visited by a user at the site. Based on the visiting statistics, the system provides links on each page to visitors' eventual goals, skipping the in-between pages. Perkowitz and Etzioni (2001) defined the *index page synthesis problem* as: given a Web site and a visitor access log, create new index pages containing collections of links to related but currently unlinked pages. An *index page* is a page consisting of links to a set of pages that cover a particular topic (e.g., electric guitars). In their system, two pages are considered linked if there exists a link from one to the other, or if there exists a page that links to both of them. To find a collection of pages that are related to each other, a clustering algorithm is applied to the Web logs. The PageGather algorithm takes five basic steps as shown in Table 4.

A requirement for the PageGather algorithm is that in its last step, a human webmaster is needed to determine the appropriateness of the generated index pages. This will likely create a bottleneck for the workflow, especially for sites that have many Web pages to be indexed. Is there a method to find out the optimal number of index pages and links to generate without too much human intervention? Aimed at solving this problem, Su et al. designed an algorithm for finding an optimal number of pages and hyperlinks to include in the index pages (Su, Yang, Zhang, Hu, & Ma, 2002). First, they applied a clustering algorithm on the Web log data similar to the PageGather algorithm. Then, instead of inspecting the resulting clusters manually, they defined a cost function to capture the users' efforts in browsing a Web site. They defined the cost for a user to access the index pages to be the *OverallCost*, which is the overall cost of the browsing Web pages and index pages. This cost has two parts. The first part *PageCost* is the cost of flipping index pages. The second part *TransitionCost* is the cost of switching from one Web page to another. Then our cost models are as follows:

$$OverallCost = PageCost + TransitionCost. \tag{5}$$

Based on the cost functions, Su et al. calculated the effect of including a variable number of index pages as shown in Figure 5. The graph shows that as the number of index pages increases, the transition costs decrease due to the number of potential target pages that are included in the index pages. However, the cost of reading through a single index page also increases due to the number of hyperlinks included. The optimal point for the overall cost can be seen to be between the two extremes in the number of index pages. A webmaster can use this optimal point to decide how many index pages to include for the Web site.

Learning to Adapt Web Sites

The Web is increasingly connected to the wireless world. The adaptive Web site system of Anderson, Domingos, and Weld (2001) takes a Web log as input and learns the users' preference models to build a more compact Web site for wireless WAP users. For example, a Web user may frequently access certain pages while almost never visit other pages. It can also observe that the user who makes a sequence of visits through a set of Web pages almost always ends with a certain page—the target page of the sequence. It then uses this information to build a utility model for the space of pages and search for a subset of Web pages that maximizes the expected utility. The end result is a smaller Web site catering to a particular user or group of users, which can then be used to dynamically generate a new Web site for other devices, such as the PDA or cell phones.

Table 4 The PageGather Algorithm

PageGather Algorithm (Server Access Logs)
Process the access log into visits.
Compute the cooccurrence frequencies between pages.
Create the graph corresponding to the matrix to find maximal cliques.
Rank the clusters found.
For each cluster, create a set of index links to the documents in the cluster.
Present clusters to the Webmaster for evaluation.

Source: Perkowitz & Etzioni (2001)

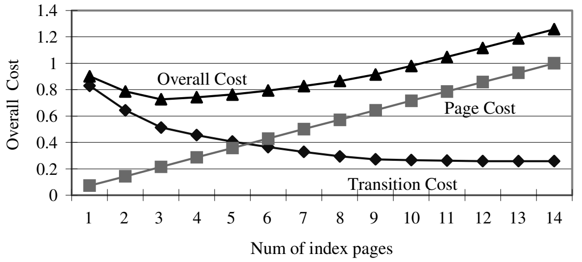


Figure 5: Overall costs computed based on the NASA Web log.

Learning from Click Streams

Any useful information about the Web users is of great value to the Web site managers. The WUM system (SP00) operates on Web logs to present the Web site managers with a graphical view of the sites' traversal patterns. The KDD CUP 2000 (Kohavi, Brodley, Frasca, Mason, & Zheng, 2000) was designed to learn the relationship between user behavior and the business goals of an electronic commerce Web site. For example, a question posted in the competition was this: Given a set of page views, will the visitor view another page on the site or will the visitor leave? A large number of machine learning tools were applied to build models for this and other related questions. For example, many researchers used decision trees, regression analysis, and neural networks to learn from the data. The result was shown to be indeed very useful in guiding the marketing activities of the company.

WEB-BASED RECOMMENDATION FOR E-COMMERCE

With the rise of the Internet, the Web has become a major medium for companies and institutions to reach out to its users and customers. On the Web it is possible to assign an ID number to a user, so that the users' behaviors can be identified. For a frequently visiting user, it is possible to make recommendations based on the users' previous behaviors or based on other similar users' behaviors.

Pearson Correlation

The basic machine learning algorithm for recommendation systems is based on statistical correlation, known as Pearson correlation. Stated simply, it makes inference on a user's preferences for new products based on ratings of other similar users for the same or similar products. Here the notions of similarity and relationship are important. The application of correlation as the basic relationship builder starts with the Tapestry system. The Tapestry system (Goldberg, Nichols, Oki, & Terry, 1992) defines collaborative filtering as a process where people collaborate to help one another perform filtering by recording their reactions to documents they read. It allows people to rank documents that users find interesting based on annotations contributed by others. Based on the given relationship between users, advisor and advisees can be discovered. Any user can serve as an advisor if he or she chooses to record their feelings about a document based on the key words they choose. If another user wishes to receive advice about a document, he or she seeks helps from other potential advisors who have ranked the document.

The requirement for inputting users' relationships makes Tapestry rather limiting in making inferences. To counter this problem, the GroupLens system attempts to automate the processing, by employing an automated collaborative filtering system (Konstan, Miller, Maltz, Herlocker, Gordon, & Riedl, 1997). Its central idea is to apply a statistical collaborative filter to automate the inference of relationship. The GroupLens system identifies advisors based on the *Pearson correlation* of voting history between pairs of users, which measures the degree with which the rating histories of two users are linearly correlated. For example, consider the three users in Table 5. In this table, users one and two have both ranked all three movies. User three ranked the first two movies *Star Wars* and *Pearl Harbor,* but has not watched *AI: Artificial Intelligence*. What should we assign as a predicted rating for *AI: Artificial Intelligence* for the third user based on the other users' ratings? This is the problem of collaborative filtering.

To capture this intuition formally, let $r_{u,i}$ be a rating of user u on product item i. Let $\bar{r}_u$ be the average rating of user u in history. Let $w_{a,u}$ be the Pearson correlation between users a and u, and let α be a normalizing factor. Then the prediction for user a's rating on item i is $P_{a,i}$.

$$P_{a,i} = \bar{r}_a + \frac{1}{\alpha}\sum(r_{u,i} - \bar{r}_u)\cdot w_{a,u}.$$

The calculation of the correlation weight factor $w_{a,u}$ is based on the formula below.

$$w_{a,u} = \frac{1}{|items|}\sum_{items}\frac{(r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{items}(r_{a,i} - \bar{r}_a)^2}\sqrt{\sum_{items}(r_{u,i} - \bar{r}_u)^2}}.$$

For example, in the Ringo (later called Firefly) system, a statistical collaborative filtering system recommendations for music albums are computed based on variations of the Pearson correlation formula and the high-ranking albums are provided for users. The result is more interested users

Table 5 User Ratings for Movies

	Ratings for Movies		
User ID	**Star Wars**	**Pearl Harbor**	**AI: Artificial Intelligence**
1	5	2	4
2	2	5	1
3	5	4	?

and higher sales for the company. Built on the same key idea, Breese, Heckerman, and Kadie (1998) experimented with a Bayesian classifier that computes the probability of membership in an unobserved class *c* based on the assumption that ratings are conditionally independent. This model can both learn the users ratings on items and users classes that are hidden. To learn the hidden class labels, an EM algorithm was used to learn the model structure with a fixed number of classes. Through experiments, the researchers have found that the Bayesian classifier takes longer time to train but have comparable accuracy in predictions.

Speeding Up Collaborative Filters

A major problem with both the Pearson correlation and Bayesian methods is that they take a long time to compute a prediction for a new user. In Chee, Han, and Wang (2001) a partition-based method is designed to speed up Pearson correlation with very large user and product databases. A tree is constructed based on a clustering algorithm, where each leaf node of the tree contains a subset of the users that are most similar to each other. Using this idea, the computation efficiency for Pearson correlations is greatly enhanced. Experiments show that the system can scale linearly with the size of the data and can produce accurate results even when the data are very sparse, which is a common problem with electronic commerce applications where the number of products and users are extremely large.

Personalization Systems

The correlation-based recommendation system is one example of personalization services on the Web. Other types of personalization services have also been extensively studied. Basu, Hirsh, Cohen, and Nevill-Manning (2001) described a system for recommending technical papers for paper review committee members so that the recommendations fit their interests. Their approach uses multiple information sources such as the information from the reviewers' home pages. Many recommendation systems have been built based on similar ideas. For example, the *NewsDude* system which we discussed earlier uses was constructed based on user preferences that correspond to both long-term and short-term user interests. Both types of interests can be learned using supervised learning algorithms. These interest models can be applied to articles that are downloaded online. Experiments have shown that they are very helpful in filtering out a large number of unwanted articles for its users.

CONCLUSION

In this chapter, we have given a focused survey on machine learning and data mining on the Web and for the Web. We have chosen to consider four related areas of machine learning and data mining for Web content, Web structures, Web usage, and Web recommendations. The amount of research activity is increasing at a very rapid rate. Many systems are not covered in this short chapter, but they can be found in various sources. Many research papers can be found from the research index Web site http://citeseer.nj.nec.com.

GLOSSARY

Association rules An IF-Then rule such that when the left-hand side of the rule holds, the right-hand side of the rule also holds with high probability.

Clustering A data mining and machine learning technique that groups similar items together.

Collaborative filtering A machine learning method to learn a user's preferences by taking into account preferences of similar users.

Data mining The process of applying machine learning to databases in order to discover interesting and useful patterns from the data.

PageRank A score assigned to a Web page to designate its importance in being a hub (contains many hyperlinks to other important pages) or authority (pointed to by many other important pages).

Supervised and unsupervised learning Techniques used to learn the relationship between independent attributes and a designated dependent class attribute (the label). Supervised learning (for example, naïve Bayesian, decision trees, and rule induction) requires the class-labeled examples as input, whereas unsupervised learning (for example, clustering) does not.

Web log mining Learning users' browsing behavior from Web log data using sequential learning methods such as Markov models and sequential association rules.

Wrapper A mapping in the form of a set of rules between semistructured texts such as HTML to structured contents.

CROSS REFERENCES

See *Data Mining in E-Commerce; Data Warehouses and Data Marts; Knowledge Management.*

REFERENCES

Argawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, & C. Zanido (Eds.), *Proceedings of the 20th International Conference on Very Large Databases* (pp. 487–499). San Mateo, CA: Morgan Kaufmann.

Aggarwal, C. Wolf, J. L., & Yu, P. S. (1999). Caching on the World Wide Web. In *IEEE Transactions on Knowledge and Data Engineering, 11,* 94–107.

Albrecht, D., Zukerman, I., & Nicholson, A. (1999). Presending documents on the WWW: A comparative study. In *IJCAI99* (pp. 1274–1279). Sweden.

Anderson, C. R., Domingos, P., and Weld, D. S. (2001). Personalizing web sites for mobile users. In *Proceedings of the 10th International WWW Conference,* 2001. Retrieved from http://www.www10.org/

Basu, C., Hirsh, H., Cohen, W. W., & Nevill-Manning, C. (2001). Technical paper recommendation: A study in combining multiple information sources. *Journal of AI Research, 14,* 231–252.

Breese, J. S., Heckerman D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)* (pp. 43–52). San Mateo, CA: Morgan Kaufmann.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 284*(5), 34–43.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Journal of Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Chee, S., Han, J., & Wang, K. (2001). RecTree: An efficient collaborative filtering method. In Y. Kambayashi, W. Winiwarter, & M. Arikawa (Eds.), *Proceedings of the 2001 Data Warehousing and Knowledge Discovery, Third International Conference* (pp. 141–151). New York: Springer.

Cao, P., & Irani, S. (1997). Cost-aware WWW proxy caching algorithms. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems* (pp. 193–206). Berkeley, CA: USENIX.

Doan, A., Domingos, P., & Levy, A. (2001). Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD Conference* (pp. 509–520). New York: ACM Press.

Dumais, S. T., Cutrell, E., & Chen, H. (2001). Bringing order to the Web: Optimizing search by showing results in context. In *Proceedings of CHI'01, Human Factors in Computing Systems* (pp. 277–283). New York: ACM Press.

Etzioni, O. (1996). The World Wide Web: Quagmire or gold mine? *Communications of the ACM, 39*(11), 65–68.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry, *CACM 35* (12), 61–70.

Han, J., & Kamber, M. (2001). *Data mining—Concepts and techniques*. San Mateo, CA: Morgan Kaufmann.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: MIT Press.

Hsu, C., & Dung, M. (1998). Generating finite-state transducers for semi-structured data extraction from the Web. *Information Systems, 23*(8), 521–538.

Joachims, T., Freitag, D., & Mitchell, T. (1997). WebWatcher: A tour guide for the World Wide Web. In *Proceedings of The 15th International Conference on Artificial Intelligence* (pp. 770–777). Somerset, NJ: IJCAI Inc.

Kaelbling, L. P., Littman, L. M., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research, 4,* 237–285.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632.

Knoblock, C. A., Lerman, K., Minton, S., & Muslea, I. (2000). Accurately and reliably extracting data from the Web: A machine learning approach. IEEE *Data Engineering Bulletin, 23*(4).

Kohavi, R., Brodley, C. E., Frasca, B., Mason, L., & Zheng, Z. (2000). KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations, 2*(2), 86–98.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM, 40*(3), 77–87.

Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence Journal, 118* (1–2), 15–68.

Kushmerick, N., Weld, D., & Doorenbos, R. (1997). Wrapper induction for information extraction. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 729–737). Somerset, NJ: Morgan Kaufmann.

Liu, B., Ma, Y., & Yu, P. (2001). Discovering unexpected information from your competitors' Web sites. In *Proceedings of the Seventh ACM Knowledge Discovery and Data Mining Conference* (pp. 144–153). New York: ACM Press.

Lou, W., Liu, G., Lu, H., & Yang, Q. (2002). Cut-and-pick transactions for proxy log mining. In *Proceedings of the 2002 Conference on Extending Database Technology.* Berlin: Springer.

Mitchell, T. (1997). *Machine learning*. New York: McGraw-Hill.

McCallum, A., & Nigan, K. (1998). A comparison of event models for naïve Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*. Menlo Park, CA: AAAI Press.

Perkowitz, M., & Etzioni, O. (2001). Adaptive Web sites: Concept and case study. *Artificial Intelligence, 118* (1–2), 245–275.

Pazzani, M. J., Muramatsu, J., & Billsus, D. (1996). Syskill & Webert: Identifying interesting Web sites. In *Proceedings of the American Association of Artificial Intelligence* (pp. 54–61). Menlo Park, CA: AAAI Press.

Pitkow, J., & Pirolli, P. (1999). Mining longest repeating subsequences to predict World Wide Web surfing. In *Second USENIX Symposium on Internet Technologies and Systems,* Berkeley, CA: USENIX.

Schechter, S., Krishnan, M., & Smith, M. D. (1998). Using path profiles to predict HTTP requests. *Journal of Computer Networks and ISDN Systems, 30* (457–467).

Shardanand, U., & Maes, P. (1995). Social information filtering: Algorithms for automating "word of mouth." In *Proceedings of 1995 ACM Conference on Human Factors in Computing Systems:* Vol. 1 (pp. 210–217). New York: Elsevier.

Spiliopoulo, M., & Pohle, C. (2001). Data mining for measuring and improving the success of Web sites. *Data Mining and Knowledge Discovery Journal, 5*(1–2), 85–114.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations, 1*(2).

Su, Z., Yang, Q., Zhang, H., Xu, X., & Hu, Y. (2002). Correlation-based Web-document clustering for Web interface design. *International Journal Knowledge and Information Systems, 4,* 141–167.

Su, Z., Yang, Q., & Zhang, H. (2000). A prediction system for multimedia pre-fetching in Internet. In *Proceedings of 2000 International ACM Conference on Multimedia,* New York: ACM Press.

Witten, I., & Frank, E. (1999). *Data mining—Practical machine learning tools and techniques with Java implementations.* San Mateo, CA: Morgan Kaufmann.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval, 1* (1/2), 67–88.

Yang, Q., Zhang, H., & Li, I. T. (2001). Mining Web logs for prediction models in WWW caching and prefetching. In *Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'01.* New York: ACM Press.

Zaïane, O., Xin, M., & Han, J. (1998). Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In *Proceedings of ADL'98 (Advances in Digital Libraries)* (pp. 19–29). Los Alamitos, CA: IEEE Computer Society.

Managing a Network Environment

Haniph A. Latchman, *University of Florida*
Jordan Walters, *BCN Associates, Inc.*

INTRODUCTION TO NETWORK MANAGEMENT

Computer networks and their interconnections with the Internet have now become integrated into the business, academic, government, and personal environment. To grasp the impact of this complete saturation, we need only to consider the network centric processes used to perform everyday functions in daily life. Purchases, investments, research, education, interaction with governmental agencies, personal communication, and many other activities are now routinely handled or enhanced by using the World Wide Web (WWW). Even though this integration is virtually seamless in today's society, if we dissect the systems used to accomplish these tasks, we find a complex "network of networks" at the core. However transparent these tasks are to end users of current technology, without competent management of the underlying interconnected networks, none of these actions would be possible. In this chapter, we discuss the fundamental design issues and operational techniques, as well as the "best practices" that are involved in managing and maintaining a network environment.

BENEFITS OF A NETWORK ENVIRONMENT

Before we get to the specifics of how networks are maintained, it is instructive to consider the root purpose of computer networks. If computer networks are deployed and used arbitrarily without regard to the real-world benefits obtained from their usage, then personal computers (PCs) and networks would become little more than "gadgets" that may complicate rather than simplify a person's life or a company's operation. For computer networks to be truly useful there must be a net gain in terms of overall efficiency, financial impact, or functionality in the completion of a given task when compared to other traditional methods of undertaking the same task, or, alternatively, there should be some innovation or creation of new resources not previously available without computer networks.

Return on Investment

The positive impact of computers and the associated networked environment on overall business and corporate functions, when properly integrated with sound, traditional business practices, is well documented. As a simple example, assume that corporation XYZ Inc. has numerous PCs connected to a common network. XYZ Inc.'s employees can instantly share or transmit large and complicated documents or other pieces of information across the network, rather than delivering this information by other traditional means. (The term "sneakernet" was coined to describe a network of persons who share information by hand delivery of documents or other information. A properly configured PC network eliminates the need for such a "network.") Thus the use of a well-designed computer network can result in significant improvement in overall efficiency for important organizational functions. As another example, when considering how to allocate the budget for equipment purchases, XYZ Inc. can better leverage resources by obtaining a few shared printers strategically located throughout its office rather than acquiring a separate printer for each of its several hundred employees. Or, XYZ Inc. could procure a few advanced printers, thereby maximizing resources for increased functionality and quality.

Innovations and Wider Reach

Another possible objective to be satisfied by deploying a computer network would be the ability to produce unique, innovative products or revolutionary ways to deliver services that makes existing products more appealing or more accessible to consumers. For instance, assume that XYZ Inc. sells auto insurance. Because of its Internet presence, XYZ Inc. can now sell insurance plans automatically via its e-commerce enabled Web site. Auto insurance is not a groundbreaking product, but the new way of offering this product makes XYZ Inc.'s services more accessible to potential clients. Now anyone in the world can do business with them, instead of just people in their local community, and the transaction is easier for the customer, who can purchase insurance any time of day or night, rather than having to meet with an agent in person during business hours.

PRINCIPLES OF NETWORK MANAGEMENT

Since the purpose of using a computer network in the first place is to achieve a net gain of resources or functionality, the methods used to manage a network should reflect these same ideals, so that maximum performance of the network can be realized while minimizing the effort required for management. To illustrate, suppose XYZ Inc. uses its computer network to efficiently share information and deliver services. The information technology (IT) department of XYZ Inc. keeps the network running at peak performance. However, the IT department must employ an inordinately large staff to handle the management load, since the network was designed poorly, without due consideration to its maintenance, and therefore the efficiencies gained by the good performance of the network are outweighed by the huge cost incurred for network management and maintenance. Even though the network functions as it should, the root purpose of the network is not realized, since there is no net resource or functionality gain in the overall operation of the organization. So, when we form a management strategy, our governing principles must be to maximize network availability, performance, and functionality while minimizing resources needed for maintenance and management.

Network Management By Design

The *management* of a computer network may be viewed as being a separate process from the *design* of a network, since in many cases the design of a network requires a higher level of skill than does the management. For example, there are separate competency certifications for "administration" of a Microsoft network (Microsoft Certified Systems Administrator) as opposed to the "engineering" or designing of such networks (Microsoft Certified Systems Engineer). To conserve resources, a firm may choose to outsource the design of their corporate network and retain a network manager to perform the day-to-day operations of the system. Even though network design and management are often separated as just described, it can be said that the management of a network begins with its design. A network may be designed with good performance in mind, but without good planning, the resources required to manage the resulting network may be prohibitive. On the other hand, a network may be designed to be extremely easy to manage, but not to deliver the performance or functionality desired by the organization. Therefore, we assert that good network management begins with a good design that accounts for acceptable performance and functionality while allowing for relative ease of maintenance.

There are an ever-growing number of factors to consider when a network is designed and an effective network management plan is developed. Emerging technologies may render inconsequential considerations that may have been essential in the recent past, or new considerations may become apparent where previously they were unimportant. For example, until quite recently, options were limited for high speed connectivity between sites separated by great distances. Because the only feasible options for remote connectivity were relatively slow links (56 kbps or 128 kbps integrated digital services network [ISDN], for instance), network infrastructure and network management strategies and applications had to be designed to accommodate these slow links. Now, however, faster connections are more affordable and available (digital subscriber lines [DSLs], for example), so exchanging information across the telephone line is no longer a major constraint. Remote management of technology systems has therefore become much more viable, thereby allowing a single IT professional to quickly troubleshoot problems from miles away (Newsome, 2002). Since network design and network management considerations change so rapidly, we will attempt to generalize these issues into a few essential categories, and then focus on how these concerns relate to the management of networked environments. The discussion that follows first focuses on local area networks (LANs), then on LAN interconnections, and finally on Internet mediated wide area networks (WANs).

LOCAL AREA NETWORK CONNECTIVITY

Network Management Issues in LAN Technologies

When you decide on a LAN connectivity solution for a particular computer network, there are many choices available. However, depending on the particular installation, these choices may be limited or some choices can be eliminated. For example, in an environment where there could potentially be a high amount of electromagnetic interference (EMI) such as a power plant, the only functional choice may be a connectivity solution that is immune to EMI, possibly using fiber optic technology rather than twisted copper wire. Or, in a building that is not conducive to a wire-based installation because of a unique interior decoration scheme (such as in a historical building where modification is prohibited) a wireless solution may be the only choice. In cases where such restrictions exist, the management techniques will be defined by the requirements of the given solution. However, in cases where these restrictions are not inherent, a network may be designed with acceptable performance, nominal cost, and minimal management effort in mind. Some issues impacting the overall management of the resulting networked environment include speed, ease of installation, upgradability, cost, and compatibility with legacy systems.

Ethernet—The De Facto LAN Standard

The most popular LAN connectivity solution is Ethernet—operating at speeds of 10 Mbps to 1 Gbps and based on the Institue of Electrical and Electronic Engineers (IEEE) 802.3 standard. Although initially deployed using coaxial cable (10Base-5 or 10Base-2), star-wired 10Base-T hubs have been the most successful Ethernet LAN technology, and the most attractive from a network management point of view. In 1996 it was estimated that 83% of all LANs were based on the Ethernet standard, representing some 120 million computers, and in 1997, 77% of all networking equipment shipped was Ethernet based (Stallings, 2002). There are several reasons for the strong support for Ethernet as the LAN technology of choice. The most important of these from a network management perspective include the overall network reliability, scalability from 10 Mbs up to speeds of 1000 Mbps, ease of management and troubleshooting, as well as low cost and almost ubiquitous availability from all manufacturers.

Classic Ethernet and Fast Ethernet

10/100Base-T Ethernet, operating at 10/100 Mbps over twisted pair cable, has virtually become the de facto LAN standard for corporations of all sizes because it satisfies key network management requirements in an extremely effective way. Ethernet provides good throughput performance for most LAN applications, it is easy to install (aside from the actual wiring, implementing this solution is almost as simple as plugging in a telephone), it is easy to upgrade (adding nodes is accomplished by simply purchasing additional hubs or switches and plugging in additional cables), it is a relatively inexpensive solution compared to other currently available technologies, and it benefits from compatibility and availability of hardware and software support (a multitude of vendors offer products which are compatible with this standard).

Classic Ethernet (10Base-T) and fast Ethernet (100Base-TX) (Johnson, 1995; Wolf, 1999) are now predominantly found in most LANs, together with a variety of intelligent hubs and switches providing half-duplex and full-duplex operation at 10 Mbps up to 200 Mbps. The network manager is responsible for selecting the network components (network interface cards [NICs], hubs, and switches) to ensure that the network is easily managed and that a clearly defined migration path exists for moving from 10Base-T to higher bandwidth networks. For example, this may require the replacement of a 10Base-T hub in a network group with a 10/100Base-T Ethernet switch. The NICs in this work group would still operate at 10 Mbps, while the switch could support 100 Mbps NICs automatically, provided that the NICs support autosensing and negotiation. In addition the switched operation localizes high intensity traffic pairs to their own switched segment thus improving overall LAN performance (Schuyler, 2001). The network manager would also want to ensure that the high quality CAT-5 twisted cabling is used even for classic 10Base-T networks, since this same cabling could then be usable for 100 Mbps and even 1 Gbps operations. In this migration the network manager would also need to ensure that the overall segment lengths and number of total repeaters do not violate the rules for proper operation of the IEEE 802.3 CSMA/CD protocol on which all flavors of Ethernet are based.

Innovative LAN Technologies for Home and Small Office Networks

Other options available for local networks in small offices and home applications include the 802.11× wireless LANs and the power line and phone line LANs (Brown, 2002; HomePlug, 2002). While these alternatives for connecting computers and peripherals are not well known in the corporate environment and may not compete favorably with industry standard, specifically designed, wired networks, they do provide some options for retrofitting existing buildings with network connectivity with no new wires—namely deploying a purely wireless system or using the existing telephone or even electrical wiring as the data communication channel. In 2002, the HomePlug Alliance released its first power line products based on the HomePlug 1.0 specification. These devices provide Ethernet and universal serial bus (USB) bridges to the power line communication bus, operating at speeds of up to 14 Mbps with virtually complete coverage in typical homes (Karagiannis, 1999). The power line LANs are thus quite attractive and competitive with the better known wireless LANs, which operate at speeds of 11 Mbps (802.11b) or 54 Mbps (802.11a) but with somewhat less reach using a single wireless access point. Even with multiple access points, the wireless solutions are less attractive because of the need for a wired infrastructure to interconnect the access points. Powerline LANs have the advantage of already having many connection points (electrical outlets) in almost every room in the home, while phone line LANs (and even cable-based connectivity) have limited connection points in only a few select rooms (Lin, Latchman, Lee, & Katar, 2002; Latchman & Yonge, 2003).

The deployment of new technologies such as wireless or power line LANs, always triggers the question of network management and maintenance. Network managers are hesitant to add newer and less well-understood technologies into the network since user support and management will certainly become more of a challenge. On the other hand, the designers of consumer grade wireless and power line LANs, acutely aware of the inexperienced personnel in home and small office networking environments, have attempted to make the new products very easy to install and manage. Even a seemingly complex matter such as enabling encryption for data and network security is easily handled by novice home users.

LAN Interconnections and Campus Area Networks (CANs)

Interconnection of LANs within a local enterprise that consists of several buildings form what is called a campus area network (CAN). Among the major contenders for backbone LAN interconnection technologies are fiber distributed digital interface (FDDI), asynchronous transfer mode (ATM), fast Ethernet, and gigabit Ethernet (Atkins & Norris, 1998). Again in this realm of operation, gigabit Ethernet is becoming a major contender, partially because of the overall familiarity with the lower speed flavors of Ethernet and the fact that the frame format is virtually identical (McLeod, 2002). While FDDI, ATM, and gigabit Ethernet all support the use of multimode or single mode fiber links, gigabit Ethernet also supports the 1000Base-TX physical layer using four pairs of CAT-5 in conjunction with a sophisticated five-level coding scheme on each pair. In this way, gigabit Ethernet may be implemented using existing CAT-5 cabling by simply installing new gigabit Ethernet switches (Kadambu, Kalkunte, & Grayford, 1998). Gigabit Ethernet is also substantially less expensive than their ATM and FDDI counterparts. It is also easier to manage since personnel who already understand classic and fast Ethernet can readily be trained in extending their expertise to this higher speed incarnation of a standard they already know well—often simply involving the use of new maintenance and troubleshooting tools.

In addition to wired CANs and LANs, it is possible to provide alternative or complementary wireless LAN and CAN solutions using a network of wireless access points (McKenzie, 2001). The proliferation of portable computing devices such as notebook and palm computers, personal digital assistants (PDAs), and even personal area networking devices, has resulted in greater demand for mobile access to network resources. In this regard, the issues of network and information security and encryption become urgent concerns in the management of the networked environment.

WIDE AREA NETWORK (WAN) CONSIDERATIONS

Networks of computers within an office complex or even some homes share resources within the LAN but users may also need to exchange information with other networks separated by possibly large distances. For example, branch offices in regional, national, or international corporations need to communicate with each other and with the headquarters on a regular basis, often instantaneously, in order to increase the overall efficiency of the organization as a whole.

Wide area networks (WANs) provide such long-distance interconnections of LANs. WANs are distinguished from campus area networks (CANs) or metropolitan area networks (MANs) in that the CANs connect several buildings in a campus setting (such as a university) and MANs provide a high speed data backbone for a large city, while WANs may cover much larger geographical regions and use different telecommunications and networking technologies to accomplish the interconnection.

The largest and best known WAN is the "mother of all networks"—the Internet, a network of networks. Thus network managers and designers often consider establishing their corporate WANs by interconnecting each network in their multiple branch offices to the Internet with a local Internet link, and then using privacy mechanisms to create what has become known as a virtual private network (VPN) (Harris, 2002). In other corporations, however, network managers insist on the establishment of a completely private WAN with dedicated leased telecommunications links between the various nodes (LANs) in their network. While inherently more costly, this option provides many advantages such as privacy and guaranteed performance when compared with shared access networks such as the Internet. Other smaller organizations simply need their single LAN to be interconnected with the Internet to exchange electronic mail (e-mail) with clients and suppliers and for e-commerce and accessing the World Wide Web for research or even to publish its own WWW content.

Options for Internetwork Connections

Interconnections between networks and to the Internet exploit the embedded telecommunications infrastructure used for circuit switched voice communication for more than a century now (Morrison, 2002). The difference is that the telecommunications infrastructure is now being used for exchanging packets of data (rather than voice only) and the links are thus often shared by many users. Users of the data network have their own packets encapsulated within "envelopes" bearing their addresses and other pertinent information to keep the information separate in transit, as well as to route the packets to the correct destination.

The public switched telephone network (PSTN) operates on the principle of circuit switching in which a complete end-to-end channel is established and held for the entire duration of the information interchange (such as a telephone conversation). Initially data services were offered using the same PSTN with adapters (modulators/demodulators, modems) for interfacing the digital data with the PSTN. Data services are more efficiently served by a packet data network (PDN) in which segments of data are routed progressively through a digital network without the concept of a circuit and with all links in the network shared by other data traffic. Examples of PDNs are X.25, and its successor frame relay, as well as the Internet itself. Traditional telecommunications links are still used to interconnect nodes in these networks but the traffic streams which traverse these networks now all share the capacity of the links.

Using the Local Telephone Line for Network Connections

The simplest and most common connection to the Internet (or from one network to another) is obtained by using a pair of analog modems connected via a voice (plain old telephone service [POTS]) line. Other variations of internetworking connections using telephone lines are the use of integrated digital services network (ISDN) and asymmetric digital subscriber lines (ADSL) connections. Both of these latter technologies use the same single pair of copper wires that provide telephone service to residential and business customers, but require special conditioning of the line to maximize the overall bandwidth, and also have restrictions on the total distance from the local exchange for which the service will work. ISDN uses a digital scheme called time division multiplexing (TDM) to provide three independent digital channels, two operating at 64 kbps and one at 16 kbps. One of the 64-kbps channels could then be used for POTS voice service and the other for dial-up or even full-time data services. The two 64-kbps channels may also be linked together to form a 128-kbps data channel. The 16-kbps channel is used for signaling and low speed data. ADSL uses frequency division multiplexing to allow the lower 4 KHz of bandwidth to continue to be used for POTS voice and then the upper bandwidth to be used for high speed (broadband) data with asymmetric speeds of some 9 Mbps to the user and 1.5 Mbps from the user.

Dial-Up Connections to the Internet

Let us imagine that a network manager has designed a 100-Mbps Ethernet-based network for an organization using hubs and switches, with about 100 computers sharing internal resources such as printers and inbound and outbound facsimile services. The network users are also able to send e-mail within the network, and the corporate Web site is outsourced to an external hosting company. Now the need is simply to provide inbound and outbound Internet e-mail services to all users in a seamless and inexpensive manner.

In this case the network manager could provide a plain old telephone service modem operating at 56 kbps to each employee who would then use a telephone line to dial-up to an Internet service provider to get access to e-mail and the WWW. Clearly this would be a very inefficient solution and would also open up security and monitoring problems for the network manager. Alternatively the network manager could set up an e-mail gateway on the LAN that would accept outgoing Internet e-mail and receive incoming e-mail for the users on the LAN, and that gateway machine could then itself be configured to make a 56-kbps dial-up connection periodically to retrieve and send Internet e-mail. This is a very low cost solution that provides e-mail connectivity as required. Depending on the security concerns with regard to interconnecting the corporate LAN to the Internet, one could even design a “stand-alone” mail exchanger with a single phone line that would alternately dial to the Internet to send and retrieve e-mail, and then itself be polled by a dial-up connection from a network connected computer to synchronize mail with the corporate LAN. In this way the corporate LAN would never be directly connected to the Internet.

To Connect or Not to Connect to the Internet?

Clearly there is a trade-off between the need to exchange information between the corporate LAN and other networks such as the Internet and the concern over network and information security and privacy. The option just outlined for using a single dial-up connection has been used successfully by relatively small organizations wanting to exchange Internet e-mail in a fairly secure and low cost manner. However, when the volume, frequency, and size of Internet e-mail are large, intermittent dial-up connections at 56 kbps tend to be inconvenient and slow, especially when there are large file attachments. Moreover, it is often desired to provide e-mail and WWW hosting and WWW browsing solutions in-house, rather than by outsourcing to external providers, and in these cases the network manager must decide on an appropriate scheme for establishing a dedicated connection between the corporate LAN and the Internet.

Dedicated Internet Networking Connections

Dedicated Internet connections, operating at speeds ranging from 64kbps to 10 Mbps, are available, via cable TV distribution systems, ADSL, satellite and even local wireless connections between the user and a nearby provider. However, for most corporate applications dedicated internetwork connections are provided by means of high capacity digital telecommunications circuits of the type normally used for interconnecting telephone exchanges or switches. The simplest of these is the digital signal level-1 (or DS-1) link, which is essentially composed of 24 time division multiplexed digital voice channels (DS-0) each operating at 64 kbps, thus making the overall data rate of the DS-1 signal about 1.544 Mbps, including management overhead. Most small- to medium-sized corporations are connected to the Internet and/or branch offices via a dedicated DS-1 infrastructure, with either multiple DS-1 links for higher capacity or fractional DS-1 capacity in multiples of 64 kbps as needed. Higher capacity links are also available with DS-2 (the equivalent of 4 DS-1s) operating at 6.312 Mbps), DS-3 (the equivalent of 7 DS-2s) operating at 44.736 Mbps, DS-4 (the equivalent of 6 DS-3s) operating at 274.176 Mbps, and DS-5 (the equivalent of 2 DS-4s) operating at 560.160 Mbps (Latchman, 1997). Depending on the distances involved, these connections can be provided over copper facilities (usually multiple pairs) or over pairs of optical fiber for longer distances. In addition to the DS-x hierarchy there is an optical communications hierarchy specification as shown in Table 1.

Clearly a network manager would need to carefully weigh the pros and cons of having a dedicated network connection between the corporate LAN and the Internet or branch offices, and whether to outsource some services to third party providers. If the decision is to establish a dedicated connection, the manager would need to consider the volume of traffic likely to be exchanged over the proposed links and factor concerns such as costs, quality of service, reliability, and security to make the final decision as to the type of connection to be supported. An alternative to fully dedicated interconnections is to use cell or frame services provided by some network operating companies to establish partially private networks using a shared digital infrastructure. Services in this class include frame relay services and asynchronous transfer mode.

Table 1 Optical Communications Hierarchy

Optical Carrier Level	Speed (Mbps)
OC-1	51.84
OC-3	155.52
OC-9	466.56
OC-12	622.08
OC-18	933.12
OC-24	1,244
OC-36	1,866
OC-48	2,488
OC-96	4,977
OC-192	9,953
OC-256	13,271
OC-768	40,000

Security Concerns and Virtual Private Networks

As mentioned earlier, network managers often design their global corporate network using the Internet as the shared interconnection network, rather than the more costly private network consisting of dedicated long-haul digital telecommunication links. While interconnecting branch offices and the corporate headquarters via the Internet has some attractive features, there are also serious security and performance concerns due to the shared nature of such connections. Indeed, some large network service providers, while contributing to the global Internet connectivity, offer global interconnection nodes on a restricted "internetwork," which is not shared by regular Internet traffic, thus guaranteeing better performance from the point of view of potential network congestion, delays, and packet loss. Another major concern is the matter of information and network security when corporate network resources are connected to shared networks such as the Internet.

Figure 1 shows a typical corporate network with a large infrastructure at the headquarters location as well as multiple satellite offices and mobile users connected back to the headquarters' resources via dedicated and dial-up Internet connections. This scenario is now quite common and serves to illustrate some of the key challenges in managing a network environment spanning multiple geographically disparate locations and that also caters to mobile users. Since the satellite offices and traveling users are connected to the corporate network via the Internet, a key concern will be the issue of security as proprietary information traverses the global Internet. Figure 1 shows an encrypted tunnel through the Internet forming a virtual private network (VPN) between all satellite offices and itinerant users.

This encrypted tunnel is formed by invoking industry standard public key cryptography schemes at the packet level, between the interface nodes on both networks interconnected via the public network. In this way information can be seamlessly sent and received securely without the need for application level encryption. Figure 2 illustrates the difference between a VPN infrastructure and individual branch offices directly connected to the Internet.

NETWORK OPERATING SYSTEMS (NOS)

In addition to augmenting functionality and performance for corporate networks, network operating systems often include features that minimize the time needed to manage a network. Some of these features include authentication, resource access, application delivery and installation, and permission and policy management. Having authentication handled by a NOS can save time for the network manager by allowing the manager to create a single account for users that can be used from any workstation in the organization, rather than having to create authentication for each individual workstation. Managing resource access can be facilitated by a NOS in that users can be granted access to network resources (files, printers, applications, etc.) as groups rather than as individuals. Applications can be set to install automatically from a server for an entire organization when triggered by user actions (log-in, for example), so that a network manager need not physically go to each workstation to install software updates. Networkwide permissions and policies can be managed by assigning individuals to groups. In this way, a network manager can grant permission to users to perform some action (change a password, for example) by simply adding the user to a group with those permissions. Network managers configure permissions for a few broad groups instead of for each user individually. In choosing a NOS that will be both functional and manageable, it is good to evaluate the features available from various NOSs to see which will lend the most assistance to a particular network implementation. To help illustrate some management approaches that can be adopted with different NOSs, we will give an overview of some common NOSs available.

Peer-to-Peer

In a peer-to-peer system, each node on the network acts as both a client (a user of some service) and a server (a deliverer of a service). Some examples of peer-to-peer NOSs are Windows 95/98/ME. Peer-based systems are often used in very small networks where funds and resources are limited. In a peer-based system, users can access files, printers, and other services from other workstations, but there normally is no central authentication system (in order to access resources, users must have an account on each peer). Advantages of this type of system include ease of setup and maintenance (for a small number of users), low cost (generally, peer-to-peer network functionality is built into the desktop operating system so additional software is not necessary), and good fault tolerance (since there is no single point of failure).

When peer-based networks are managed, it is important to create some structure to define the manner in which resources are accessed, since no logical structuring is done for you. Necessary decisions must be made as to how users' files will be stored (centrally or in a distributed manner), how user accounts will be managed (created as needed per workstation or all at once), how permissions

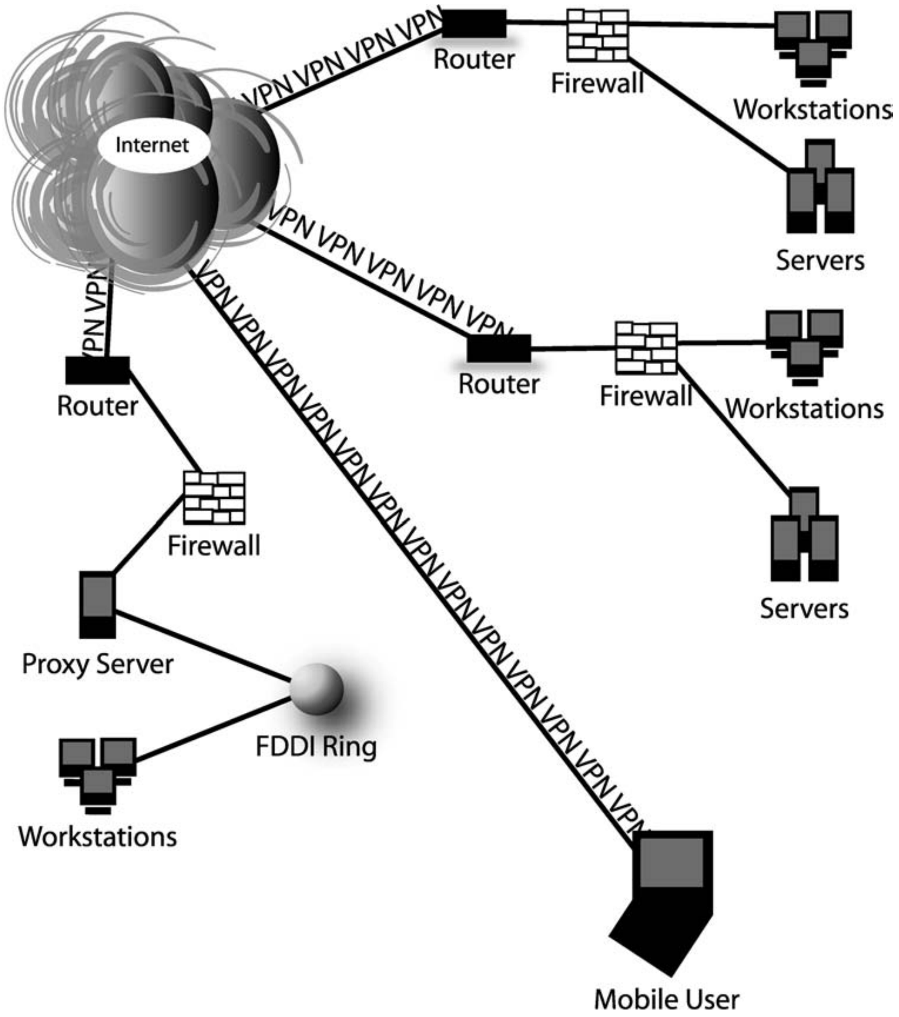


Figure 1: Virtual private network for network interconnection.

will be handled, and how updates and maintenance will be performed on workstations.

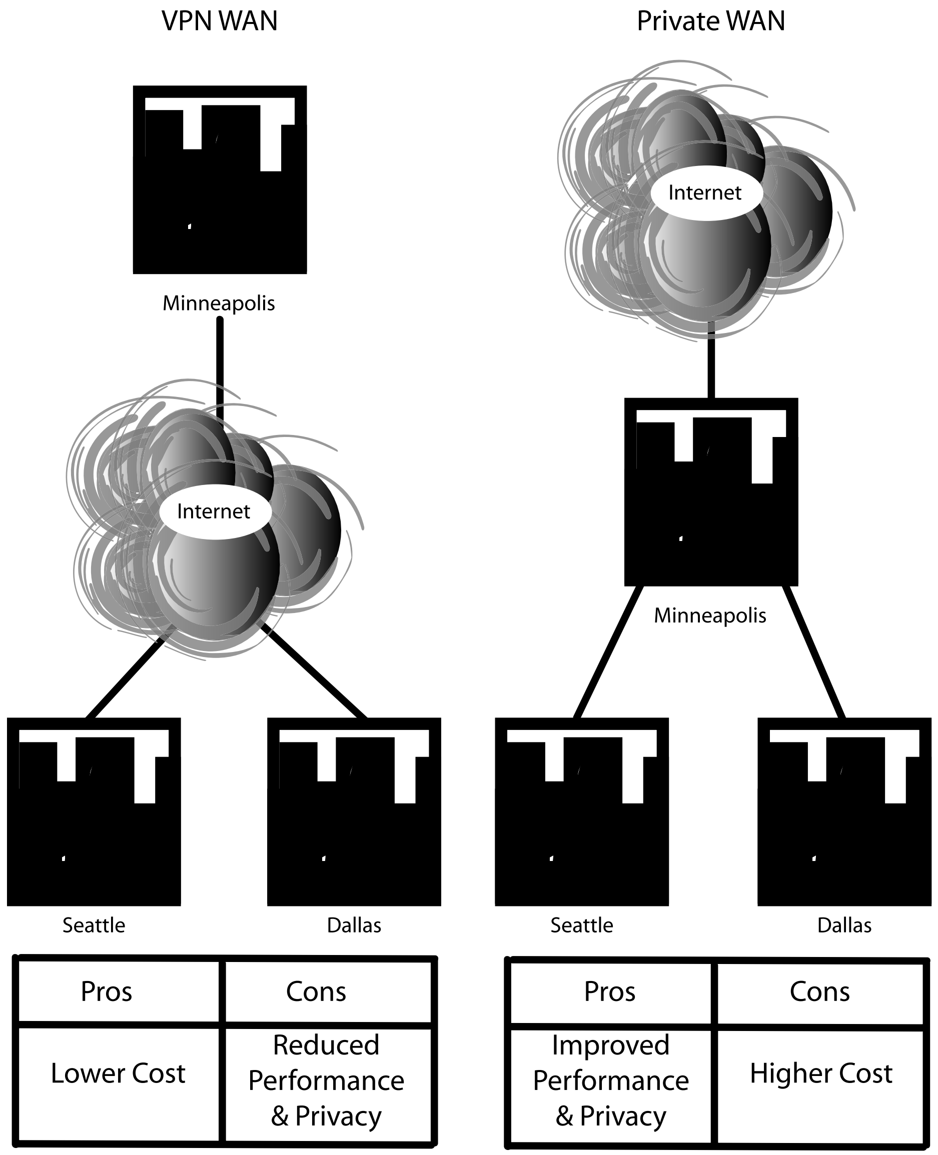
Client/Server

Client/Server NOSs are common for corporate networks where there are more than 10 users. Some examples of client/server NOSs are Windows NT/2000 Server and Novell Netware. Client/Server systems rely on the processing power of both the client and server nodes to some extent, although networkwide services (file storage, printing, messaging, authentication, etc.) are provided by the server (Spanbauer, 1999). Normally, each client node (workstation) has a small software program that runs when the workstation boots up and that will allow the end user to authenticate (log in) to the server.

There are many advantages of a client/server system from the perspective of network management. For example, user accounts can be created in one place (the server) so that users can use the same log-in credentials no matter what node they use to log-in. Wherever a user logs in, the network manager can define triggered actions in a log-in script for installing or updating applications, updating workstation configuration, delivering important instructions or messages, updating virus definitions, or virtually any other task the manager chooses.

Mainframe/Thin Client Systems

Whereas client/server systems involve processing on both the client and server sides, mainframe and thin client systems use the processing power of the server almost exclusively. Although some type of software on the client side is still necessary, normally the only processing done on the client side is whatever is necessary to connect to the server. Instead of using processing power to run applications, the client just does enough to exhange user input (keystrokes, mouse movements, etc.) and user output (screen displays) between the client and server. Although we are grouping mainframe and thin client systems together, it could be said that mainframe systems are a subset of thin client systems. Any system where application processing is done on the server side could be termed a "thin client" system, and mainframe systems



Pros	Cons
Lower Cost	Reduced Performance & Privacy

Pros	Cons
Improved Performance & Privacy	Higher Cost

Figure 2: VPN WAN vs. private WAN.

are just one example of this. In mainframe systems, the software on the client side is normally termed a "terminal emulator," since historically, mainframe systems are accessed by a "dumb terminal," which is simply a screen display and a keyboard that connects to the central mainframe. One drawback of this type of system is that application compatibility is restricted to programs that run on the proprietary operating system of the mainframe or server. More recently, the concept of a "terminal" that uses processing power on the server side for applications has been expanded to include full desktop PC functionality. Now, with products such as Citrix Metaframe and Microsoft Windows Terminal Services, any application that can be run from a PC desktop can also be run from a thin client connected to the terminal server.

Advantages of mainframe and thin client systems from the perspective of network management include complete centralized management. Instead of being required to maintain hundreds of individual PC workstations, a network manager only needs to be concerned with a single computer, the mainframe, or a terminal server. All troubleshooting can be done from the server side.

ASPECTS OF NETWORK MANAGEMENT

User Management

Computer networks are functional only to the extent that people can use them effectively. A network manager has the responsibility to make technology accessible for end users. An observed weakness on the part of many network managers is their loss of focus on deploying useful technological solutions from the users' perspective. It is important for network managers to ensure that, while the network is managed efficiently, the needs of the end user are given full consideration.

There are virtually an unlimited number of issues to consider related to management of network users. In the following, some of the major areas are considered.

User Accounts

The first step in giving the user access to network resources is the creation of some type of log-in credentials. Traditionally, this involves a user name and a password, but depending on the security required for a particular installation, it may also include fingerprint recognition or other high security identification methods. While it may also be possible to enable open networking without the need for creating separate user accounts, this is strongly discouraged except in the most simple of networked environments. The use of a unique user name and a strong password, with users regularly forced to change and select passwords that are not easily guessed or found in dictionaries, is commonly accepted as providing adequate security for most environments (Kay, 2000).

Naming

Although any unique user name is generally acceptable within the limitations of the particular NOS, it is recommended that the network manager design a standard *naming convention* for user accounts. Any convention likely to yield unique user names is acceptable. For small organizations, the convention may be <first_initial><last_name>, for example. Using this naming convention, Albert Einstein's user name would be aeinstein. For larger organizations where initials and names are more likely to be duplicated, a more sophisticated naming standard may be necessary, such as <first_initial><middle_initial><last_initial><last_four_digits_of_social_security_#>.

Automation

Depending on the NOS, there are various utilities that can be used to automate the creation of user accounts. When a network administrator creates a Linux or Windows 2000 Server account, for example, a script is often run to fill in properties of the user account that would otherwise be entered manually. Instead of the network manager having to specify every user property individually, the script prompts the manager for the unique information for that user (name, password, etc.) and automatically sets other properties (file access permissions, home directory path, log-in time restrictions, etc.), thus making the process of account creation easier and more standardized.

Privileges

It is by means of the user account that the network manager can grant a user access to various network resources (printers, files, etc.). In addition to automating the user creation process, various NOSs allow for automation of user assignment to access network resources. Rather than simply having a database of user accounts, passwords, and individual permissions, a logical *directory* is created, where access to network resources can be set, not simply by explicit assignment to a user or group of users, but to a *location* in the directory (Lais, 2000). When a directory is well designed, little or no explicit user assignment to network resources is necessary. Users are automatically given access to network resources by virtue of their assigned location in the directory. For example, using Novell Directory Services, a directory structure can be created as shown in Figure 3.

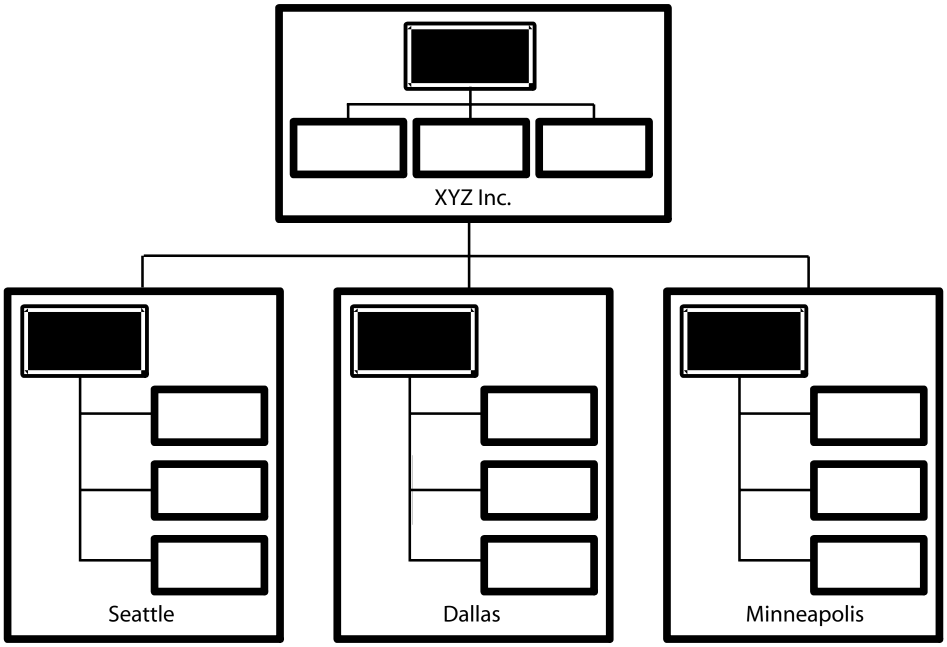


Figure 3: Typical Novell Directory Services structure.

Since appropriate access permissions have been granted to the various *container* objects, there is no need to do *any* explicit assignment of rights to individual users. Users are automatically given access to files, printers, and other resources when their user accounts are created in the given container. This approach to user management saves time for the network manager and improves overall efficiency.

Application Distribution

One of the most time consuming tasks in setting up a PC workstation is installing applications. To minimize the time needed to install applications on a multitude of workstations, some software manufacturers include the capacity for their setup programs to use a configuration file that shows what settings the network manager wants to apply. Using this method, network managers can simply start the installation process and then wait for it to finish. Since the configuration file contains all of the desired options, there is no need for the network manager to give full attention to the setup of the workstation to answer questions about what directory the program should be installed to, who the software should be licensed to, etc. The network manager has already answered these questions in advance in the configuration file.

An innovative feature of some NOSs is the ability to completely automate the installation of applications. Instead of loading installation disks and stepping through the installation process for each individual workstation, automatic installation can be set to run upon a user's login or some other event. For example, with Windows 2000 based networks, a difference file can be generated by running a script that notes the state of a model PC before an installation, and again afterward. The difference file lists all of the files added or modified during the installation, and this file is consulted during the automatic installation to determine what files and settings to add or change on the workstation during the installation.

Another approach to automating workstation setup and application installation is disk *cloning*. Instead of sampling the system state before and after a particular application is installed, the entire hard drive or partition is copied to an *image* file. This file is then used to create a clone of the sampled workstation by copying the disk image to the workstation to be set up. Instead of spending hours installing a workstation's operating system, software patches, hardware drivers and applications, network managers can simply spend a few minutes cloning a workstation. One limitation of disk cloning is that normally the model of the workstation and the workstation to be cloned must be identical in terms of hardward configuration. Some examples of disk cloning utilities are Norton Ghost by Symantec and Zenworks Imaging by Novell.

Network Policies

Even when operating system and application installation are automated by means of cloning or otherwise, there still may be a need for workstation configuration as to security settings, user restrictions, or the like. Instead of having to travel to individual workstations to make these settings, network policies are used by some NOSs to establish or configure various settings on a workstation. For example, if a network manager wants to make sure that users do not have the ability to change display settings on workstations, the manager can enable a policy that will restrict access to a workstation's display configuration utility. For Microsoft Windows 95/98/NT, these policies are configured using a utility called the System Policy Editor. This utility generates a policy file that is accessible to workstations so that when a user logs in, the policy is enabled on the workstation. For newer NOSs such as Windows 2000, group policies are used so that different policies can be enabled for different users or groups. These policies are configured using the group policy snap-in for the Microsoft Management Console.

Management Utilities

While NOSs provide various functions that are useful to network managers, these functions are normally enabled and configured by using proprietary utilities included with the NOS. For example, Windows 2000 uses the Microsoft Management Console to configure various NOS settings. Novell Netware comes with several utilities, such as Netware Administrator and ConsoleOne. Increasingly, management of networks is accomplished using Web-based utilities. For example, Novell Netware has a utility called iManage, where many network management functions, such as printer setup and remote control, can be performed from any accessible WWW browser on the network. In addition to the management functionality built into NOSs, other products are available that bundle several additional network management functions together as a suite. Examples of this include Novell Zenworks and Microsoft Systems Management Server.

E-MAIL COMMUNICATION

E-mail—A Major Application in the Network Environment

Of all the ways the Internet has come to be used through the course of its development, perhaps the most widely used service, having the most practical benefit is electronic mail. In business, it is a quick way to contact a large number of customers, colleagues, vendors, or staff with new products, information, orders, or instructions. At home, it is an inexpensive way for family and friends to exchange news, pictures, thoughts, or plans with loved ones on the next street or across the country or the world.

E-mail Clients

There are many variations in how e-mail applications function from the perspective of the end user depending on the choice of e-mail *client* used to send and receive e-mail. Although there are software applications whose only function is to send and receive e-mail, more and more frequently e-mail functionality is bundled together with software offering additional functionality. For example, Microsoft Outlook is not only an e-mail client, but also a personal information manager, so that e-mail, task and appointment scheduling, project tracking, and contact management can all be done centrally. The Web browser suite Netscape Communicator comes with an e-mail client called Netscape Messenger, so that Web browsing, Web design, and e-mail can all be done from one set of applications. Additionally, Web-based e-mail has also become popular. Many Web-portal sites such as Yahoo and Microsoft Hotmail provide e-mail accounts that can be accessed from any compatible Web browser connected to the Internet. The advantage to the end user is that e-mail can be sent and received without doing any configuration of a proprietary e-mail client and without having to leave the familiarity of the Web browser used to access the WWW.

Of course to enable the exchange of Internet e-mail from the corporate network, the network manager will have to establish a presence on the Internet in terms of a corporate Internet domain. This will be necessary regardless of whether the e-mail is being hosted externally or on a local server on the LAN with a dedicated Internet connection. Since this is a key required step we will discuss the major procedures involved in establishing an Internet domain for Internet e-mail and WWW services.

Providing Internet e-mail and WWW hosting for an entire organization basically involves the establishment of a fully certified Internet domain (or a subdomain) for the organization, with pointers to domain name servers as part of the domain name system (DNS).

The domain name servers are programmed so that all references to http://www.XYZinc.com and e-mail for users@XYZinc.com are forwarded to computers that are permanently connected to the Internet with their respective Internet protocol (IP) addresses stored in the DNS tables (Kallum, 1995). The computer with IP address corresponding to http://www.XYZinc.com and to the mail exchanger designated for users@XYZInc.com may be on the corporate LAN and accessible via a dedicated Internet connection and firewall scheme as described earlier. Alternatively, the network manager may elect to outsource mail and WWW hosting services to a reliable third party with regular updates to the corporate WWW site and frequent upload and download of corporate Internet mail. In this latter case the DNS entries for the domain XYZInc.com would reflect the IP addresses of the outsourced servers, and XYZ Inc. would then need to have only a low speed dedicated interconnection, a dial-up link to the Internet, or in some cases a link that is disconnected outside of working hours. The required servers for e-mail and the corporate WWW site are always online at outsourced locations permanently and reliably connected to the Internet.

Establishing a Corporate Internet Domain

Before the corporate WWW site http://www.XYZInc.com or the e-mail address user@XYZInc.com becomes active, the domain name XYZInc.com must be registered with an appropriate domain registrar by providing the following key information and paying the appropriate annual fees:

Organization name and address of the party that will own the domain;

Name, address, and e-mail address for the administrative contact, billing contact, and technical contact; and

Primary and secondary domain name servers for the domain name system. Both IP addresses and fully qualified Internet names are required.

The registration of a domain name with the DNS management hierarchy ensures that when this new domain (XYZInc.com) is referenced, for example, as part of a WWW address or an e-mail address, the appropriate servers can be correctly identified by their Internet protocol addresses. The DNS server then simply maps the host names such as *www* (referenced as *www.XYZInc.com*) to an IP address by way of a lookup table maintained by the designated DNS servers.

Once the domain name is registered and the DNS is active, a mail exchange (MX) DNS entry can be set up on the designated domain name server, to direct all mail to a certain mail server. If the mail server to which the MX record on the DNS server points is a computer on the local LAN of the company, then this computer would run a mail server program which would accept and send Internet e-mail for the users of the corporate network.

WWW ACCESS

Another responsibility of network managers is to ensure that all authorized users have reliable access to the Internet for WWW services. In a basic sense, access to the Internet is a matter of having the IP protocol running on a PC along with an appropriate network interface card connected to a routed network on the public Internet. However, due to security vulnerabilities and the limited number of IP addresses, it is not usually practical to allow each PC workstation to have its own unique address on the public Internet. Indeed even for organizations that have the luxury or the necessity for most if not all corporate computers to be fully integrated into the Internet, namely having a publicly addressable IP address, the network manager has to manage this IP space carefully, with full regard to the issues associated with IP routing and the implications of classless IP addresses as well as the transition from the older address format (IPv4) to the newer IPv6 addressing schemes.

Firewalls and Proxy Servers

There are a variety of methods used to provide Internet access to users while addressing the potential challenges of having a host on the public Internet. One approach that addresses the security issue is to install a software- or hardware-based firewall system so that private data are unavailable from the public side of the firewall. Another possible solution is to install a *proxy* server. With a proxy server, one host machine is configured as the only node authorized to exchange WWW traffic with the public Internet. The WWW proxy server fields and delivers WWW requests from other workstations, which are thus insulated from security threats and hackers, yet they still can access the public network. Another approach that addresses both the security and the finite address issues is to use network address translation (NAT). With NAT, a type of router is used that has an address on the public Internet. On the private side, the NAT-enabled device may function as a dynamic host configuration protocol (DHCP) server that dynamically assigns addresses to hosts on the private network. As data are delivered from the public network to the private, the NAT router dynamically assigns addresses for each internal (private) node. From the public side, the entire network appears as a single node, while the private network transparently connects to the Internet.

Monitoring and Content Filtering

Even if data on a private network were protected, a network manager should also be concerned about how the Internet is used by internal staff. Filtering Web content before it gets to the desktop is one way of ensuring that users do not access inappropriate content on the WWW. Desktop utilities such as Web-Blocker are available as freeware for business and personal use, and others are sold commercially. Another technique that can supplement Web filtering is logging user activity while online. If desired, a network manager can periodically check an activity log for inappropriate Internet usage and take appropriate action if necessary.

MOBILE USERS

Another consideration for network managers is how to handle users who are mobile. Mobile users can be separated into at least two categories. First, there are users who may rotate between several different locations, using a different PC at each location. In this scenario, a network manager would ensure that either the user has separate user accounts for each location, or if a WAN-enabled directory is being used for authentication, a user alias can be used so that a single user can obtain access to different resources depending on location. Second, there are users who roam while carrying along the primary PC in the form of a laptop or handheld computer. In this case, the network manager will have to make sure that the mobile PC is configured to interface with varying network topologies (10Base-T, 100Base-T, 10Base-F, Token Ring, dial-up, etc.), in addition to making resources available by location.

A feature of NOSs that makes mobility easier for users to handle is the roaming user *profile*. A user profile allows a user to have the same look and feel of the user's interface regardless of log-in location. This is accomplished when a set of configuration files on the user's workstation are copied to and from the server whenever the user logs in. Since these configuration files are stored on the server, the user's personal settings are preserved no matter where the user logs in.

USER SUPPORT

In addition to configuring the network and monitoring its performance, a network manager should form a strategy for supporting end users. Inevitably, questions or problems will arise due to hardware failure, software bugs, or user error. Helpful network managers will provide end users with a definite point of contact so that problems can be resolved quickly and effectively. Proactive user support strategies include thorough user training on system software and hardware and creation of an intuitive user interface so that applications, files, and resources can easily be accessed.

In addition the network manager will be concerned with the selection of hardware and operating systems for the file and application servers in the network as well as an appropriate backup regime to protect mission critical corporate data from loss due to equipment failure. The choices of operating systems for use in the network (servers as well as clients) include the Microsoft Windows environment, Apple Macintosh environment, and various flavors of the UNIX operating systems. Since the network resources (especially the servers) can be seriously damaged by electrical power outages, appropriately sized uninterruptible power supplies should also be provided, with the capability for graceful system shut down in the event of catastrophic power failures.

CONCLUSION

There are many resources available to network managers to make the job of administering networks less time consuming and ultimately more effective (Wisniewski, 2001). Although we have not covered every possible network configuration and software setup, we have outlined concepts and general techniques that can be applied to a broad range of network types. With technologies appearing at the speed of "Internet time," network managers benefit themselves by becoming well acquainted with the general strategy features of updated and newer NOSs and network management utilities. By having a good appreciation of the concepts of network management and staying abreast with new technologies as they become available, network managers can serve well their organizations and the users who depend on them. Looking to the future, one interesting aspect of network management that will become progressively more important is related to the issue of quality of service for multimedia traffic such as broadcast and interactive voice, video, and gaming services. The convergence of these services will be the driving force for many of the demands on the emerging home and office computer network environment.

GLOSSARY

10Base-T A media access specification operating at 10 Mbps over twisted pair cabling. This specification, along with its other varieties (including 100Base-TX), form the predominant networking infrastructure for modern LANs.

802.11x A set of protocols specified by the IEEE for wireless LAN connectivity. 802.11b, also known as Wi-Fi, has thus far been the most common specification in use commercially and supports data rates of up to 11 Mbps. The 802.11a specification is gaining wider acceptance and operates at data rates up to 54 Mbps. Other specifications such as 802.11e and 802.11i will address issues such as security and quality of service.

Asymmetric digital subscriber line (ADSL) A high speed digital access mechanism using regular telephone lines whereby higher data rates are sent to the subscriber and substantially lower data rates are supported from the subscriber to the network.

Asynchronous transfer mode (ATM) A switching technology in which data packets are transmitted in 53-byte cell units. Each cell is handled asynchronously, with buffering and queueing before being transmitted.

Campus area network (CAN) A network generally spanning an area the size of a campus composed of several buildings in relatively close proximity.

Carrier sense multiple access with collision detection (CSMA/CD) The method generally used in Ethernet defining how multiple hosts share a common medium. "Carrier sense" implies that hosts intending to transmit will "listen" first to see if the medium is currently in use. "Multiple access" denotes the fact that multiple hosts share the medium. "Collision detection" describes the process used to handle conflicts when multiple hosts attempt to transmit simultaneously.

Dynamic host configuration protocol (DHCP) A protocol which is used by a server to assign IP addresses and other networking parameters to enable access automatically to network resources.

Directory A database of network resources and user information needed to authorize persons to access hardware and software on a network.

Ethernet A medium access control protocol using CSMA/CD whereby nodes on a network "compete" for access to send packets. This is normally implemented on a physical or logical bus topology.

Fiber distributed data interface (FDDI) A network connectivity specification using fiber optic medium and that provides bandwidths of 100–1,000 Mbps over distances up to 250 km. FDDI is generally used in MAN and WAN applications, since it provides for transmitting data across relatively large distances.

Frequency division multiplexing (FDM) A method used to allow multiple signals to share a single medium, where each signal uses a separate frequency band. FDM is currently in popular use in broadband networks.

Firewall A combination of software and hardware used to restrict access to and from internal network resources based on address, authentication, or port.

Frame relay A data communication service for intermittent traffic, which operates between the typical low data rates offered by ISDN and the higher data rates supported by ATM. Unlike earlier packet switching protocols, in frame relay, error handling is done on an end-to-end basis.

Gigabit Ethernet Ethernet networks providing 1,000 Mbps throughput. While this technology is gaining in popularity, even faster bandwidth networks are in development, providing 10 Gbps. Currently, 1,000 Mbps transmission speeds are generally used for shared uplinks. However, it is increasingly more common for networks to provide 1,000 Mbps to the desktop.

HomePlug Alliance A consortium of technology companies committed to the successful development and launch of power line networking devices and solutions for home networks. The alliance is responsible for the HomePlug 1.0 protocol, which provides for 10 Mbps power line LANs. A new protocol, HomePlug AV, operating at 100 Mbps, is expected to be released at the end of 2002.

Institute of Electrical and Electronic Engineers (IEEE) A worldwide organization of engineers and researchers in the field of electrical engineering. A major service provided by the IEEE is the development of common standards and protocols, such as those relating to LAN connectivity and computer bus circuitry.

Integrated services digital network (ISDN) A digital scheme for a completely digital end-to-end telephone service supporting digital voice video and data over regular phone wires.

Local area network (LAN) A network of interconnected computers contained within a single location, such as room or a small building.

Metropolitan area network (MAN) A network generally spanning a city or metropolitan area.

Network address translation (NAT) A protocol allowing multiple nodes to gain access to the Internet while using a single public Internet address. Packets addressed to the single public Internet address are forwarded to internal private nodes by "translating" the destination address of the packets into the private address of nodes on the "inside" of the NAT device.

Network interface card (NIC) A hardware device that allows a PC or other equipment to participate as a node on a network.

Network operating system (NOS) An operating system with built-in capability to participate in a network as a node.

Proxy server A server that intercepts all requests for data from the Internet from internal network nodes. The server allows or rejects requests based on bandwidth, content, or location thresholds configured by the network manager.

Public data network (PDN) A network whereby a vendor will provide WAN connectivity to its customers. Rather than using the Internet, a vendor will use private links to provide service only to its customers. Traditionally, this has been the preferred WAN connectivity solution for organizations, since WAN communication over the Internet has been viewed as insecure. However, with the emergence of secure VPNs, organizations are becoming more comfortable using the Internet to connect remote locations.

Public switched telephone network (PSTN) The worldwide telecommunication network commonly used for voice telephone communication.

Quality of service (QoS) The required performance threshold to deliver a particular network service at an acceptable data rate.

Time division multiple access (TDMA) A method allowing multiple signals to simultaneously be carried across a medium, where "time slots" are allotted to each independent channel.

Universal serial bus (USB) A specification allowing computer peripherals to be installed (or uninstalled) on a "hot swap" basis. In addition to convenience, other advantages of USB include the allowance for multiple devices (up to 127) to share a single USB port via hubs or "daisy chained" devices.

Virtual private network (VPN) A data connection between sites or locations whereby each site or location is independently connected to the Internet and a direct, private data connection is obtained using encryption and "tunneling" protocols.

Wide area network (WAN) A network that spans an area wider than a single city.

CROSS REFERENCES

See *Guidelines for a Comprehensive Security System; Local Area Networks; Multiplexing; Public Networks; Virtual Private Networks: Internet Protocol (IP) Based; Web Quality of Service; Wide Area and Metropolitan Area Networks.*

REFERENCES

Atkins, J., & Norris, M. (1998). *Total area networking: ATM, IP frame relay and SMDS explained.* New York: Wiley.

Brown, B. (2002). Dual-mode wireless: Beating 802.11g to the punch. *PC Magazine, 21,* 46.

Harris, S. (2002). *IP VPNs—An overview for network executives.* Retrieved June 7, 2002, from http://www.nwfusion.com

HomePlug 1.0 protocol description (2002). Retrieved November 5, 2002, from http://www.homeplug.org

Johnson, H. W. (1995). *Fast Ethernet: Dawn of a new network.* Englewood Cliffs, NJ: Prentice Hall International.

Kadambu, J., Kalkunte, M., & Grayford, I. (1998). *Gigabit Ethernet: Migration to high bandwidth LANs.* Englewood Cliffs, NJ: Prentice Hall International.

Kallum, C. (1995). *Everything you wanted to know about TCP/IP—Tutorial.* Santa Clara, CA: 3 COM Corporation.

Karagiannis, K. (1999). The importance of USB. *Electronics Now, 70,* 8–9.

Kay, R. (2000). *Authentication.* Retrieved June 6, 2002, from http://www.computerworld.com

Lais, S. (2000). *Directories.* Retrieved June 6, 2002, http://www.computerworld.com

Latchman, H. (1997). *Computer communication networks and the Internet.* New York: McGraw–Hill.

Latchman, H. A., and Yonge, L. (2003, April). Power line local area networking. *IEEE Communications Magazine, 41*(4), 32–33.

Lin, Y., Latchman, H. A., Lee, M. K., & Katar, S. (2002). A power line communication network infrastructure for the smart home. *Wireless Communications, 9*(6), 104–111.

McKenzie, M. (2001). Wireless networks find a home at work. *PC World, 19,* 150.

McLeod, R. G. (2002). Gigabit nets for small business. *PC World, 20,* 36.

Morrison, G. (2002). Circuit-switched network lives on. *Electronic News, 48,* 26.

Newsome, T. (2002). Test industry responding to signs of the times. *Electronics News, 48,* 20.

Schuyler, M. (2001). Measuring network traffic. *Computers in Libraries, 21,* 55–57.

Spanbauer, S. (1999). Windows NT. *PC World, 17,* 273.

Stallings, W. (2002). *High speed networks and Internets—Performance and quality of service* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall International.

Wisniewski, S. (2001). *Network administration.* Englewood Cliffs, NJ: Prentice Hall International.

Wolf, M. (1999). From fast to faster: Ethernet IC vendors get Gig-E with it. *Electronic News, 45,* 12.

FURTHER READING

Flannigan, W. (1990). *Guide to T1 networking.* New York: Telecom Library.

Managing the Flow of Materials Across the Supply Chain

Matthias Holweg, *Massachusetts Institute of Technology*
Nick Rich, *Cardiff Business School, United Kingdom*

INTRODUCTION

The Supply Chain

The supply chain as concept has been widely discussed and was first promoted as a source of competitive advantage by Michael Porter of the Harvard Business School in the 1980s (Porter, 1985). The logic that underpinned the rise of supply chain management and competitive advantage was born from the notion that "supply chains compete, not individual companies" (Christopher, 1992).

Traditional management thought did not promote the benefits of integration and collaboration with suppliers and distributors, preferring instead to promote "mistrust" of such companies and the use of "power" to bargain down the costs of products and services in a "zero sum" approach. In addition, traditional management thought has always focused on the staking out of a competitive position by the firm—but, as only one node in a total flow process that provides materials to customers, it is difficult to compete effectively with a poorly integrated supply and distribution network. Improvements to the supply chain (upstream with suppliers and downstream with distribution partners) represented a logical means of translating business efficiency into a much broader "intercompany" and supply chain approach to competitive advantage. In addition, supply chain management and the integration of trading partners brought with it access to new knowledge and a communication channel that could improve chain performance by mutual development, strategy sharing, targeted operations improvement, and the ability to bring new products to market quickly. Supply chains, not individual companies, actually compete, and the success of the supply chain in providing materials, products, and services profitably determines the viability of all members of the chain.

The management literature from the late 1980s promoted competitive advantage based on a supply chain approach grounded in mutuality, collaboration, and the commercial benefits of "working together." To be "world class" and to design an effective supply chain therefore meant going beyond simple price negotiations and instead examining all potential partners to test whether they had the necessary capabilities and processes to support the company in its pursuit of customer satisfaction and market growth. At the heart of these deliberations concerning the selection of the "right type" of suppliers and distributors is the ability to flow products and information between companies and to ensure that inventory never stood still. The objective was therefore to design new systems that used the minimal amount of time to satisfy customers at the lowest overall costs and to make a profit in doing so.

Within the supply chain, two main flows are distinguished: (a) the information flow conveying demand information from the customer back upstream to the manufacturers and their supplier tiers (b) and the physical or material flows delivering and converting the material into products, until it reaches the final customer (see Figure 1).

These two flows are integral, and interlinked, and can only be improved by treating information and physical products as a single system within which information "triggers and paces" the release of materials to flow through production and delivery functions (orders, schedules, etc.) to the customer. The details of these flows, and how they interact are discussed in detail in the following two sections.

Current Business Trends

New technologies and new business models are affecting the design and management of the supply chain, and it is important to understand how these issues influence the performance of any given supply chain. First, there is a general trend toward compressing the time involved in designing, manufacturing, and delivering products. This notion has been popularly termed *time-based management,* and the concentration on eliminating delays in the supply chain has been promoted widely as a source of competitive advantage (Stalk & Hout, 1990). This advantage

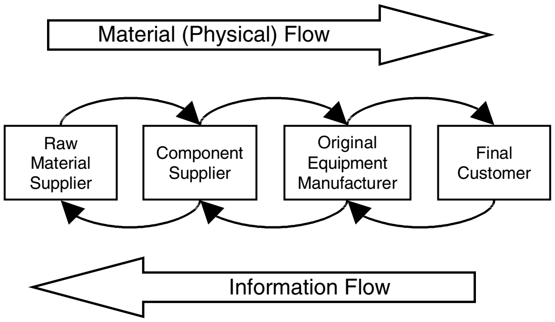


Figure 1: Basic supply chain.

is therefore based on "speed-to-market" and on getting material flow and product ideas from the design concept to the customer "just in time" (JIT; Ohno, 1988). A JIT approach to the physical flows of existing product therefore increases competitive advantage by lowering the costs of inventory buffers, reducing obsolescence, and increasing flexibility by making better use of factory capacity.

Second, the concept of partnering with suppliers and customer relationship management (CRM) has challenged the traditional adversarial approach favored by many Western producers (Dowling, 2002). The "partnering approach" acknowledges that there is a dependency between supplier and customer organizations throughout the chain, and therefore rather than "fighting" between companies it is more important to work together to improve customer service performance (and profits) by collaborating (MacBeth & Ferguson, 1994). Such an approach is common to the manner in which the world-class companies work with their supply chains and is an approach that has supported the competitive advantage enjoyed by many Japanese manufacturers in markets such as electronics and vehicle production. The objective of such collaboration is therefore to improve faster than competitors and to switch from short-term, price-driven negotiations to a long-term approach to collaboration and a steady profit stream for all members of the supply chain. The stability that a partnership brings therefore allows investment decisions to be taken with a greater level of certainty that the company will continue to trade with the supplier and vice versa. In effect, the restriction of information by traditional adversarial purchasing is replaced with greater transparency of information to encourage the right decisions and behavior throughout the chain.

Third, interest in the concept of world-class manufacturing and supply chain management has been greatly accelerated by benchmarking studies of comparative performance (conducted during the 1990s). These studies confirmed that in sectors such as vehicle production, Japanese producers (often termed "lean" producers) operated at vastly superior performance levels than Western firms (Womack, Jones, & Roos, 1990). The competitive disadvantage of the West therefore resulted in a "productivity challenge" to catch up with the performance of Japanese firms in order to remain competitive. The process of learning about how world-class companies operate helped identify a number of best practices concerning production system design, human resource management, and a collaborative approach to the supply chain. The practices of these world-class companies were termed "best practice" and, even today, companies still benchmark themselves against these world-class competitor companies, and even world-class companies from completely different industry sectors. Benchmarking against the best helps a firm find out which practices support high-performance supply chain management and how best to emulate them.

Fourth, the outsourcing of manufacturing tasks to suppliers or low-labor countries implies that businesses must develop a supply chain management capability to control the manufacturing operations that were once "in house" and "made." The outsourcing of such manufacturing, to allow the company to concentrate on its core competencies, needs careful management to ensure that material continues to flow properly. In some cases, outsourcing included only the retention of brand ownership with the entire product being subcontracted. The risks of affecting the "value of the brand" by poor supplier performance prompted a greater involvement in the development of the supply chain and a strategic dependency between firms.

In the automotive industry, outsourcing has also included a restructuring of the supply chain and an increasing amount of modularity in product designs (Ulrich, 1995). Employing a "modularity strategy" involves the supplier being given the task of manufacturing a much larger and specific module for the vehicle, such as a dashboard or the vehicle interior, which previously would have been built by a number of smaller manufacturers. The vehicle assembler thus has fewer suppliers of much larger modules that can be simply installed into the vehicle at the assembly track. In this manner, modules arrive and are simply "slotted into" the vehicle, which again helps to eliminate "build time" at the assembly track. In other sectors, such as consumer electronics, modularity is also been used as a means of offering increased product variety. This is often coupled to the assemble-to-order strategy, which is discussed later.

Fifth, new computer innovations and e-business have a major impact on the current state of supply chain management and the business model employed by manufacturing companies. Three main areas of e-business can be identified. First is B2C (business-to-customer) using online ordering and product configuration tools. Examples include online retailers such as Amazon.com or online configuration Web site such as that operated by Dell Computers. Second is B2B (business-to-business) e-commerce between companies, which allows for communication of point-of-sales data throughout the supply chain, as well as real-time inventory and order data. Examples include the Efficient Consumer Response (ECR) movement in the grocery sector and the Quick Response initiative (often also referred as Quick Response Manufacturing, QRM) in the textile sector. Such communication in the supply chain can be enabled through the Internet but commonly uses dedicated extranets or other forms of electronic data interchange (EDI). Finally, the E2E (engineer-to-engineer) component of e-commerce facilitates collaboration between the various companies in the product design stage. Most important for supply chain management is the B2C aspect, because it enables the link between the customer online to the manufacturing and distribution process, as well as the B2B aspect, because it enables communication of demand and planning information without delay in the supply chain. It is important to note that although electronic communication has great potential in streamlining the information flow in the supply chain, the challenge of effectively managing the physical flow remains and is critical for the success of any organization—whether it is a "dot-com" or a traditional company (Lee & Whang, 2001).

For customers, e-business allows products to be "last-minute configured" exactly to their requirements and satisfies the need for a "portfolio" of products to offer the new e-market for manufactured goods. Also, online auctions can be conducted, as in the case of some automotive trade exchanges, such as the recently established Covisint.com Web site. The application of e-business for the supply chain therefore increases competition amongst suppliers of commodity or "nonstrategic" supplies and ensures that the best prices, on a global scale, can be negotiated. For strategic suppliers, on whom the buying company is dependent, e-business offers a new communication channel and the ability to access important information databases (such as design and engineering) and to synchronize efforts across the supply chain wherever that activity is conducted.

Effective Supply Chain Management

Managing the physical flow is the most basic form of supply chain management that adds value for the supplier and buying organizations regardless of online or conventional operation. The ability to control the physical flow of products is, however, highly dependent on an accurate flow of information between companies and across the entire supply chain to enable synchronization. Therefore, the design of the supply chain and the careful selection of companies with whom to partner is just one side of a bigger design puzzle. The key to solving the puzzle and raising performance is to apply the same level of effort to the design of communication channels between companies. If such attention is not provided, suppliers will act on bad information, resulting in frequent interruptions to material flow and thus inefficiency. Efficient online order processing is undoubtedly a major enabler to improved information exchanges and an indirect form of competitive advantage, yet if it is not coupled with a reliable and rapid physical order fulfillment process (i.e., the manufacture and distribution of the product), then little is gained. Instead, the fast processing of data will not have raised customer service and will not attract loyalty from the customer base. The art of an effective supply chain and information system design is in the sharing of data and the filtering of misleading information that could cause products to be produced either too soon or too late to meet the delivery requirements of customers. Either way, the costs of the supply chain will rise, and excess inventory will result, adding costs but no value to the customer.

We now outline the basics of both material and information flows, their interactions, and discuss the key challenges that must be met to design efficient material flows for sustainable and competitive supply chains in the Internet age.

THE MATERIAL FLOW BASICS

Basic Features

The material flow in supply chains comprises physical inventory, production processes, and logistics or distribution processes. Each material flow or value stream (the combined processes that make a product) must be combined to optimize the customer service level provided at the minimum cost. The key challenge is to minimize the inventory, while utilizing the production capacity to the maximum extent (Simchi-Levi, Kaminsky, & Simchi-Levi, 1999).

Physical Inventory in the Supply Chain

There are various forms of inventory that exist in any given supply chain. The obvious ones are the raw materials and subcomponent stock needed to produce parts that ultimately make a finished product. While being processed, the material is referred to as "work-in-progress" (WIP) inventory, which is of particular interest because it is a good indicator of the quality of production control. The higher the inventory, the more interrupted the flow of goods and therefore the longer the production cycle time. In addition, work in progress carries substantial risks, and if orders are canceled or the business is closed, this inventory cannot be sold (unlike that of finished goods and raw materials). As such, it is the management of WIP that determines the effectiveness of the overall production system. The final form of materials is converted finished goods inventory (FGI), and this is an indicator of how close the production process is aligned to customer demand. Highly efficient manufacturers with reliable processes therefore have much lower FGI and instead use the effectiveness of the production system to be responsive to customer demand.

There are several reasons inventory is held. First of all, cycle stock refers to the inventory needed to cover the time period between two material deliveries and thus is nonstationary, or continuously turning over. As such, this inventory is essential for sustaining the manufacturing operations. The longer the time period that has to be covered, the more cycle stock that must be held. A range of replenishment techniques can be used, which will not be discussed here (see Silver, Pyke, Peterson, & Miltenberg, 1998). Second, there is safety stock, which is inventory held in addition to the cycle stock to cover for eventualities such as machine breakdowns or unreliable supply. Third, pipeline stock refers to the inventory in the distribution and retail channel (i.e., in the warehouses and shelves in the shops). This inventory comprises both cycle and safety stock elements.

For traditional operations planning and production control, large amounts of WIP resulted from large batches of production, resulting from a hypothesis that large batches were more economical. Such a system has many weaknesses when attempting to compete in the modern market for manufactured products. Large batches increase costs (inventory), and because asset capacity is consumed in making large lots of standard products, the responsiveness of the factory is reduced. Instead, the modern JIT approach is to design a physical production process that is better able to respond to demand than to try and push large batches of production (against a schedule) to meet what the planners think that the customer will buy. Working to demand using small inventory buffers is a characteristic of the "lean production" Japanese manufacturing approach to material flow. This production system is operated by world-class companies such as the Toyota Motor Corporation (the subject of the benchmarking activities of the 1990s mentioned earlier).

Within the three classifications of raw materials, work in progress, and finished goods forms of inventory there is another classification of product. This classification concerns the value and volume of all materials and products and allows insights into designing the most effective form of conversion process. This form of inventory analysis is often called an ABC classification or the "runners, repeater, strangers" of factory volume. Parts are classified according to their volume in the production system and multiplied by their monetary value to determine their overall impact on the production system. The ABC classification distinguishes A-parts, or runners, as the highest impact products manufactured at the factory. These products are produced continuously or very frequently and as such form the basic loading of the factory. In addition, if these products are not available "on demand," then customer service will diminish.

For B-parts, or repeaters, the product is made on a regular basis but less frequently than A parts. Here, managers must decide how often and in what batch size to make these products and often decide to hold slightly higher amounts of finished product and to manufacture in slightly larger batches every planning time period. For example, running products may be manufactured every week (to replenish at least a week's worth of finished goods inventory), whereas repeaters may be made every 2 weeks (to replenish a finished goods inventory of at least 2 weeks worth of sales volume).

Customers infrequently order the strangers, or C-products, and therefore managers often decide to hold finished products to allow customers instant satisfaction of their order requirements. The production of these C products is likely to be in a large batch, however, and the batch will be launched only when the available finished product has reached a minimum level. This form of classification is useful and helps to maintain an appropriate factory loading and ensures products are available (but not in excessive quantity) to meet demand.

In any inventory analysis, it is important to consider the volume, value, and frequency of the product to determine the right approach for creating the appropriate supply chain for the product. The more frequent and valuable, the more the product should be linked to customer demand and actual orders. For low-value, low-volume parts, however, it is likely that holding stock is the more economical option (such as for nuts, bolts, etc.). Each production system must be designed to meet the unique market characteristics of the firm. Where inventory is deployed to raw, work in progress, or finished goods, it is there by design to allow maximum customer service, to protect manufacturing lead time, and to cover for catastrophes in the supply chain. By stabilizing the levels of inventory and operating to these conditions, the pace and frequency of production and supply can be synchronized (each ABC product within the firm will be an ABC for the supplier). In this manner, inventory is used to stabilize and regulate the production and supply systems and when an unplanned event occurs, it should be obvious that something in the total system has failed. Under the traditional system of large batches and no knowledge of demand patterns, inventory simply materializes and this "stability" is not achieved, therefore it is difficult to determine when things have gone wrong.

Production Processes

The production process is the actual conversion of the materials and subcomponents into finished products. Examples include metal pressing, welding, painting, subassembly, and final assembly operations. The actual material conversion process is generally dictated by technological factors in terms of "how" products are converted, and it is the manner in which the production system is planned that is of most concern. This feature of a production system includes how processes are scheduled (the information trigger that determines how manufacturing operations are instructed when and what products to make).

The "correct" way to schedule a process is subject to a number of schools of thought, but basically there are two extremes and many hybrids between them. The most common distinction between the two extremes is whether a "push" or a "pull" approach is taken. Push systems use production forecasts, not actual orders, to determine the factory and supplier schedules. Products are made on predicted usage, not on actual orders, and "batched" into large lot sizes before being pushed from raw materials all the way into the warehouse. At the warehouse, the product will wait until an order is received from a customer.

Advocates of push systems argue that this is beneficial because production is more efficient when large batch sizes can be operated. The main risk is that if the forecast is too high, inventory of products that are not selling will increase; if the forecast is too low, there will be no inventory buffer of products in demand, and therefore orders cannot be fulfilled. Push systems are commonly used in conjunction with MRP systems (material requirements planning), which have evolved into MRP II system (manufacturing resource planning) and later into ERP systems (enterprise resource planning), yet essentially push systems still apply the same logic (Wight, 1995).

MRP systems are a common approach to manufacturing scheduling and use a master production schedule (which is an amalgamation of sales forecast and customer orders) to generate time-phased production orders for each work centre on the shop floor. These systems also calculate the quantity and timing of order to be released to material and component suppliers (Schonberger & Knod, 1997). In effect, the schedule takes each order and deducts the amount of time needed at each stage in the production process to determine the time that it must be available for processing. To illustrate this point, if assembly takes 2 days and the customer expects the product on the 10th of the month, then, because assembly is the final operation, the materials must be at assembly by the 8th of the month at the latest to meet the order. Such MRP systems can therefore calculate complex requirements for products at different due dates, going through all the factory work centers in a fraction of time that it would take a human being. Although the MRP approach seems logical, in practice the uncertainties of production processes and capacity issues require additional buffers in the systems, hence MRP systems tend to result in longer lead times and higher inventories than the next approach.

The alternative is the pull system, in which the customer, who pulls the product through the supply chain, drives the schedule. This approach is based on the just-in-time philosophy derived from Toyota in Japan (Monden, 1983), and was subsequently known as "lean manufacturing" in the West (Womack & Jones, 1994; Womack et al., 1990). Under this approach, rather than pushing products to the finished goods warehouse, the business allows products to be taken from the warehouse to meet customer orders. When a certain point is reached, a request for resupply is placed on the final operation, which takes materials and converts them to replenish the stock holding. Each subsequent process takes the relevant material from their in-bound buffer (known as a "kanban") and replenishes the buffer of their customer work center. This trigger to make products goes all the way to the supplier and allows instant customer satisfaction with a minimal amount of finished inventory. The lean approach therefore aims to produce products at the same rate as the customer needs them, which is called "Takt time," and frequent small-lot manufacturing enables a flexible means of resupplying what has been sold. Lean manufacturing has since migrated into many sectors beyond automotive manufacturing and offers a wide tool set for the analysis and improvement of value streams and manufacturing operations (Hines, Lamming, Jones, Cousins, & Rich, 2000; Rother & Shook, 1998).

The hybrid approaches to production scheduling include elements of the push and pull systems that are combined in a logical manner to support the stability of material flow in the factory. An illustration of this approach is the situation in which the beginning work centers at a factory require large batches but the finishing operations are fast and can process in small lots. Under these conditions, the two parts of the factory could be separated such that the initial stages work to a schedule and fill the buffer that separates the work centers from the finishing processes. The finishing processes work to the replenishments needed by the warehouse, however, and therefore will treat the inbound buffer as a large kanban. The finishing processes therefore convert more frequently much smaller lots of production from the beginning processes. The objective of the primary and finishing operations is to reduce batch sizes, the ideal batch size for both being a *single* product (also referred to as single piece flow), such that a single order from a customer triggers a single product being made by the factory operations.

Logistics and Distribution

Logistics and distribution is the counterpart to the manufacturing processes in the material flow process, with the main difference being that the product itself is not altered during the transportation process with the possible exception of finished goods packaging. The main goal of the distribution function is to ship, at minimal cost and with minimal lead time, products to the next tier or to the customer in the supply chain. To achieve this goal, managers often have to trade off between load efficiency and delivery frequency, similar to the batch-size–inventory trade-off in the manufacturing process.

Distribution and logistics have been widely explored (e.g., Bowersox & Closs, 1996) and the design considerations for managers include the location of warehouses and decision on the mode of transport (truck, rail, ship) and the size of the transport. Recent developments in this area include the use of "milk-rounds" for distribution in which finished products are delivered frequently to many destinations (and materials for the factory could be picked up). Another development is "cross-docking" of transport to allow bulk movements of common products to a destination from which another transporter will take its requirements for delivery. Cross-docks are therefore facilities used to reassemble truck or rail loads, so that both the collection and delivery runs are efficiently loaded. Cross-docks may also be combined with a milk-round system to support frequent deliveries to customers, which is typical of the systems used by supermarkets and vehicle assembly factories.

Best-performing logistics systems in the automotive sector, such as Nissan Europe, for example, can achieve delivery windows of "every 2 hours" into the assembly plants, thus reducing the effective inventory between supplier and manufacturer to less than 1 day. Interestingly, the common perception of Nissan is that no further financial benefit can be gained from reducing the pipeline inventory below 1 day. Yet the quality benefit of tightly integrating the suppliers into the JIT production and delivery schemes is the main driver behind this concept.

Often, logistics schemes are run by external or "third-party" service providers. These are independent and specialized logistics companies that work as subcontractors to the manufacturers and suppliers. For example, if a customer orders a Dell computer, it is likely that the final delivery will be made by DHL or TNT Express, rather than by a Dell van. In the auto sector, such third parties, such as Exel Logistics or Ryder, are charged with managing entire regional collection schemes from several hundred suppliers. In recent years, these management responsibilities have even been extended into the management of entire supplier parks (whereby the supplier collocates around their customer's assembly plants). In these cases, the logistics company is in charge of not only managing the site itself, but also with managing the logistics of the park as well as delivery into the assembly plants, which involves coordination of other third-party logistics company. In this case, the "lead logistics partner" is also referred to as a "fourth-party logistics provider" and might not even execute any actual transportation tasks.

INFORMATION FLOW BASICS

Information Flow—the Pacemaker

The information flow throughout the supply chain is the "pacemaker" that drives the system and triggers questions about products such as what to make, in what quantity, and by when. The problem with scheduling is therefore a complex one, because multiple trade-offs must be considered and because the variety of components and materials needed for every finished part makes the design problem a major task to review. More important, however, is that the difference in lead time between the time needed to make a product and the time the customer is prepared to wait for it must be considered. It is generally the customer's willingness to wait that is collapsing today, which means that manufacturers must generate high quality and rapid delivery of products in increasingly shorter time frames. An unfortunate consequence of the Internet is that this technology has fuelled the impatience of the consumer and customer to the point that they both expect immediate satisfaction of their needs.

What Constitutes Information?

Within the information flow, several types of information are important to the design of an effective supply chain. First, there are the actual orders, which are commercially binding. These orders are sent by the customer either on individual ("one off") or on a regular basis. For example, every new vehicle buyer will order his or her car individually, yet Ford or General Motors will order the parts needed to build cars on a regular, daily basis from their suppliers. The individual customer order does not provide enough visibility or time to plan production in time to meet the delivery required. Therefore, forecasts are also generated to give further notice of upcoming demand at such manufacturing sites. Forecasts, however, are "best guesses" or interpretations of the future demand for products, and when these forecasts deviate from the actual demand that arises, a forecast error exists. This error has a direct impact on the amount of inventory in the system, so forecasts should be kept to the minimum time horizon possible to increase the relative certainty of actual demand. The longer the forecast time horizon, the greater the risk of error being produced and a mismatch in what has been made against what was wanted. Many techniques exist to help with this form of analysis, of which the P:D ratio analysis is one of the most important in designing effective supply chain systems.

The P:D Ratio

A more detailed model of how to cope with the difference between the lead time to produce a product (P = production lead-time) and the time the customer is willing to wait for it (D = demand lead-time) was proposed by Mather (1988). He called this the "P:D ratio," and it compares product and delivery times to suggest implications for the design of manufacturing operations and the manner in which orders should be fulfilled. This model (shown in Figure 2) is also referred to as the order penetration point or decision point analysis, because it shows how far the customer order penetrates the system and "pulls" the product. Put simply, the model highlights the differential in production and "expectation" time and therefore identifies where a manufacturing facility should locate its main buffer of inventory to allow customers to be satisfied without having to make to order for each customer.

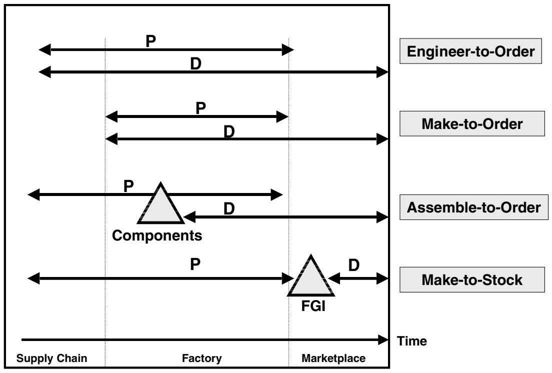


Figure 2: The P:D ratio. Based on the P:D ratio by Mather (1988).

The model shows many points in the manufacturing supply chain where inventory may be held. On the far right-hand side of the diagram is an order penetration point that means all manufacturing is "pushed" and driven by forecasts to replenish a major buffer in the finished goods area. This point marks the interface of "push" and "pull" in the supply chain; in this instance, the customer pull is from the finished goods buffer. At this point, the pull is not truly effective, and customer orders are simply taken from finished goods buffers and dispatched to awaiting customers. In case the customer is not willing to wait until the products are manufactured, the only option is to hold finished goods inventory. This is the case in supply chains such as in the grocery supply chain, in which customers expect milk bottles to be on the supermarket shelf and are not willing to wait until the bottle is filled, shipped, and put onto the shelf while they wait. The challenge in this "make-to-forecast" scenario is to replenish the stock as efficiently and quickly as possible, to minimize production and inventory cost. A good example of this is the recent Efficient Consumer Response initiative in the grocery sector, which uses electronic point of sales data to replenish the stores (Kurt Salmon Associates, 1993; see also http://www.ecr-net.org.)

If the customer is willing to wait for the product assembly, yet the overall production of all the components is outside the customer's tolerance, an "assemble-to-order" approach should be used with a "penetration point" at the work in progress buffer. At this point, the product is assembled to customer order using an inventory buffer at parts level. Every time an order comes in, a customized product is assembled, and the parts buffer is replenished on a forecast basis. The most commonly known assemble-to-order system is the Dell approach, in which the computers are assembled to order within a few days, buffering the long replenishment lead-time for components from distant suppliers. If the customer had to wait for all components to be made to order, the overall lead time for a

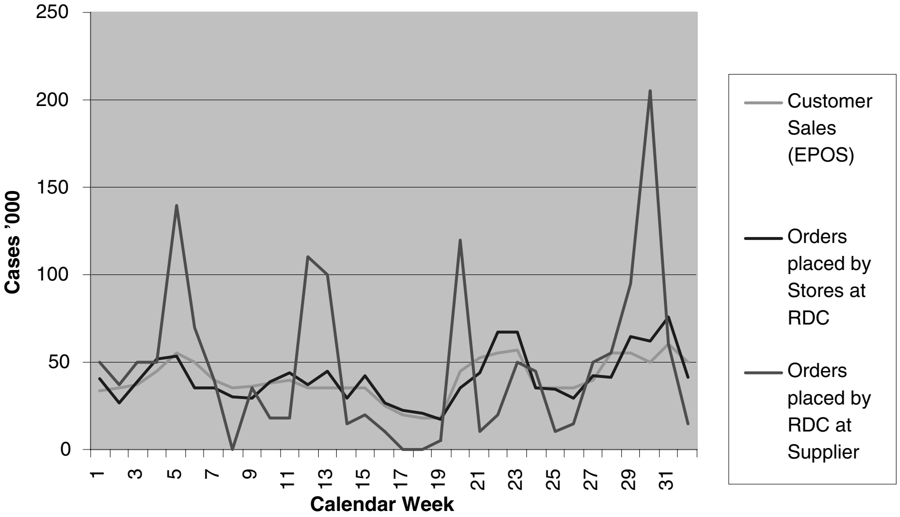


Figure 3: Demand amplification example. Source: Holweg & Pil, 2001.

computer would be several weeks, as opposed to several days.

For certain products such as motor vehicles, however, which require a much greater range of components, holding stock for all of these would be too expensive. Also, customers are generally willing to wait 1–2 weeks for their vehicle to be delivered, and hence a "make-to-order" system can be used. Ideally, a make-to-order system would hold no inventory at the finished product or component level, because production is only initiated on customer request (Holweg & Pil, 2001).

For highly specialized products, such as oil rigs, for example, the product is not only made to order, but also designed specifically. This case is also referred to as "engineer-to-order." Hybrid cases between engineer- and make-to-order are also common, as, for example, in the aerospace industry, in which planes are of an overall standard airframe design but are highly customized in many respects.

The P:D analysis is important in determining the right "mode of supply chain" and where best to position inventory to protect the production lead time and the customer's expectation of service. It should be noted that any of these order fulfillment approaches could be coupled with an Internet-based order entry interface, and Dell, Amazon, and ECR are only few examples. A list of relevant Web pages is provided in the Further Reading section.

THE KEY PROBLEMS IN THE SUPPLY CHAIN

Supply chain management possesses several key challenges that pose a threat to customer service. These challenges include excess inventory and ultimately cost. Both of these failings relate to structural problems and system effects that have long been known to many business managers yet have not been solved by managers in many modern supply chains. One of the most commonly found effects is the so-called demand amplification or "bullwhip" effect (Lee, Padmanabhan, & Seungjin, 1997). As shown in the example in Figure 3, the demand pattern is distorted as it is transmitted through the supply chain, and this information distortion amplifies as it is passed from tier to tier down the supply chain.

The main reason for such distortion, which leads to excess inventory and cost, is the delay of information before it is transmitted. Another problem is the treatment of data when it is received. Often order information does not match certain requirements, such as batch sizes for manufacturing and other policies, and therefore the information is "rounded up" to the nearest convenient batch size for manufacturing. To illustrate this process, a business may receive an order for 250 fax machines, but because the minimum batch size to make these products is 300, then 300 will be made and new orders will be issued to suppliers. If the policies that are established with suppliers means that orders are placed for a minimum of 1,000 plastic molded machine bodies and keyboards, then the initial order for 250 products has now been amplified to 1,000 units of supplied parts. If these ordering processes are completely automated and rarely investigated by managers, these problems will continue to exist and continue to "trigger" production and movement of products that are not really needed. The result is WIP inventory and FGI inventory, which, if not immediately sold, incurs costs (storage being just one part of a much larger amount of cost incurred). The other problem is that this creates a need for capital, because these products will have been paid for and will wait for a certain period until a customer purchases them, and only then is the company paid. Meanwhile, the stock has to be financed.

Further problems are variability and uncertainty of material supply and within the manufacturing process. Any source of uncertainty will drive planners toward "playing it safe" and hence lead to overordering or making some "just in case." Ultimately, uncertainty results in excess inventory, either in the form of materials or finished products. Poor asset maintenance and frequent breakdowns of machines will therefore increase these problems and result in material movements that stop-and-start but do not flow. The quantification and elimination of variances and problems with manufacturing processes is the focus of the Six Sigma approach that has recently increased in popularity as a management technique of companies such as Motorola, Johnson and Johnson, and many other world-class corporations (Bicheno, 2000).

To summarize this chapter, it is important that managers understand the concept of the supply chain both in terms of the internal supply chain of machines and processes that convert materials and also the external chain that involves the integration of material and information flows. This chapter also promotes an understanding of the internal and external chain in terms of where best to position inventory (for customer service) as well as the critical integration of information channels to support the flow of products. The next section continues to explore the challenges of supply chain management.

KEY CHALLENGES IN MANAGING THE SUPPLY CHAIN

Any supply chain faces the challenge of meeting—or even exceeding—customer expectations. Only if customers' requirements are met can a supply chain be sustainable for the long term. The key is to manage the supply chain effectively for sustained competitiveness in terms of cost and customer service. Thus, it needs to be understood as a single entity of material and information flows that must be managed and designed holistically. Despite the great variety of supply chains in all economic sectors, there are a number of common challenges or principles for the effective design of material flows.

First, delays of information and material need to be eliminated whenever possible. Delays not only reduce responsiveness to customer demand, they also worsen the demand pattern distortion and are a main cause for the bullwhip effect. The identification of delay points is important if the correct countermeasures, policies, and information flows are to be established and managed (or, at the very least, monitored).

Second, inventory in any form needs to be questioned as to its purpose. In reality, only a fraction of the stock being held in supply chains is actually necessary for the efficient operation of the system. The reminder results from failures (in understanding customers, in setting minimum production and transport batch quantities, in poor asset maintenance), and this inventory incurs cost and hides quality problems. Even worse, the bulk of inventory hides problems that managers must know about if these problems are to be solved and material flow throughout the entire chain is to be improved. Toyota refers to this inventory as "dead material," and it is this inventory, waste, and costs that are the focus of lean production.

Third, uncertainty of any kind needs to be identified because uncertainty leads to poor decision making and a suboptimal flow of materials. Whether through Six Sigma programs or through joint improvement projects with suppliers, information and knowledge about the negative impacts of uncertainty are valuable and assist the design of robust supply chains. Uncertainty about customer demand, the basic information that drives any supply chain, should be understood and any areas of uncertainty investigated to perfect performance.

Fourth, companies at the top of the supply chain (closest to the customer) should try to promote the customer demand signal to the maximum extent possible. Many companies settle for the easy make-to-stock approach, yet pull systems require far less inventory and are much less likely to show the bullwhip effect that is evident when customer demand is not understood and communicated with all businesses in the supply chain. These demand-driven supply chains, as characterized by Dell Computers, outperform their competitors if these rival firms push materials to forecast. For a JIT movement of materials throughout a supply chain, customer information must be packaged into a common format for all suppliers. This information packaging should be such that any distortion is "filtered out" by the company at the head of the supply chain to allow all other businesses to achieve a "flat" and "level" demand. In this way, a stable demand pattern allows materials to flow without generating the chaos and uncertainty of information that is erratic, infrequent, or constantly amended.

Fifth, other decisions, such as whether to source locally or from an international low-cost supplier, need to be made with the supply chain in mind. In many cases, the latter option seems at first glance to be the better cost option. Long-haul or global sourcing has serious disadvantages in terms of lower flexibility and responsiveness, however, and can incur premium airfreight needed to cope with schedule fluctuations and expediting. These charges (for problems associated with information flow) often outweigh the benefits of the lower labor cost.

Further policy decisions include the control of the manufacturing system and investments in technology that allow the "delayed differentiation" of products and postponement of the order penetration point. In this scenario, the product is customized at the last possible moment in the production process in order to meet customer specifications. This system represents the scheduling of materials to a given point at which the product is then pulled to the exact specification of the customer. The "customization" approach may require additional capacity or technology to eliminate unnecessary finished goods and allow a responsive finishing operation to supply a highly integrated logistics distribution system. These features may also need to be considered at the product design stage. Other policies within this approach include the designed "modularity" of the product and other considerations concerning design for manufacture (Pine, 1993).

Finally, most supply chains actually do not end with the sales to the final customer, because in many cases a recycling chain exists in which the original equipment is dismantled and reused is some form. Again, a management design of this element of the chain must be considered and thus environmental considerations are also challenges that managers of modern supply chains face.

CONCLUSION

Managing the physical flow in the supply chain is key to a successful supply chain strategy. Physical products needed to meet customer orders must be kept moving and should be interrupted as little as possible. To ensure products are kept moving, the information that triggers manufacturing must be reliable and timely. A rapid information flow, as over an extranet or the Internet, will not provide any competitive advantage if the physical manufacture and delivery is slow, unreliable, or not cost-competitive. The Internet will also inhibit the flow of products if the data transmission contains errors or is sent so quickly that the supply chain cannot respond in time. The latter scenario creates a "permanent sense of crisis" for suppliers, and this is undesirable because decisions will be made simultaneously at different companies, increasing the probability that a vital input will not arrive when all others do.

An effective material flow is therefore based on certainty, predictability, and a reliable standardized process of information exchange that is known to every supplier, distributor, and work station. Therefore, an integration of the physical good and information chain is essential, and this system should be shared within (internal supply chain) and beyond the factory (external supply chain). In reality, it is difficult to analyze, manage or optimize any supply chain by treating these design issues in isolation. The results of poor analyses and an incorrect design of the supply chain can therefore be catastrophic and, far from improving customer service and generating competitive advantage, will serve to reduce the reputation of the firm. An integrated approach to the design of the physical and information supply systems is important and can be achieved by a systematic understanding and analysis of how the current supply system performs. All too often, these failings result from poor information and delays and are manifest in terms of long lead times and excess inventories, both of which add to cost but have no value for the customer.

In this section a number of design features have been highlighted, together with the key concepts behind the material and information flow processes that support an effective "order fulfillment strategy." These design issues and the analyses presented in this chapter serve to identify the manner in which the supply chain should be configured to meet the unique constraints of the individual business (its market, technology, and material requirements). For most businesses, this process of design involves an existing system for which information is readily available. Typically, this information about the supply chain and its performance has not been analyzed as a single entity. Instead, functional managers will know elements of the overall system and its problems.

Only when these managers are brought together to redesign the supply chain will these information bases be combined and a true picture of performance achieved. As such, supply chain management requires cross-functional management problem solving such that the sales department explains the process of recruiting orders, the production control department explains how this information triggers material conversion, and the purchasing department explains how these two activities generate supplier schedules. When this integrated design approach is undertaken, it is possible to optimize the flow of materials and information and to create a robust system within which all managers understand how best to keep material moving. This combination of management-level staff is also important if the external suppliers to the firm are to become integrated with the new material flow design and to benefit from it.

Many challenges exist in this process of redesign, and each design is specific to a given company. This process must be undertaken, however, if managers are to fulfill their role as business planners and exercise control of the manufacturing process. At the heart of the design process is the need to integrate the physical and the "triggers" stimulated by information exchange. In an era of compressed manufacturing times, it would be nonsense for delays in information exchange to ransom the material flow. The Internet may well eliminate this problem, but it will also add to the challenges of managing the supply chain if it is not integrated by design. Communicating bad information quickly is a recipe for chaos. Overall, the future will see more electronic information sharing between companies in the supply chain and the Internet, when used to transmit information that has been validated, will offer great possibilities for competitive advantage to those businesses that can translate quality information into the highest levels of customer service. To achieve this, without excess inventory, means sharing data across all tiers in the supply chain and harnessing, by synchronization, the material flow activities at each manufacturer. The Internet, as an enabler for "high-performance" data flow is a reality, as many of the world's leading companies have demonstrated.

GLOSSARY

Bill of material (BOM) A listing of all the subassemblies, intermediates, parts, and raw materials that go into a parent assembly, showing the quantity of each required to make an assembly. A BOM is used in conjunction with the master production schedule to determine the items for which purchase requisitions and production orders must be released.

Finished goods inventory (FGI) Completed products that are not sold or allocated to immediate shipment to the customer.

Forecast Estimation based on extrapolation of historic data and further assumptions about future events. The deviation of the forecast from the actual demand is known as the 'forecast error'.

Just-in-time (JIT) A philosophy of manufacturing based on planned elimination of all waste and continuous improvement of productivity. JIT encompasses the successful execution of all manufacturing activities required to produce a final product, from design engineering to delivery and including all stages of conversion from raw material onward. The primary elements of JIT are to have only what is required when needed; to improve quality to zero defects; to reduce lead-times by reducing setup times, queue lengths, and lot sizes;

to revise incrementally the operations themselves; and to accomplish this at minimum cost.

Kanban A method of just-in-time production that uses standard containers or lot sizes with a single card attached to each. It is a pull system in which work centers signal with a card that they wish to withdraw parts from feeding operations or suppliers. The Japanese word *kanban* loosely translates as "card," "billboard," or "sign." The term is often used synonymously for the specific scheduling system developed and used by the Toyota Corporation in Japan.

Level scheduling In traditional management, a production schedule or master production schedule that generates material and labor requirements that are as evenly spread over time as possible, a core concept of lean production. Also referred to as 'production smoothing'. In some cases, finished goods inventories are used to buffer the production system against seasonal demand.

Logistics The management of the storage, transportation and delivery of goods along the supply chain. Often distinguishes 'inbound logistics', i.e., the movement of components, parts, and materials from suppliers to the original equipment manufacturers, and 'outbound logistics', which relates to the movement of finished (or near finished) products from the manufacturing plant to the customer.

Material requirements planning (MRP) A set of computer-based techniques that uses bill of material data, inventory data, and the master production schedule to calculate requirements for materials. It makes recommendations to release replenishment orders for material.

Milk-round A logistics concept in which parts are collected from several suppliers before being delivered to the point of use. In comparison to direct deliveries from all suppliers, smaller and more frequent shipments can be made. Milk-round delivery is an integral part of the just-in-time concept.

Order fulfillment process The process from order entry to delivery of a product, including order processing, manufacturing, and distribution. (also called the product delivery process).

Pull (system) In production, a pull system is the production of items only as demanded for use or to replace those taken for use. In material control, a pull system is the withdrawal of inventory as demanded by the using operations. Material is not issued until a signal comes from the user. In distribution, a pull system is a system for replenishing field warehouse inventories in which replenishment decisions are made at the field warehouse not at the central warehouse or plant.

Push (system) In production, a push system is the production of items at times required by a given schedule planned in advance. In material control, a push system is the issuing of material according to a given schedule or issuing material to a job order at its start time. In distribution, a push system is a system for replenishing field warehouse inventories in which decision making is centralized, usually at the manufacturing site or central supply facility.

Quick response manufacturing A system of linking final retail sales with production and shipping schedules back through the chain of supply; employs point-of-sale scanning and electronic data interchange and may use direct shipment from a factory.

Scheduling The general process of assigning orders to available production resources. In the automotive context, for example, the compilation of orders from the order bank into weekly and daily production instructions for the particular assembly plants.

Supply chain The process from the initial raw materials to the ultimate consumption of the finished product linking across supplier-user companies or the functions within and outside a company that enable the value chain to make products and provide services to the customer (synonyms: value chain, value stream).

Third-party logistics The use of external companies to perform logistics functions that have traditionally been performed within an organization. The functions performed by the third party can encompass the entire logistics process or selected activities within that process.

Value stream The specific activities required to design, order, and provide a specific product, from concept to launch, order to delivery, and raw materials into the hands of the customer. 'Value stream mapping' refers to the identification of all the specific activities occurring along a value stream for a product or product family.

Waste Any activity that consumes resources but creates no value. The seven wastes are Taiichi Ohno's original enumeration of the wastes commonly found in physical production. These are overproduction ahead of demand, waiting for the next processing step, unnecessary transport of materials, overprocessing of parts because of poor tool and product design, inventories greater than the absolute minimum, unnecessary movement by employees during the course of their work, and production of defective parts.

Work in progress (WIP) A product or products in various stages of completion throughout the plant, including all material from raw material that has been released for initial processing up to completely processed material awaiting final inspection and acceptance as finished product. Many accounting systems also include the value of semifinished stock and components in this category.

CROSS REFERENCES

See *Developing and Maintaining Supply Chain Relationships; Electronic Procurement; International Supply Chain Management; Inventory Management; Strategic Alliances; Supply Chain Management; Supply Chain Management and the Internet; Supply Chain Management Technologies.*

REFERENCES

Bicheno, J. (2000). *The lean toolbox* (2nd ed.). Buckingham, England: Picsie Books.

Bowersox, D., & Closs, D. (1996). *Logistical management: The integrated supply chain process.* New York: McGraw-Hill.

Christopher, M. (1992). *Logistics and supply chain management: Strategies for reducing cost and improving services.* London: Pitman.

Dowling, G. (2002). Customer relationship management: In B2C markets, often less is more. *California Management Review, 44* (3), 87–104

Hines, P., Lamming, R., Jones, D. T., Cousins, P., & Rich, N. (2000). *Value stream management—strategy and excellence in the supply chain.* London: Financial Times/Prentice-Hall.

Holweg, M., & Pil, F. (2001). Successful build-to-order strategies start with the customer. *Sloan Management Review, 43* (1), 74–83.

Kurt Salmon Associates. (1993). Efficient Consumer Response: Enhancing customer value in the grocery industry. Washington, DC: Food Marketing Institute.

Lee, H. L., Padmanabhan, V., & Seungjin, W. (1997). The bullwhip effect in supply chains. *Sloan Management Review 38* (3), 93–102.

Lee, H. L., & Whang, S. (2001). Winning the last mile of e-commerce. *Sloan Management Review, 42* (4), 54–62.

MacBeth, D., & Ferguson, N. (1994). *Partnership sourcing, an integrated supply chain approach.* London: Pitman.

Mather, H. (1988). *Competitive manufacturing.* Englewood Cliffs, NJ: Prentice-Hall.

Monden, Y. (1983). *The Toyota production system.* Portland, OR: Productivity Press.

Ohno, T. (1988). *The Toyota production system: Beyond large-scale production.* Portland, OR: Productivity Press.

Pine, J. B. (1993). *Mass customization: The new frontier in business competition.* Boston: Harvard Business School Press.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance.* New York: Free Press/Macmillan.

Rother, M., & Shook, J. (1998). *Learning to see: Value stream mapping to add value and eliminate muda.* Brookline, MA: Lean Enterprise Institute.

Schonberger, R., & Knod, E. (1997). *Operations management: Customer-focused principles.* Boston: Irwin/McGraw-Hill.

Silver, E., Pyke, D., Peterson, R., & Miltenburg, G. (1998). *Inventory management and production planning and scheduling* (3rd ed.). New York: Wiley.

Simchi-Levi, D., Kaminsky, P., & Simch-Levi, E. (1999). *Designing and managing the supply chain.* New York: McGraw-Hill.

Stalk, G., & Hout, T. (1990). *Competing against time: How time-based competition is reshaping global markets.* New York: Free Press.

Ulrich, K. (1995). The role of product architecture in the manufacturing firm. *Research Policy, 24,* 419–440.

Wight, O. (1995). *Manufacturing resource planning: MRP II.* New York: Wiley.

Womack, J., & Jones, D. T. (1994). From lean production to the lean enterprise. *Harvard Business Review, 72* (2), 93–104.

Womack, J., Jones, D. T., & Roos, D. (1990). *The machine that changed the world.* New York: Rawson Associates.

FURTHER READING

Homepage of the Efficient Consumer Response movement in the grocery industry: http://www.ecr-net.org

Homepage of the Collaborative Planning, Forecasting and Replenishment (CPFR) concept, which allows for online co-ordination of production forecasting and scheduling: http://www.cpfr.org

For more information on lean thinking and value stream mapping: http://www.lean.org

For general information on manufacturing and supply chain management: http://www.ame.org

For general information on MRP systems, production and inventory control: http://www. apics.org

For more information about lean thinking and the latest research into the management of lean manufacturing and supply chain systems: http://www.leanenterprise.org.uk

Marketing Communication Strategies

Judy Strauss, *University of Nevada, Reno*

INTRODUCTION

Would you click on the banner ad in Figure 1? Probably not, but it ran for 12 weeks on HotWired.com, and 30% of all viewers did click on it. The AT&T ad was among the first 13 Web banners in history, all debuting on October 27, 1994. The first Web pages emerged in 1993, even before Netscape and Internet Explorer arrived, providing companies a way to communicate with target audiences. The early pages were dubbed *brochureware* because they closely paralleled material in company brochures. Among the sites to begin in 1994 was *Jerry's Guide to the World Wide Web*—later named Yahoo! From that point, the gold rush was on. Sites such as ESPN SportsZone sold more than $1 million in advertising in 1995, shocking the advertising world into believing that consumers were actually looking at Web pages and that the Internet was well suited for advertising. Next to arrive on the scene were the eyeball counters. Firms such as the Audit Bureau of Circulation (ABC), auditors of print media circulation, joined Internet Profiles and others to measure Web site traffic for firms and advertisers. Next, professional organizations such as the Coalition for Advertising Supported Information and Entertainment (CASIE) formed to help the advertising industry create standards for banner ad sizes, ad cost models, and audience measurement. Finally, the government came on the scene with policies to protect customers while preserving free speech online. Online marketing communication has changed quite quickly in less than 10 years, but its evolution is far from over.

Current User Context

The Internet is ideal for marketing communication because companies can reach 530 million consumers worldwide in the B2C market ("Internet Users Will Top 1 Billion," 2002). Although this represents only 8.5% of the world's population, there is much greater penetration in industrialized nations and among particular demographic and psychographic market segments. For example, nearly 60% of the U.S. population uses the Internet. Beyond that, most businesses in industrialized nations have Internet access to facilitate business-to-business marketing communication. As noted in other articles in this encyclopedia, customers are demanding when it comes to online communication. A few of the more relevant trends affecting online marketing communication follow.

The balance of control has moved from marketers to consumers. Consumers will not wait more than 5–7 seconds for Web pages to download, and they won't spend much time looking for information on a Web site. Furthermore, if the price is not right, they will click away to the competition. Consumers do tend to return to familiar sites, but they are not as brand loyal online as in the brick-and-mortar world. Unless they are extremely motivated to find a particular item, consumers want it easy, fast, and convenient. They also want self-service access to product, account, and other information from any device at any location—24 hours a day, seven days a week—via computers, handheld personal digital assistants (PDAs) such as the PalmPilot, cell phones, or televisions (iTV).

The Internet also created a space for consumers to disseminate their brand experiences and perceptions widely. Many Web sites, such as Amazon.com, provide space for customers to post book reviews, thus playing a role in the success of a new book. Whether consumers post complaints on Internet bulletin boards such as the Usenet (accessed at GoogleGroups) or ePinions.com or send e-mail to their friends on distribution lists, they have a greater influence on brand images than ever before. The result is that marketers must monitor consumer online chat

Figure 1: 1994 AT&T Banner Ad on HotWired.com. Source: Original art reproduction; published with permission from AT&T.

and be accurate in their marketing communication brand promises or be found out under the bright light of the Internet.

Information overload is another important trend. In what some call the *attention economy,* "information is essentially infinite, but the demand for it is limited by the waking hours in a human day ("Encyclopedia of the New Economy," 2002). E-mail in-boxes overflow, Web sites proliferate, and the traditional media and direct mail add to the pile of messages. In this context, marketers must be clever to help consumers find relevant information quickly. For example, Travelocity.com allows users to log onto the wireless Web via a PalmPilot, enter a flight number, and receive instant update on the flight departure time and gate.

Current Marketing Context

Marketers must cut through the haze of online and offline information and persuasive messages to grab the attention of relevant targets at the right time and place. They accomplish this by understanding many basic marketing communication principles and then by using technology to create efficient and effective messages.

A Few Important Basic Principles

Marketers must first have a thorough understanding of their prospect and customer needs and media habits, both in general and as individuals. For example, a return visitor to Amazon.com gets a special page displaying book recommendations based on previous purchases. Second, marketers understand that customers form brand images through many brand contact points. A Sharper Image customer might buy and use a product purchased on the Web site, then e-mail or call the toll-free number to complain about a problem, and finally return the product to the brick-and-mortar retail store. Every contact with an employee, a Web site, a magazine ad, a catalog, the store physical facilities, and the product itself helps the customer form an image of the firm. Successful marketers build and manage a consistent brand image at all these contact points by integrating online and offline strategies. For example, Dell Computer uses a combination of Web site, personal selling, telephone sales, traditional advertising, online advertising, e-mail, and more, to build its brand and product sales. Third, marketers attempt to build long-term relationships with valuable customers. It is 5 times more expensive to acquire a new customer than to retain a current one, so firms direct much of their marketing communication efforts to the process of ensuring customer satisfaction and repeat patronage. Brand loyal customers are much more receptive to a firm's marketing communication than are noncustomers. For more information on how this is done, see the article on customer relationship management in this book.

Finally, marketers use integrated marketing communication (IMC) to create a single-minded focus on the brand promise by delivering consistent messages in all customer communication. Controllable marketing communication tactics in the IMC toolbox include online and offline advertising, direct marketing, sales promotion, marketing public relations (MPR), and personal selling. By definition, personal selling is not possible on the Internet, although the Internet can be used to generate sales leads. Also, a few firms, such as Lands' End, use live chat between a customer service representative and online user—closely simulating the personal selling environment. Although this chapter focuses on online marketing communication, in practice it is part of a larger marketing context.

Technology Enhanced Marketing Communication

Innovative technologies boosted marketing communication effectiveness and efficiency in many interesting ways as discussed in this chapter. Important technologies include the Internet and its multimedia display tool, the Web; databases to hold information; new Web development, browsing, and e-mail software to facilitate Internet communication; and a plethora of digital receiving devices for viewing multimedia messages.

From a marketer's perspective, the Internet is similar in some ways to other media. Like print media, it carries advertising messages along with editorial content from media publishers such as *People* magazine. Also, many corporate Web sites can be compared to catalogs and product brochures except that they are less expensive to print and the messages can constantly be revised. E-mail is similar to postal mail, but faster and less expensive. The Internet can send multimedia communication messages, just like television and radio, but only about 20% of home users have the broadband capability to download video as fast as television does.

The Internet is unique among media in several ways. First is its interactive ability: live chat, e-mail, and response forms on Web sites allow consumers to communicate with firms. Also, Web servers interact automatically with user computers for the purpose of pushing customized Web pages and ads. Second, the Internet can reach global markets in a way not possible without using a multitude of traditional international media. Third, the Internet is a time moderator. This means that companies and customers can communicate 24 hours a day, seven days a week, at their convenience. This feature opened a wide range of possibilities including entertainment and information on demand and customer account and transaction self-service. These consumer conveniences make marketing communication a lot more challenging because of customers' control of the mouse, and the difficulty of gaining customer attention in a sea of information. Finally, computer and Internet technologies allow marketers to measure customer behavior in ways that were previously impossible. Marketers can follow customers as they surf Web sites, view ads, and purchase. When added to customer purchase data from brick-and-mortar retailers and telephone calls to the company, marketers can learn how to retain satisfied customers and sell them more products. In addition to individual customer data, marketers review aggregate data to find marketing opportunities and refine current strategies. These technologies are neither cheap nor perfect, however. Some firms have failed because of the huge costs, and Web site audience measurement numbers are still plagued by technology barriers such as Web site caches, a procedure that stores recently viewed Web pages on a company server to be sent to the next viewer instead of requesting a new page from the Web site. This saves on Internet load but undercounts actual visits to the site.

Although the Internet is widely used for transmitting communication, databases are the information storehouses used to prompt marketing communication messages. Firms build extensive customer, prospect, and product databases for creating personalized e-mail and customized Web pages. Thus, Amazon.com can present an individualized Web page at a moment's notice by reading a previously placed cookie file on the user computer, accessing the customer record in the database, and sending appropriate product and customer account information to create the Web page—all in a nanosecond.

The following sections present four important forms of online marketing communication: advertising, direct marketing, sales promotion, and marketing public relations.

ADVERTISING

eDietShop manufactures a variety of gourmet diet foods, including sugar-free products for diabetics. The founder, Steven Bernard, estimates that there are 15 million diabetics in his market and devised an online advertising plan to draw them to his Web site (Rosen, 2001). He arranged banner ad exchanges and sponsored e-mail newsletters for groups such as DiabeticGourmet.com. As well, he used key-word advertising on search engines such as GoTo.com, so that users saw his banner ad when searching using the terms "sugar free" and "low fat." Bernard negotiated a pay-for-performance pricing model so that he only paid for visitors who clicked on the search engine ad, visiting eDietShop.com. In late 2001, he spent $10,000 a month on online advertising to generate $150,000 in sales. Bernard's recent advertising buy on the cable TV Food Channel drove traffic and site transactions such that the company now enjoys revenues 10 times those experienced 2 years earlier.

This example demonstrates how carefully targeted online advertising, when combined with traditional media, can build awareness, drive traffic, and increase sales for an online retailer. These and many other online advertising tactics are discussed in the following sections.

Advertising Spending for Various Media

Online advertising in the U.S. grew quickly from its humble beginnings in 1994 to an estimated $7.5 billion in revenues during 2001 (Figure 2). Expenditures dropped nearly 7% from 2000 to 2001, reflecting the U.S. economic recession. Compared with the 9.8% drop in all U.S. advertising expenditures, online advertising fared well and continues to do so ("No surprise," 2002). It is notable that most online advertising concentrates on a dozen top portal and news sites; for example, the top three, Yahoo!, AOL, and Excite, sold $791 million (10.5%) of the online space (Saunders, 2002). Similarly, in April 2002 three industry sectors placed 62% of all online advertising: Web media (25%), retail goods and services (24%), and financial services (13%) ("Leading industry advertisers," 2002).

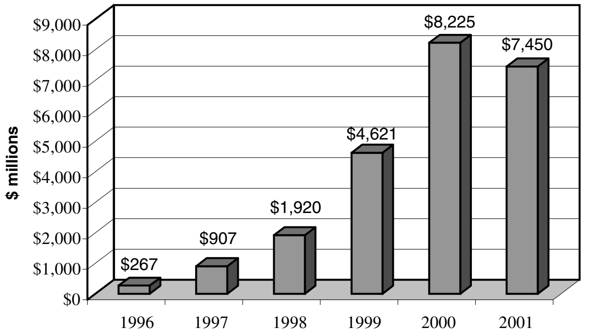


Figure 2: Annual Internet advertising spending: 1996–2001. Sources: Data are from PricewaterhouseCoopers. (2002). *IAB Internet advertising revenue report: Third quarter 2001 results.* New York: PricewaterhouseCoopers New Media Group; and "eMarketer designates Interactive Advertising Bureau/PricewaterhouseCoopers quarterly Internet ad revenue report as its benchmark source for online ad revenues (2002).

On average, an advertiser only spends 7.5% of his or her media budget on the Internet ($7.5 billion total Internet of an estimated $99.8 for all advertising spending). There are few *average* firms when it comes to Internet advertising, however. A small number of advertisers, especially niche firms, spend nearly all of their budgets online, whereas others rarely use the Internet (such as large packaged goods firms). As shown in Figure 3, online advertising has surpassed outdoor expenditures (e.g., billboards, bus cards), but traditional media still receive the lion's share of advertiser budgets. Why is this? First of all, although the Internet reaches 30–60% of the residents in the world's industrialized nations, it does not yet enjoy the nearly 100% penetration of television. Second, although advertisers can easily track user behavior online, Web site audience measures as a whole are still imperfect. Without good measures, advertisers and agencies do not know how

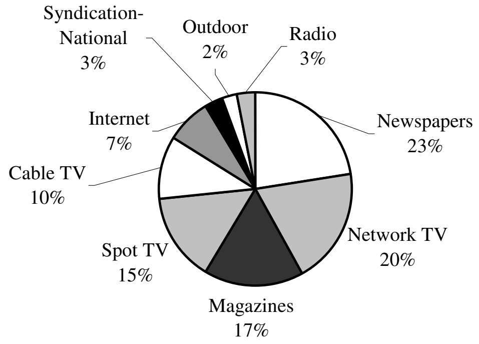


Figure 3: Proportion of spending for various media of $99.8 billion in 2001. Source: Data are from "A Classic Scenario: Bad News, Good News," (2002, June 18), *eMarketer.*

Table 1 Strengths and Weaknesses of Various Media

CRITERION	TV	RADIO	MAGAZINE	NEWSPAPER	DIRECT MAIL	WEB
Involvement	Passive	Passive	Active	Active	Active	Interactive
Media richness	Multimedia	Audio	Text and graphic	Text and graphic	Text and graphic	Multimedia
Geographic coverage	Global	Local	Global	Local	Varies	Global
CPM	Low	Lowest	High	Medium	High	Low
Reach	High	Medium	Low	Medium	Varies	Medium
Targeting	Good	Good	Excellent	Good	Excellent	Excellent
Message flexibility	Poor	Good	Poor	Good	Excellent	Excellent

Source: adapted from Strauss and Frost (2000). CPM = cost per thousand.

many of their target can be reached at a Web site. Some of the problems include inability to measure Internet usage from work (i.e., behind corporate security systems); inflated measures due to search engine robots and other automated software hitting Web pages 24 hours a day; and problems identifying unique users from Internet service providers such as America Online. The Interactive Advertising Bureau and others are working to solve these problems, and when coupled with measurement companies such as Jupiter Media Metrix, audience figures presented by Web companies continue to improve. Finally, firms select from a variety of media because each medium has unique properties to reach particular markets and achieve specific communication objectives (Table 1). Notably, television is best for creating broad awareness and building brands with the masses, radio and newspapers are best for local advertising, and the Internet, cable TV, and direct mail are good for narrow targeting. Of these, the Internet is best at reaching individuals with personalized two-way communication for a low price per thousand impressions.

Online Advertising Formats and Vehicles

Most people think of banner ads as the only type of Internet advertising, yet they were only 36% of online expenditures for the first 9 months of 2001 (PricewaterhouseCoopers, 2002). Banner use has declined in favor of online sponsorships (27%), classified ads (16%), and a number of new types of online advertising (Figure 4). The standard banner size is 468 wide by 60 pixels tall (as depicted in Figure 1), but recently advertisers have been using larger formats, such as *skyscraper* (160 by 600 pixels) and *large rectangles* (360 by 300 pixels). (A pixel is one dot of light on a computer screen.) Sponsorships are particularly interesting online because advertisers create content for Web pages that is context relevant. For example, Kraft provides recipes using its brands, paying to insert them onto sites for women. This technique blurs the line between editorial and advertising content, something most traditional media will not do. On one hand, it is better for customers because they get more relevant content, but conversely, sponsored content can undermine the perception of content credibility, especially if used at media Web sites. On corporate sites, consumers perceive it all as advertising anyway, so probably little harm is done.

Key-word advertising at search sites and e-mail are two important ways to advertise on the Internet, as represented in the eDietShop example. E-mail advertising is usually a text message an advertiser pays to include in someone else's e-mail newsletter or other communication. Yahoo! offers free e-mail because of the inclusion of text ads in the e-mail exchanged by customers. Keyword advertising is a brilliant strategy because it presents banner ads or links on a search query return page that are relevant to the topic the user is researching. Google.com takes it a step further by ordering the search query return page key-word banners by popularity. Thus, the most relevant ad tops the list of four or eight on the page.

New online advertising formats emerge continuously, showing that evolving technology creates many new opportunities. Some of these include a variety of banner shapes, sizes, and interactive options, text sponsorship of news to wireless devices, branded games, and Shoshkeles. The latter are screen interrupts—floating images that run through the Web page to grab user attention. The first floating ad featured the Energizer Bunny, hopping though the page text and graphics. A recent floater appeared at Speedvision.com. After arriving at the page, the screen slowly darkened to the point of making the text

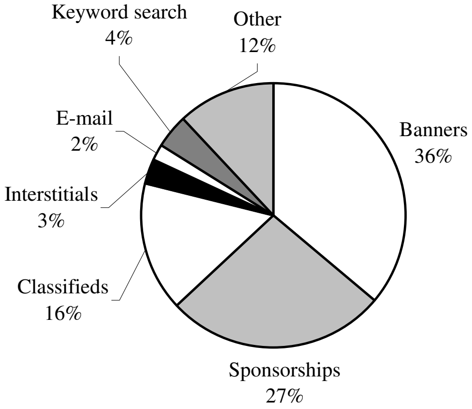


Figure 4: Proportion of online advertising spending by format for first nine months of 2001. Source: PricewaterhouseCoopers (2002). *IAB Internet advertising revenue report: Third quarter 2001 results.* New York: PricewaterhouseCoopers New Media Group.

unreadable, and then a hand emerged at the lower left corner and sparked up a Zippo lighter to illuminate the page again before disappearing. These Shoshkeles are novelties that have big impact, but when the novelty diminishes, consumers may perceive them as annoying interruptions to their online tasks. Finally, a new gimmick involves contextual advertising. This employs highlighted words on a Web page that are linked to either a pop-up window ad or the advertiser's site. Originally thought a perfect example of relevant advertising, contextual ads are now discussed as the next generation of spam because of the way they interfere with the text and lure innocent users into unknowingly clicking on something unwanted. It should be noted that each new advertising format creates an initial surge of user excitement and attention; as the cutting edge becomes mainstream its ability to capture eyeballs diminishes.

Advertising Pricing Models

Most advertisers and media use a pricing model called cost per thousand (CPM). It is calculated by dividing the advertisement cost by the number of viewers, and multiplying by 1,000. It is particularly powerful because advertisers can use it to compare advertising efficiency across a multitude of media vehicles to find the best buy. Unlike traditional media, only 50% of online advertising is purchased using the CPM model (PricewaterhouseCoopers, 2002). When applied online, CPM becomes cost per thousand *impressions;* an impression is one rendering of an ad on the user's computer screen, either in text or graphics. Departing from most traditional media, 13% of online advertisers pay based on performance, and the remainder use some combination of the two models. Performance-based payment includes schemes such as payment for each click on the ad, payment for each conversion (sale), or payment for each sales lead. Obviously, this type of pricing is attractive to advertisers because they only pay for effective ad placements; Web publishers take a bigger risk with performance pricing because their revenue depends partially on the strength of the advertiser's banner and product itself.

How much does it cost to advertise online? Typical prices are $7 to $15 CPM (Hallerman, 2002). Google charges $8 to $15 per thousand impressions for placement on its key word search (called AdWords). The number of dollars an advertiser can spend at Google will widely vary depending how popular the key words are, ranging from $10,000 a month to half a million dollars. Compared with typical prices for traditional media ads, Web advertising costs about the same to reach a thousand viewers as on national television, more than on radio, and less than in newspapers and magazines. Obviously this varies widely based on a number of factors and tends to rise with more narrowly defined and hard to reach targets. More important than cost, however, is advertising effectiveness.

Advertising Effectiveness on the Internet Medium

There is much debate about the effectiveness of online advertising compared with other media, and this keeps some advertisers at bay. The answer to whether online advertising works depends on the specific question, starting with the most basic: whether the advertiser wants to employ advertising for a direct response or for brand building. The Internet started as a direct marketing medium, and when viewed from this perspective, banners are virtually ineffective with their average 0.5% response click through. There are lots of exceptions, however (e.g., the Mexican Fiesta Americana Hotels got a 10.2% click through by narrow targeting of Americans living in seven eastern states who had just purchased an airline ticket to Cancun and were online between 2 p.m. and 7 p.m. Monday through Wednesday; Strauss & Frost, 2000). Also, using Nielsen/NetRating statistics, AdRelevance studied lag time between viewing an ad and purchasing at the advertised site—a metric known as *conversion* (Saunders, 2000). The study found that although 61% of those who did click purchased within 30 minutes, 38% purchased within 8 to 30 days after viewing an ad but without clicking on it. Thus, when people click, they do buy.

The Internet is now viewed by many as an effective branding medium; banners have been found in many research studies to increase brand awareness and message association and to build brand favorability and purchase intent (see http://www.iab.net). In three studies, online ads that were bigger, that contained rich multimedia, or that were placed as interstitials delivered an even greater impact. Interstitials are ads that display before a page loads—during the transition from page to page. For instance, skyscrapers and large rectangles were found to be 3 to 6 times more effective than standard size banners in increasing brand awareness (Pastore, 2001).

There is increasing evidence that online advertising works well with traditional advertising, thus helping to validate the concept of integrated marketing communication. Unilever advertised its Dove Nutrium Bar in print, television, and online. Using an outside research firm, it found that increasing the number of online impressions from 6 to 12 over a 6-week period increased branding effectiveness by 42% ("Internet Is Powerful," 2002). It further found that each of the three media were effective at building brand awareness and purchase intent, but that online was the most cost-efficient.

Performance Metrics

The only way to tell whether online initiatives are effective is to measure the results. The decision about what to measure arises from the marketing communication goals. If the firm wants to build brand awareness or position the brand, a survey will measure whether the numerical goals were met. Conversely, the Internet is particularly well suited to behavioral measures. A few common metrics are displayed in Table 2.

As an example, the catalog and online retailer iGo purchased more than $3 million in advertising from a number of sites in 2000 (Strauss & Frost, 2000). Their budget estimated 647 million impressions with an average $4.64 CPM. Before the purchase, iGo estimated an average 0.54% click through for 3.49 million visitors to the iGo site. If they could get a 1.27% conversion, this would result in 44,255 orders. At an average order value

Table 2 Commonly Used Performance Metrics for Advertising Effectiveness

METRIC	DEFINITION/FORMULA
Response Time	Time between sending e-mail or viewing ad and click-through response
Cost per Action (CPA) 1	CPA = ad cost ÷ number of people taking action, such as registering or purchasing at the site
Click-Through Rate (CTR)	Number of clicks as percent of total impressions CTR = Clicks ÷ Impressions
Cost Per Click (CPC)	Cost for each visitor from ad click CPC = Total Ad Cost ÷ Clicks
Conversion Rate	Percent of people who purchased from total number of visitors Conversion Rate = Orders ÷ Visitors
Cost Per Order Equation (CPO)	Cost of each order resulting from click-through visit CPO = Total Ad Cost ÷ Orders
Customer Acquisition Clost (CAC)	Total marketing costs to acquire a customer
Average order value (AOV)	AOV = sales in dollars ÷ number of orders

of $140, they estimated this campaign to create a $67.83 cost per order, $0.86 cost per click, and $566,895 in profits. The big bonus, however, was the additional 3.5 million customers that might be lured to repeat purchase. Advertisers continually use metrics in this manner to fine-tune their online advertising campaigns. Fortunately, technology helps with the counting.

E-MAIL DIRECT MARKETING

Using a list of 200,000 fans who opted into a marketing communication list at the 'N Sync Web site, Zomba e-mailed a message announcing the new *No Strings Attached* album 1 month ahead of its national release (Weintraub, 2000). The e-mail featured a downloadable video with a brief voice message from band members and a sampling of one song. More than one third of the recipients downloaded the video, nearly 90% clicked on one of the embedded links to get more information, and thousands shared the e-mail with their friends. This e-mail campaign helped build anticipation and demand for the album: In its first week, 2.4 million copies were sold.

This example demonstrates the Internet's power to build databases and use them for marketing communication that builds brands and transactions (when the fans signed up online, their contact information went into a customer database for later use). Marketing databases and their uses are discussed in other chapters in this encyclopedia, so it will suffice here to say that they are extremely important for generating personalized marketing communication, delivered at the right time to the right person over the Internet. This section focuses on e-mail because it is a strong direct marketing communication tool unique from all others (such as interactive banner ads).

E-mail allows users to send messages from one computer to another over a network such as the Internet. With 8 billion e-mails a year flying over the Internet, the typical user spends nearly an hour a day managing e-mail—more than one third of all time online. Marketing related e-mail is 22% of a typical Internet user's in-box (Mardesich, 2001). About half of this is requested, and the other half is unwanted spam. Many marketers use *outbound e-mail* to send sales announcements, offers, product information, newsletters, and other marketing communication to e-mail addresses—sometimes to one individual (called 1:1 marketing), sometimes to large groups. *Inbound e-mail* allows customers to interact with marketers by asking questions, placing orders, and complaining about products and services.

Outbound E-mail

Although e-mail is often considered the Internet's "killer app," e-mail marketing is still in its infancy. As shown in Figure 5, there are approximately 150 million postal mail addresses in the United States that can be associated with specific individuals ("The value of a Corporate,"

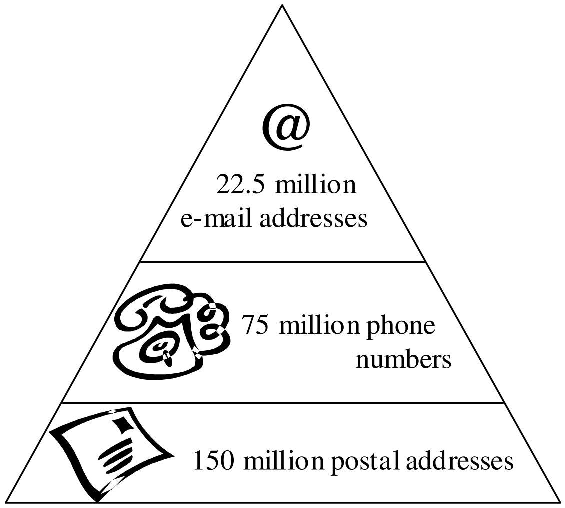


Figure 5: Number of outbound direct marketing channels matching database names. Source: Data are from "The value of a corporate e-mail address" (2001).

Table 3 Metrics for Electronic and Postal Mail

	E-MAIL	POSTAL MAIL
Delivery Cost per Thousand	$30	$500
Creative Costs to Develop	$1,000	$17,000
Click-Through Rate	10%	N/A
Customer Conversion Rate	5%	3%
Execution Time	3 weeks	3 months
Response Time	48 hours	3 weeks

Source: Jupiter Communications as cited in "E-mail and the different levels of ROI" (n.d.).

2002) for direct mail purposes. About half of those can be matched with individual phone numbers for telemarketing purposes, but only 15% of postal addresses can be associated with e-mail lists that will identify particular individuals. Although nearly 50% of the U.S. population has one or more e-mail addresses, at this time it is difficult to match a list of these with individual customers and prospects in a firm's database.

Regardless of this difficulty, many marketers include e-mail marketing in their communication campaigns. This is partially because e-mail is faster and cheaper than postal mail (Table 3). Another advantage of outbound e-mail to individuals is the high level of automated personalization. In addition to using a customer's name, e-mail can be sent based on some event trigger and include details of the customer's previous purchasing behavior. For example, Amazon sends e-mail to selected customers to announce a new book by a previously purchased author. Newsletters sent to larger groups of customers represent another successful use of e-mail marketing. Eighty percent of U.S. online customers enjoy receiving electronic newsletters ("e-Consumers prefer eMail," 2001). As with individual e-mail, the content is as important as identifying the right target for an effective campaign.

According to the Jupiter Communications study in Table 3, e-mail enjoys a 10% click through to the sponsor's Web site and a 5% conversion rate. In actuality, e-mail effectiveness varies widely, and some believe that these metrics are declining. Nevertheless, click-through rates as high as 90% have been reported by Ticketmaster for a mailing to Bruce Springsteen fans who had already bought an upcoming concert ticket. Success goes to those who send relevant information on a timely basis to consumers and business customers who want to receive it.

Through *permission marketing,* marketers identify individuals who are interested in the firm's offerings as well as demonstrate respect and an understanding that recipients can control the information flow. To build good relationships, companies allow users to *opt in* (ask to be put on the mailing list) as well as to *unsubscribe* (take themselves off the list). Some marketers use *opt out* (users uncheck a box on the Web page if they don't want to receive e-mail). This practice doesn't always create goodwill because most users scan a page quickly and don't notice opt out boxes, thus receiving e-mail they did not want. One study conducted by DoubleClick found that 88% of online customers have made a purchase after receiving permission based e-mail, and 40% of the study respondents said they were loyal to particular online retailers because of this e-mail (DoubleClick Inc, 2001).

Managing Privacy

Privacy is a critical concern when managing customer databases. Savvy firms collect data from customers, use it to provide relevant information and services that make customer lives easier, and build trust by keeping this information private. This means not sharing customer e-mail addresses or personal information with others unless given permission, and many firms have received a huge customer backlash when they violated this principle. It also means guarding credit card information. A good privacy policy is in the marketer's best interest because a shared e-mail list soon loses its power through overuse, and customers prefer to do businesses with firms they can trust. Many online firms use third-party seals, such as available at Truste.com, to validate their well-thought-out and communicated privacy policies.

In addition to information privacy, firms must honor customer wishes about not intruding on the privacy of their private spaces. Telemarketers who have displayed no regard for customer privacy when they call a person's home have set the groundwork for customer ire over spam. As previously mentioned, opt-in databases yield higher response rates than non-permission-based lists because of the respect shown for customer privacy desires.

Inbound E-mail

There is only one critical thing to say about inbound e-mail: If marketers publish e-mail addresses on a Web site or print materials, they must answer inbound e-mail communication in a timely manner. Otherwise, they are better off not publishing e-mail addresses. Several studies have found that up to half of all companies do not answer their e-mail from consumers, or send inappropriate automated responses. For instance, Jupiter Communications found that 40% of online-only retailers took longer than 3 days or did not answer inbound customer service inquiries ("Value sites win," 2002). Conversely, 46% responded within 24 hours. Obviously, this form of direct-marketing communication is time intensive, 1:1 marketing that can make a big impact on customer satisfaction. For these reasons, many firms have Frequently Asked Questions (FAQ) and other procedures for customers to use self-help problem solving before sending e-mail. For example, Dell Computer asks customers to enter a service code into a Web page query; this identifies both their customer record and specific computer model so that automated relevant responses can be sent as a first step to solving the problem.

Viral Marketing

Viral marketing is a powerful e-mail variation with an awful name; it refers to e-mail messages specifically designed for recipients to forward to friends, family, and others on their e-mail lists. Viral marketing is an electronic version of traditional word-of-mouth product communication. Like all e-mail campaigns, viral marketing costs little, can include engaging multimedia elements, can be highly targeted, and the results can be easily tracked. The main advantage, however, is that it works. Hotmail is a viral marketing legend because the free e-mail service had signed up 12 million subscribers within 18 months of introduction, primarily because of a line similar to this on all e-mail messages: "Join the world's largest e-mail service with MSN Hotmail. Click Here." This is the current Hotmail message line because the firm was sold to Microsoft (MSN) for $400 million in stock a few days after the 18-month mark. There are many viral marketing success stories, especially in the entertainment industry—for example, both the *Blair Witch Project* film's success and the six-month 0 to 50% MP3 adoption rate have been largely accredited to viral marketing.

Viral marketing must be carefully crafted to work, however. To be successful, marketers must give away something useful or fun, make it easy for users to send it to others, and must tap into basic consumer needs or desires. Also, marketers must be ready to handle the wildfire of orders or requests should it be successful. Hotmail, for example, quickly had to upgrade its server capability to handle all the new e-mail accounts. Finally, some marketers emulate viral techniques when they pose as consumers and post favorable product comments on Web sites, such as ePinions.com. Similarly, some firms pay consumers to rave about products online. These practices will reflect poorly on marketers if discovered and demonstrate that the Internet is still a place where buyers must judge information sources carefully.

Short Message Services (SMS)

SMS is a way to communicate via the Internet without using e-mail software. Marketers and others send messages up to 160 characters in length that are stored until receivers open their Internet connection, receive, and view them. These are different from instant messaging, in which senders and receivers must be online at the same time in a 1:1 chat type of format. SMS's big use occurs with mobile phone users, not computer users. For example, any cell phone user can send an SMS to a friend by simply entering the message and phone number, and the friend's phone might make a ringing sound when it arrives. By one estimate, as many as 200 billion short text messages will appear on mobile phones worldwide on a monthly basis by the end of 2002 (Silk, 2001). The biggest usage occurs in Europe and Asia: In the United Kingdom alone, SMS is the technology of choice for peer-to-peer communication, with users currently sending an estimated 1 billion messages a month.

The big question is how to commercialize SMS without spamming users. Companies such as McDonalds, Cadbury, 20th Century Fox, and MTV have already used this new technique and enjoyed conversion rates of up to 10% (Silk, 2001). That is, users can click on an embedded SMS hyperlink to purchase products or perform other requested actions. To be successful, it is important that the messages be permission based, short, personalized, interactive, and fun. One study of 3,300 people in 11 countries discovered four important criteria for consumer acceptance: choice to receive messages (permission-based), control over viewing or deleting individual messages, filter options to bypass selected messages, and mutual benefit such as receiving discounts (Pastore, 2002). An even more powerful approach is to get users to initiate the SMS, as demonstrated in the sales promotion example that follows.

SALES PROMOTION OFFERS

Heineken, the global beer brand, used wireless Internet technology to capitalize on the British pub tradition of quiz nights. Typically a quiz night consists of a loyal pub customer shouting out a series of questions to which other customers answer on paper score sheets. Winners receive free pints or meals. Using a combination of online and offline promotion, Heineken placed point-of-purchase signs in pubs inviting customers to call a phone number from their cell or other mobile device and type in the word "play" as a text message (SMS). In response, the customer received a series of three multiple-choice questions to answer. Getting all the questions right scored a food or beverage prize to be redeemed by giving a special and verifiable number to the bartender, which 20% of all players realized. "Feedback was that it was a great promotion . . . consumers found it fun and sellers found it to be a hook," said Iain Newell, marketing controller at Interbrew, which owns the Heineken brand.

Online sales promotion offers can build brands, increase Web site visitors and length of stay at the site, build databases, and support increased online or offline sales, as in the Heineken example. As with other types of marketing communication, most promotional dollars are spent offline; however, a few important online tactics are emerging in the consumer market: coupons, sampling, contests, and sweepstakes (such as Heineken used).

Coupons

Web sites such as hotcoupons.com, coolsavings.com, and valupage.com offer searchable and printable coupon databases, organized by type of retailer and zip code. For example, participants in CoolSavings include JC Penney, Toys 'R' Us, Domino's Pizza, and H & R Block. Other firms offer e-mail coupons. J. Crew, the catalog retailer, provides e-mail coupons as an apology to customers who complain via e-mail; the customer simply provides the coupon number to receive a discount on the next purchase. In their survey of nearly 5,000 consumers, Valentine Radford found that 55% of online users prefer to receive e-mail coupons, compared with 30% preference for newspapers and 18% for snail mail ("e-Consumers prefer eMail," 2001). For some reason, however, the online coupon market has not grown to realize its earlier expectations. Perhaps this is because straight product discounts are easier for e-commerce firms to fulfill than are coupons. The author's observation is that most Web coupon sites provide this service as an afterthought accompaniment to their print coupon services.

Sampling

The Internet is particularly well suited for sampling of digital products. For instance, software firms commonly offer 30-day free trials, betting that potential customers will invest time in learning the software and appreciate its value through use. Software trials at Adobe.com, music samples at CDNow.com, daily stories only at NYTimes.com, and press releases with choice tidbits about research findings at Gartner.com are all examples of firms that sample products to entice purchase.

Some firms selling tangible products also use the Internet for offering free samples and facilitating fulfillment. For example, when Procter & Gamble decided to relaunch Pert Plus shampoo, it created a Web site to offer the samples. Within 2 months, 170,000 people visited and 83,000 requested samples (Williamson, 1999).

Contests and Sweepstakes

These tactics build excitement about a brand and entice users to visit and hang around for awhile at the firm's Web site. Many sites offer contests—a game of skill or chance and sweepstakes. For example, Orbitz.com drew 1.9 million customers in its first month of operation during June 2001 partly because of its huge sweepstakes and the accompanying radio advertising. Every site visitor that registered was eligible for the free round-trip ticket given away every hour, 24 hours a day, 7 days a week, for six weeks (Orbitz, 2001). According to Jupiter MediaMetrix online sweepstakes appear to be growing in popularity: In early 2001, 36% of the most successful new Web sites were sweepstakes sites (McAllister, 2001).

MARKETING PUBLIC RELATIONS (MPR)

During the week of the ninth MTV Music Awards in 2000, MTV.com site traffic increased by 48% to 1.1 million visitors as people logged on to learn more about the stars nominated for awards (Carr, 2000). Capitalizing on the Web's strength to provide brand information and build community, the 10th MTV Movie Awards invited viewers to cast their votes at MTV.com for award nominees during a 3.5-week period before the 2001 event. After the televised event, the Web audience could log on to the Web site to see digital photos and video clips of event celebrity interviews, to read gossip, to provide opinion about the event, and to vote for the best dressed actors and favorite event moments. Using viral marketing, MTV.com launched the "2001 Movie Award Sweepstakes," with its VIP trip to the event, complete with makeover and wardrobe. An MTV.com user was entered into the sweepstakes each time he or she sent a friend to the site. The upcoming eleventh annual 2002 MTV Movie Awards will be seen by 382 million households in 165 countries and 18 languages.

As previously mentioned, MPR is the brand-related portion of all public relations activities, directed to customers and prospects. A few of the more important online MPR tactics include the Web site content itself, product publicity, online community building, and online events.

Web Site

Every organization, company, tourist destination, or brand Web site is considered a public relations (PR) vehicle for reaching customers and other stakeholders. This is not advertising, because the company is not paying to place its message on someone else's site. A Web site is MPR because it serves as electronic brochure, including current product and company information. Web sites can perform retailing functions when they include e-commerce transactions, but every site provides a PR function as well. For example, Butterball's site (http://www.butterball.com), which features cooking and carving tips, received 550,000 visitors in 1 day during a recent Thanksgiving week.

It is not enough to build, maintain, and put a Web site online; an organization must draw traffic to the site to achieve its objectives. There are many ways to do this using online and offline marketing communication techniques; however, one is unique to the online environment: search engine optimization (SEO). Nearly 47% of all Web users claim that the most common way they find products or online stores is through search engines (Jupiter Media Metrix, 2001). When presented with a query result page, consumers do not usually look beyond the 10 links on the first page. Thus, many firms use SEO to ensure that their site will appear on the first results page from relevant searches. To do this, first firms decide what key words describe their Web sites. Then they use these key words in hidden HTML tags sought by search engines (Meta tags). Finally, they carefully craft the text on their pages to reflect this content, including even purposefully using different spellings of key words (e.g., email and e-mail). These and other SEO strategies, when coupled with keyword advertising buys at search engine sites, help build awareness and draw traffic to a firm's site.

Brand Publicity

A huge part of MPR includes news releases and other product publicity included in the media and on Web sites not owned by the firm. Most media firms accept news releases electronically, thus saving time, printing, and postage. The top-10 general news Web sites attract 42 million annual visitors, and online services charge less than $300 to submit electronically a company's news release to a carefully targeted listing of news sites (Helperin, 2001). This does not include the millions of wireless users worldwide who receive SMS headlines on their handheld devices. The firm's own Web site is especially important for disseminating news during a crisis, as when traffic to the Web sites of Firestone and its parent, Bridgestone, increased from a small number to 258,000 the week after Firestone announced a massive tire recall (Carr, 2000). Note that not all product news releases are in print form on a Web site; many firms Webcast a live presentation by a firm's spokesperson, complete with live questions and answers. Michael Dell, Bill Gates, and many others use this technique.

Community Building

Community building involves maintaining good community relations or contributing to a sense of community within the customer base. Community building is more than online chat rooms and e-mail. It involves firms gaining a following of devoted customers, such as the Harley Owners Group (HOG), Napster.com users, and Macintosh computer users. Consider the film industry's ability to build community with *Star Wars*, *Star Trek*, and more recently *The Blair Witch Project*. The latter is notable because the community was not built by the film's producer, but a grass-roots movement formed around a film that got mixed reviews. Viewers loved seeing the film and consuming fictional artifacts online (journals and videotapes). Everyone in the large community talked about it in school, in families, and in e-mail, and these interactions were at least as important as the actual movie-going experience.

Online Events

In addition to integrating a Web site with an offline event, such as the MTV Movie Awards, many sites host events directly on their Web sites. The Victoria's Secret Online Fashion Show made history in 1999 when it drew 1.2 million visitors, thus, increasing Web traffic by 82% after advertising in the *New York Times,* Super Bowl football game, and other traditional media. After solving many technical glitches, Victoria's Secret continued the tradition in subsequent years, Webcasting steaming video from places such as Cannes, France. On November 13, 2001, the models performed first in front of a live audience, then the show was telecast on ABC television network 2 days later, and finally highlights and photos were shown online. Site traffic doubled the week of the fashion show (Bailly, 2001). Online events such as these and many others will become even more widespread as broadband connectivity adoption expands.

INTO THE FUTURE

Jupiter Media Metrix, Forrester Research, and others predict that by 2006 U.S. companies will spend somewhere between $35 and $60 billion annually on digital marketing campaigns such as the online tools described in this article. Digital campaigns integrate online advertising, promotions, and e-mail strategies. The following are five important trends the author expects to have a huge impact on online marketing communication in the next 10 years.

Convergence

The medium is no longer the same as the receiving appliance. For instance, television programming will not always be restricted to viewing on television sets. When broadband Internet access to homes reaches more than 50%, television, radio, and other high bandwidth programming can be received by the masses over the Internet. Coincidental to this is appliance convergence, which is already occurring in the handheld market with PDA and cell phone convergence. This trend will continue with television sets, radios, and personal computers. It is uncertain what the entertainment and information receiving devices will eventually look like, but it is certain that marketing communication must adapt to flow with the content users want.

Mobile and Wireless Marketing Communication

Weighing in at only 1% of online advertising (Figure 4) are Interactive TV and wireless ads. Although the former seems dead in the water, wireless ads may well see stunning future growth. For example, one study revealed that 88% would be responsive to receiving an electronic coupon on their mobile wireless device when near a retailer that could redeem it (Pastore, 2002). This involves a combination of GPS technology to locate the roaming user and Internet connectivity to a handheld device. Soon a consumer nearing a McDonald's outlet will receive a message offering a free soft drink with the purchase of a meal, thus reaching the customer when and where the communication can provoke the desired response of stimulating a purchase. If the movie *Minority Report* is any indication, consumers could someday receive customized messages without GPS and handheld devices. This film depicted computers reading consumer retinas while they walked in the mall and then delivering customized messages: wireless taken to the extreme.

Information on Demand

The U.S. passed legislation that all television programming will soon be sent in digital format. When combined with the growth in broadband adoption and increasing customer control over information, this puts the final nail in the coffin for marketing control over consumer attention. Customers will request information and entertainment on their schedules, and marketers will have to be increasingly clever to attract customers with communication messages. In this climate, brand loyalty is translated into relationship capital—it becomes critical to the firm's success.

Micromarketing and Privacy Rules

The trend toward 1:1 marketing made possible by Internet and database technologies will continue. Companies will learn by observing individual customer behavior and use those data to present relevant and timely offers. In this climate, companies that build trusting customer relationships by honoring personal information privacy will ultimately win.

Customer Disappearing Act

It has been said that moose instinctively know when hunting season has begun in Alaska because they suddenly disappear from view. Customers, tired of telemarketing, spam, and other unpleasant marketing communication techniques, are beginning to disappear like moose. For example, 37% of Internet users who have been online for 3 years and use the Net daily use fake e-mail addresses when registering at Web sites (Mardesich, 2001). Ad blocking software, such as AdSubtract, already has 2 million customers that use the software to block banner ads from being delivered with Web sites. This parallels TiVo and other television commercial blocking recorders. It probably will not suffice to educate consumers that if they don't accept advertising, they will pay more for programming and information because they bought that argument for television and now pay anyway for Cable, not recognizing the fine points. Creating new advertising formats that outsmart the blocking software, such as pop-up windows, is a cat-and-mouse game that doesn't help increase customer satisfaction. Thus, marketers will need to develop new models for communicating with prospects and customers. These models will respect customer needs for information and entertainment, A^3—anytime, anywhere, anyway.

GLOSSARY

Advertising Nonpersonal communication of information, usually paid for and usually persuasive in nature, about products (goods and services) or ideas by an identified sponsor through various media.

Average order value (AOV) The dollar sales divided by the number of orders.

Click-through percentage The percentage of clicks from total number of impressions.

Conversion rate An equation that tells a firm what percentage of the visitors to the Web site actually purchased, calculated by the number of orders divided by number of visitors.

Cookies Small text files that sites put on the user's hard drive for future access by the Web server.

Cost per action (CPA) The cost for an advertisement divided by the number of people taking some action, such as registering or purchasing at the site.

Cost per click The total advertisement cost divided by number of clicks on an ad or hyperlink.

Cost per order The total ad cost divided by the number of orders.

Customer acquisition costs (CAC) Costs for advertising and all other marketing costs divided by number of customers.

Cost per thousand (CPM) an equation used to compare the efficiency of competing media space, calculated by dividing the advertisement cost by the number of viewers, then multiplying by 1,000.

Floating ads Advertising images that run through a Web page screen, interrupting the content to grab user attention (called Shoshkeles by the firm introducing the concept).

Integrated marketing communication (IMC) A cross-functional process for planning, executing, and monitoring brand communications designed to profitably acquire, retain, and grow customers.

Key word advertising A form of Web advertising in which advertisers purchasing particular query words at search engine sites so that banners will appear when those words are queried by users.

Marketing public relations (MPR) Brand-related activities and nonpaid, third-party media coverage to influence target markets positively.

Opt-in A form of e-mail marketing that offers consumers incentives to accept information in e-mail messages, usually by checking a box on a Web page.

Opt-out A form of e-mail marketing that forces users to uncheck a Web page box if they do not want to receive e-mail messages.

Permission marketing Presenting marketing communication messages only to people who agree to receive them.

SMS (short message services) An inexpensive and convenient short text message sent to any digital receiving device.

Spam Unsolicted and unwanted incoming e-mail that shift the content filtering process from sender to receiver.

Viral marketing The electronic version of word of mouth involving e-mail messages specifically designed for recipients to forward to friends, family, and others on their e-mail lists.

CROSS REFERENCES

See *Data Mining in E-Commerce; Electronic Commerce and Electronic Business; Intelligent Agents; Online Communities; Personalization and Customization Technologies; Web Site Design; Wireless Marketing.*

REFERENCES

A classic scenario: Bad news, good news (2002, June 18). *eMarketer.* Retrieved July 5, 2002, from http://www.emarketers.com

Bailly, J. (2001). Victoria's Secret Website doubles traffic. Retrieved April 30, 2002, from www.fashionwindows.com

Carr, L. (2000, October 30). Events move millions to the net. *The Industry Standard,* 190–191.

DoubleClick Inc. (2001). E-mail marketing proves effective. Retrieved January 30, 2002, from http://www.doubleclick.com

eConsumers prefer e-mail newsletters and coupons (2001). *eStatNews.* Retrieved December 15, 2001, from http://www.emarketer.com

E-mail and the different levels of ROI (n.d.). Retrieved November 17, 2001.

Encyclopedia of the new economy (2002). *Hotwired.* Retrieved February 20, 2002, http://hotwired.lycos.com

eMarketer designates Interactive Advertising Bureau/PricewaterhouseCoopers quarterly Internet ad revenue report as its benchmark source for online ad revenues (2002). Retrieved April 25, 2002, from http://www.iab.net

Hallerman, D. (2002). Online ad pricing: Count heads or count results. Retrieved April 25, 2002, from http://www.emarketer.com

Helperin, J. (2001, May 1). Spinning out of control? *Business 2.0,* 54–57.

Internet is powerful complement to traditional advertising media (2002). Retrieved December 20, 2002, from http://www.advantage.msn.com

Internet users will top 1 billion in 2005 (2002). *Computer Industry Almanac.* Retrieved February 20, 2002, from http://www.c-i-a.com

Leading industry advertisers (2002). *NetRatings Report.* Retrieved April 20, 2002, from http://adrelevance.com

Mardesich, J. (2001). Too much of a good thing. *The Industry Standard.* Retrieved May 15, 2002, from http://thestandard.net

McAllister, M. (2001). Online sweepstakes are gaining popularity. Retrieved April 21, 2002, from http://www.digitrends.com

No surprise: Total ad spending down 9.8% for 2001, compared to results in 2000 (2002). Retrieved April 24, 2002, from http://www.cmr.com/news/2002/030602.html

Orbitz (2001). Press release. Retrieved April 26, 2002, from http://212.133.71.16/hsmai/news

Jupiter Media Metrix (2001, December 10). As cited in presentation by Lisa Morita, Senior VP and GM, Online Business, Overture, at Internet World.

Pastore, M. (2001). Banners can brand, honestly they can, part II. Retrieved December 20, 2002, from http://www.cyberatlas.internet.com

Pastore, M. (2002). Incentives still key to mobile advertising. Retrieved December 20, 2002, from www.cyberatlas.internet.com

PricewaterhouseCoopers (2002). *IAB Internet advertising revenue report: Third quarter 2001 results*. New York: PricewaterhouseCoopers New Media Group. Retrieved April 10, 2002, from http://www.iab.net

Rosen, L. (2001). Web ads that work—really. *Fortune Small Business, 11,* 95–96.

Saunders, C. (2002). CMR: Web ad spending dropped 14.7%. *Internet News*. Retrieved April 10, 2002, from www.internetnews.com

Silk, J. (2001, December 13). SMS marketing finds its voice. *Internet World Show Daily, 7,* 32.

Strauss, J., & Frost, R. D. (2002). *E-marketing* (2nd ed.). Upper Saddle NJ: Prentice-Hall.

The value of a corporate e-mail address (2001). *eContacts Media Kit*. Retreived June 4, 2002, from http:// www.econtacts.com

Value sites win, customer service losses (2002). *Cyberatlas*. Retrieved April 10, 2002, from http://www.internet.com

Weintraub, A. (2000, October 16). When e-mail ads aren't spam. *Business Week,* pp. 112, 114.

Williamson, D. A. (1999, June 28). P & G's reformulated Pert Plus builds consumer relationships. *Advertising Age,* 52.

Marketing Plans for E-commerce Projects

Malu Roldan, *San Jose State University*

INTRODUCTION

With the dot-com meltdown in the early 2000s came a renewed emphasis on business fundamentals. No longer was it conceivable to ride the Internet frenzy hype machine and obtain funding on a bright idea sketched out on a dinner napkin. The hunt for profits, as opposed to revenues, underscores the need for extensive analysis, better planning, and solid credentials to buttress the case for the viability of a given e-commerce implementation. The marketing plan is essential to building this case. This chapter provides a primer on the development of a plan for marketing an e-commerce product or service. It provides the rationale behind the marketing plan and provides practical guidelines for producing a document that can serve as a roadmap for bringing your product or service to market and beyond. The guidelines provided here are pertinent to either a business-to-consumer (B2C) or business-to-business (B2B) endeavor, and throughout, examples of applications for either type of project are described.

A marketing plan encapsulates your current definition of your product or service, an assessment of the competitive nature of your industry and of the market segments that you are targeting, a plan for promoting your product or service to these market segments, a mapping of distribution channels for delivering your product or service, and methods for evaluating how well you are doing on your plans. Since the marketing plan is a document with a straightforward structure, there are many software packages available to help you organize the document. However, as is typical with most plans, the structure of the document is the least problematic part of the endeavor. Writing the plan is only a small piece of the entire process. It is the background research and creative thinking that is essential to building a credible and useful plan. In the final section of this chapter, we will discuss some of the software packages available for supporting your marketing plan development. Most of this chapter will provide guidelines for the background work that will take up most of your effort in developing your plan.

KNOW YOUR PRODUCT OR SERVICE

At the very start of your market plan development, it is useful to articulate a working definition of your product or service and some key attributes. Although it is a mistake to aim for a rigid definition of your product/service at this point, it is still quite useful to build a definition to help focus your initial work on a marketing plan. This provides a baseline from which future versions of your product/service and plan will evolve. It will also help to use this initial definition as a basis for building a one- or two-sentence description of your product/service that has some depth and specificity. This can help you maximize your crucial first few minutes of conversation with potential partners, customers, and investors.

Your working definition should include a general description of the product/service that answers the following questions:

What does the product/service look like? What are its key components?

What problem or need does the product/service address?

What are the unique attributes of the product/service that differentiate it from similar ones available on the market?

What are the benefits that this product/service provides to consumers and/or business customers?

Prepare a detailed description of your product/service based on your answers to these five questions. Then capture the essence of the description in a one- or two-sentence definition that you can use to introduce your product/service to interested parties. Lastly, start working on a phrase that can serve as the slogan for your product. Although this slogan is likely to evolve as you refine the definition of your product or service, developing it early on helps to infuse your efforts with attention to branding and promotions.

Once you have an initial definition of your product/service, it would be useful to consider enhancing it by incorporating some capabilities generally associated with e-commerce endeavors. Two capabilities that can impact marketing activities in particular are bricks-and-clicks approaches and embedded intelligence.

Bricks-and-Clicks Approaches

Any e-commerce endeavor would be well served by a consideration of the level to which it will integrate its online and offline operations. E-commerce initiatives have been shown to benefit from such integration, particularly when the organizations involved have been cognizant of the pitfalls and opportunities it presents. Gulati and Garino (2000) identify four areas to consider for integration—brand, management, operations, and equity. On a limited basis, integration can involve promoting an online product/service using bricks-and-mortar channels (e.g., billboards, in-store promotions) or letting customers pick up or return products ordered online at a bricks-and-mortar facility. On a more extensive basis, integration may involve a shared brand, management, and information systems. Office Depot is an example of a company that took on a highly integrated approach, primarily because, thanks to its catalog business, it had the luxury of already having the infrastructure to support extensive product ordering and order fulfillment. For its B2C business, Office Depot made its catalogs available online, benefiting from the ability to update product information and pricing dynamically, and allowing consumers the flexibility of ordering and obtaining delivery of their products in the manner most convenient to them. For its B2B channels, Office Depot provides its largest customers with customized applications that allow employees at these customer companies to order products directly, within the parameters of preset authorization levels (Gulati & Garino, 2000). Other companies have achieved similar arrangements through partnerships with online category managers (Sechler, 2002) and next generation online B2B exchanges (Parmar, 2002; Thomas, 2002).

Furthermore, mobile computing technologies are making it possible to allow customers to travel around in bricks-and-mortar spaces while maintaining connections to online resources via wireless connections. As the infrastructure to support mobile computing matures, it will enable new ways of designing and marketing products/services. For example, products that are difficult to describe using online technologies (e.g., how clothes fit, the texture of fabrics for a sofa, how a delivery truck drives) can be tested by customers at a bricks-and-mortar facility such as a store or dealership. With the growing availability of broadband wireless connections for handheld devices (through various versions of the 802.11 standard), information downloaded in real time to handheld devices can augment the rich information gained through this actual contact with the product. Once the customer has made a decision about the product, a customized order can be made using the handheld device.

Services that can take advantage of the location of a customer can also benefit from the application of mobile technologies. A common example of this type of service is sending to a customer's handheld device recommendations regarding products and services available close to his/her location. Marketing efforts also benefit from location-based mobile services. For example, B2C retailers can cross-sell products to consumers across an entire shopping area such as a mall. As consumers walk through the mall, allowing their purchases to be logged on the handheld device, they can receive messages promoting special prices and deals on products and services that complement the items that they have just purchased. Another possibility would be consumers logging on to services that provide them with comparison pricing for an item they are planning to buy. As a marketing ploy, retailers may choose to send them incentives for driving a few extra miles to purchase the product at another store. B2B companies can apply the same principle to send to their customers' handheld devices, information on product upgrades, sales promotions, and maintenance alerts as they come in close proximity to currently installed products. Alternatively, the same information may be beamed to a sales person's device as he/she makes a routine visit to a customer's facility.

To develop a bricks-and-clicks approach for your product/service, consider a trip to a bricks-and-mortar facility that provides a similar product/service. Observe the processes used to deliver the product/service. This will give you preliminary insight into the expectations that customers have when purchasing these items or services. Consider how you might enhance these processes by introducing portable information technologies. Also, think about how you might allocate some aspects of the process to an online application, a bricks-and-mortar facility, or both in combination.

Embedded Intelligence

As microprocessors have gained greater power and miniaturization, we have seen more of them becoming a part of the objects of our daily lives. Everything from cars to printer cartridges to clothing has been enhanced by embedded microchips that provide monitoring information and customizable features. When integrated, embedded microchips increase the information component of every product and every device used in the process of providing a service. As such it opens up the possibility of increasing the flexibility of a product/service to a level that it can be customized practically "on the fly." These "smart products" (Glazer, 1999) allow companies to postpone the final configuration of a product, enabling greater customization and alleviating issues brought on by uncertainties regarding the preferences of customers (Lee, 2002). For example, a microchip makes it possible to remember the seat settings for different drivers of an automobile. These can be automatically set based on the weight of the driver as he/she slides behind the wheel. High-end microchips such as those manufactured by Xilinx, often installed in B2B applications, feature reconfigurable logic that can be modified via the Internet or other networks, even after installation (Lee, 2002).

The ability to tweak a product or service on the fly, if necessary, is a great enabler of one-to-one marketing. One-to-one marketing (Winer, 2001) is most easily envisioned with a pure information product, such as a Web-based information service. The information that is provided to a given customer can be customized to a level that makes it unique to each customer, based on past behavior or settings that the customer opted into at an earlier date. This is made possible through the use of personalization engines and/or plain old databases and scripting languages. When a product and/or bricks-and-mortar service is involved, one-to-one marketing can be achieved by integrating intelligence into the product or service, using embedded microchips, much like the driver's seat on a car is customized. Additionally, the monitoring capabilities enabled by embedded chips make it possible to collect information that can be used to determine the appropriate time to send reminders ("time to change the battery") or marketing messages—assuming that the proper permissions have been granted by the customer. These monitoring capabilities assume that the devices in which the chips are embedded also have the capability to patch onto a network and connect with a company server, either continuously or occasionally, via a cradle, infrared, 802.11x, or similar connection to a data communication line (with or without a desktop computer).

To determine whether your product/service might benefit from embedded intelligence, first think about how it might be customized to meet customer preferences. Would it be possible to design the product/service in modules so that some of its components can be modified or replaced with other components to create a customized product or service? Would it be possible to provide some information related to a product's operation or capabilities via a small display? Finally, could customers find benefit in remote monitoring of the product/service?

IDENTIFY THE COMPANY'S GOALS

Establishing the goals of the organization upfront will help place boundaries around the range of approaches considered in the marketing plan. The set of approaches can vary widely for different goals. A goal couched in terms of sales and revenue will result in very different strategies from those resulting from a goal that emphasizes profits. The former will emphasize aggressive sales tactics and pricing strategies that undercut the competition (oftentimes selling goods below cost). The latter requires prudent inventory control and pricing, and fiscal responsibility.

Goals can take the form of financial results such as revenue targets, profit margin targets, market share, growth in revenue, profits, and market share, or cost reduction. Alternatively, you may choose to take a benchmarking approach. That is, matching the investments that the rest of the companies in the industry are placing on various strategies, technologies, marketing approaches, and research and development.

Once your goals have been established and agreed upon by your core development team, develop measures that you will use to track how well you are meeting these goals. Put together an evaluation schedule that identifies measurable achievements over time. An example is setting target sales figures to be compared against actual sales every month. Ideally, this would be part of a larger set of measures that your company will use to track its progress—not only with its marketing plan but also with establishing the entire business. Beyond tracking stock prices and sales, it is useful to include more traditional measures such as key expense items, profit margins, and cost of goods sold. The advice of a comptroller can go a long way toward lowering the risk of long-term misallocation of investor funds. Fiscal responsibility, when applied prudently, can facilitate effective use of your company's capital without stifling the creativity and enthusiasm of your core team. In the wake of the dot-com meltdown, this balanced approach is gaining greater cachet.

MARKET RESEARCH

While a good product/service definition forms the cornerstone of a strong market plan, solid market research is the foundation on which the market plan builds its credibility and depth. The extent and sophistication of the research that goes into a given market plan will have a strong impact on how the product/service and the company is viewed by potential partners and investors. The research should aim to define and describe both the target market for the product/service and the competitive climate in the company's industry. The first two parts of this section provide a discussion of various types of information that could be gathered and models used to build descriptions of the target market and the industry's competitive climate. The third part will discuss methods that may be used to gather the data to build these descriptions.

Target Market

Research into the target market should result in a rationale for the selection of the market segment or segments to be targeted, the size of that market in dollars and/or number of customers, and guidelines for how to design and market the product to the targeted segment(s). Information and data pertinent to this analysis include demographics, psychographics, technology life cycle, preferred marketing channels, and the context in which potential customers experience or purchase the product/service.

Demographic data are key to determining the size of a B2B or B2C target market as well as identifying the best way to reach the target market. To start off, determine the profiles of the types of customers most likely to buy your product/service. Typical demographic categories to include in these profiles are age, gender, household income, geographic location, ethnicity, and level of education. Demographic categories can be used to find data to estimate the size of the market. Published sources such as those listed in Table 1 typically provide findings on the sizes of various markets based on these demographic categories. Both offline and online target market estimates are available for most of these categories. Oftentimes, building an estimate of your target market will take a combination of the estimates for several demographic categories. For example, you may have to infer the size of the online population for individuals with a given level of education by first taking the total number of users online and multiplying that by the percentage of the total population that has achieved the target level of education.

Table 1 Sampling of Online Resources for Market Research

Name/URL	Description
Business.com http://www.business.com	Business-oriented directory and search engine
CyberAltlas http://cyberatlas.internet.com	News and statistics on e-commerce and Internet trends
LexisNexis http://www.lexisnexis.com/default.asp	Leading full-text database service for legal, medical, and business records and news
Ibiblio http://www.ibiblio.org/	Online public library
Neilsen/Netratings http://www.nielsen-netratings.com/	Free summaries of top-level e-commerce and Internet statistics. Online market research services.
Nua http://www.nua.com/surveys/	Internet trends and statistics
StatMarket http://www.statmarket.com	Global Internet user trends
emarketer http://www.emarketer.com/	E-business statistics and information
Securities Exchange Commission http://www.sec.gov/	Filings of public companies
Bureau of Labor Statistics http://www.bls.gov/	Business-related U.S. government statistics and studies
U.S. Census Bureau http://www.census.gov/	U.S. census data

Psychographics are a useful method for developing greater understanding of consumer markets. Psychographics aim to characterize the psychology of target customers. Consumers are often categorized according to their values, motivations, beliefs, attitudes, and lifestyles (Peltier, Schibrowsky, Schultz, & Davis, 2002). Psychographics have had somewhat controversial application due to the difficulty of measuring such categories using traditional methods such as focus groups and survey research. Recently there has been some renewed interest in using psychographics to inform the application of data-mining techniques in analyses of large consumer databases.

The life-cycle stage of the core technology of your product or service is important to enriching your understanding of the customers that you are targeting. Moore (1999) proposed a set of strategies to match the life-cycle stage of a given technology—ranging from a niche-focused strategy for technologies just beginning to enter the mainstream, to one focused on building market share if a company's product happens to be considered—or is vying to be—the de facto standard in the industry. Determining the life-cycle stage of your technology will help you decide whether your customers will be mostly made up of early adopters like technology experts and visionaries or mid- to late adopters like pragmatists and conservatives. The early adopters will be more likely to forgive technical difficulties and appreciate your product/service for its leading edge appeal. Mid- to late adopters will demand reliable products and a close, proven fit with their business needs. Build a realistic assessment of where your product lies in the life cycle, both in terms of reliability and proven practical value. This will be invaluable in developing a marketing plan that provides the right message to your target market. For example, early adopters are likely to appreciate marketing messages that emphasize the product's technical features while mid- to late adopters are more likely to respond to marketing messages that stress the product's business applications and proven track record. It will also help you determine how to move from one life-cycle stage to another. Ultimately, success in the mainstream market is the most direct means of garnering superior profits from your endeavor.

Determining which marketing channels that members of your target market tune into can be an important resource saver for your company. Market segments can display variations in their profiles of media usage, as has been shown for different ethnic groups (Greenspan, 2002a). It is, therefore, most efficient to focus marketing messages on the channels viewed most frequently by your target segment. Furthermore, if your company decides to take a bricks-and-clicks approach to marketing your product/service, one possibility is that you would aim to move individuals from an offline to an online channel. Such approaches call for marketing campaigns that publicize the online site via traditional offline media channels like billboards and television. A common example of this is the inclusion of a company's Web site URL in all offline marketing materials.

Lastly, determine where, when, and how your target customers will access and purchase your product/service. Understanding this context fully will be important for easing adoption of your product/service by your target market segment(s). It is also pertinent to building appropriate bricks-and-clicks products/services and marketing strategies. Intuit provides an excellent case to support this point. The company spent many hours observing its target customers balancing their checkbooks and handling their finances at home on the kitchen table or in the den. Using this information, Intuit was able to build a software package, Quicken, which has displayed amazing endurance—even in the face of challenges from Microsoft! Although this endurance can be attributed to the high switching costs involved in moving from Quicken to another package for managing home finances, it is also the case that Intuit's attention to context resulted in a product with a highly intuitive interface that was easily integrated into the home finance activities of its customers. A methodology for achieving this close-fit, market validation, will be discussed in the last part of this section.

Competitors

Knowledge of your competitors is essential to building a clear statement of the unique benefits provided by your product/service, whether you are marketing to a consumer or business market. Building this knowledge involves an analysis of your industry, your competitors' products/services, and your competitors' marketing methods.

As with your product definition, it is important to delineate up front the specific industry that your expect your company to impact with the introduction of your product/service. Limiting the scope of your industry analysis in this manner will make it much easier to identify relevant competitors and markets. It would also be beneficial to consider the global reach of the industry you are analyzing. With the global reach of the Internet, most any e-commerce endeavor will have global exposure once it goes online. A useful way to structure this part of your research is with a SWOT analysis. This juxtaposes your company's strengths and weaknesses with the opportunities and threats found in your target industry. A SWOT analysis answers the following questions:

What are the strengths of our business and products/services that we can build upon?

What are the weaknesses of our business and products/services that we need to address?

What opportunities are available in our target industry?

What threats do we currently face or expect to arise in the future in our target industry?

A more extensive analysis of the target industry can be achieved by conducting a five-forces analysis (Porter, 1975). With this analysis you will identify the current rivals, potential entrants, suppliers, buyers, and substitute products for your target industry. After identifying these elements in the industry, further analysis is conducted to assess the level of rivalry in the industry, the relative ease or difficulty of entry into the industry, the power wielded by suppliers and buyers relative to target industry participants, and how well substitute products stack up against the target industry's products. Table 2 provides a sampling of questions that may be used to conduct the latter analysis.

Jap and Mohr (2002) underscore the importance of understanding the type of relationship a company has with its suppliers and distributors. The tenor of these existing relationships can influence how well a company can leverage the efficiencies arising from the application of e-commerce technologies to improve information sharing, extend a company's ability to reach new customers, and enable dynamic pricing. For example, they contend that partners who have long-term, complex, cooperative relationships may benefit greatly from efficiencies brought about by improved information sharing and ability to reach new customers. However, the application of dynamic pricing may undermine the long-standing relationships and trust built by the partners.

Table 2 Sampling of Questions to Be Used in a Five-Forces Analysis

Force	Sample Questions
Barriers to entry	Do current industry participants benefit from economies of scale? Do new entrants have to invest large financial resources to compete? Is it difficult to get access to distribution channels?
Power of suppliers	Are there few suppliers? Is the supplied product unique, differentiated, or associated with high switching costs? Could the supplier integrate forward into the industry's business?
Power of buyers	Does the buyer group purchase in large volumes? Are the industry's products standardized? Could the buyer group integrate backward and make the industry's products?
Threat of substitute products	Are there substitutes for the industry's products? Do the substitute products have lower prices than the industry's products? Do the substitute products perform better than the industry's products?
Industry rivalry	Is industry growth slow? Does the industry's product or service lack differentiation and have low switching costs? Are the industry rivals diverse in strategies, origins, and "personalities"?

An in-depth understanding of your competitors' products and marketing methods is an essential complement to the global understanding of the industry that one can gain from both the SWOT and five-forces analyses. Building this understanding involves something akin to informal anthropology. Although it will be possible to collect much information from secondary sources, oftentimes the best information can only be gathered firsthand by visiting both online and offline facilities of competitors, speaking with key management and employees if possible, analyzing the features of their products, following their pricing strategies, and viewing their online and offline marketing programs. Particularly when you are trying to enter a market composed of smaller players, there is often a lack of specificity in the information provided in published sources.

As you are gathering your data, organize your information in comparison tables, lining up elements of your products and marketing strategies with those of your competitors. An initial list of comparison points should include key product features, pricing, sales channels, marketing channels, promotions, management structure, description of operations, sales, and revenue numbers. Expect that this list will expand as you learn more about your competitors and their products/services. Review the tables periodically and highlight items where you might be able to build an advantage or where some work needs to be done to strengthen things on your end. Attempt to sum up the key strengths and weaknesses of each competitor, including your own company. Summarize the information further by capturing each competitor's "personality" in a phrase (e.g., low-cost producer, boutique service, dark horse).

Market Research Methods

A wide range of sources is available for your market research efforts. It is possible to get a lot done at little or no monetary cost, although much effort must be expended in sifting through the volume of information available that in general has only tangential significance to the research you are conducting. Often, a small investment in purchasing a well-targeted research report is a large time-saver, if you have the budget to do so. However, these secondary sources may still provide inadequate information, especially when you are building innovative, untested e-commerce products/services. In the latter situation, primary research methods such as surveys or market validation will be invaluable in eliciting information on the viability of your product/services directly from your potential customers. The following discussion will begin with an overview of secondary research sources then discuss primary research methods, particularly surveys and market validation.

Secondary research sources have become significantly richer and more accessible with the advent of the Web. Unfortunately, this has also led to the proliferation of unreliable data and hence greater difficulty with finding well-supported and relevant information. For many people, the first step in any search is the use of a search engine such as Google (http://www.google.com). While this is a convenient and generally fruitful approach, it is also hardly the most efficient way to find information, offline or online. Specialized databases, available online or via your local library, often provide results that are more focused and relevant to the research that you are conducting. Table 1 provides a sampling of some useful online resources. It is also worthwhile to ask for advice from a reference librarian in a large public library to find out whether you have legitimate access to specialized databases that cover your topics of interest. If you are conducting the research for a large company, you are likely to have subscriptions that give you access to publications of consulting firms with e-commerce practices. Otherwise, you can purchase the reports if your budget allows it or you can work with limited summaries of the reports available at the consulting firms' Web sites, via consultant and trade newsletters, or at sites that track e-commerce trends such as Cyberatlas. Other rich sources of information are Web sites of government agencies such as the Securities Exchange Commission, the Bureau of Labor Statistics, and the Census Bureau, and Web sites of industry trade groups. You can find results from a wide range of public opinion polls from the Roper Center's online poll collection found at http://www.ropercenter.uconn.edu. Lastly, consider visiting some complaint sites (e.g., http://www.untied.com for United.com) where disgruntled customers give vent to their frustrations. These sites may be a good source of ideas for how to improve your product/service—based on complaints about a similar offering from another company (Goldwag, 2002).

Secondary resources, as listed above, will often fall short of providing the material that you need to build a comprehensive understanding of your market. When this happens, it will be useful to conduct some primary research of your own to collect data of greater relevance. Two avenues for doing this are survey research and market validation. With online services, it is possible to conduct informal and formal surveys at a fraction of the time and expense that they used to require. Newsgroups on your subject of interest are great settings for doing informal surveys, for quickly testing an idea or getting quick clarification on some details of your research—as long as you refrain from communications that are promotional in nature. Formal surveys can be conducted, generally for a nominal fee, on Web services such as Zoomerang (http://www.zoomerang.com) and Netreflector (http://www.netreflector.com/). Other sites like eFocusGroups (http://www.e-focusgroups.com/) will allow you to ask questions and test your ideas in front of virtual focus groups.

Market validation is a product development and market research methodology developed by Frank Robinson over decades of consulting with firms bringing new products to market (http://www.ProductDevelopment.com). It is an intensive, structured research strategy designed to result in the rapid evolution of a product/service strongly tied to customer preferences. At the same time the process helps the core project team build a solid understanding of its target markets and appropriate market strategies. Market validation immerses your core project team in several waves of meetings where prototypes of your product/service are presented to potential customers in the context in which they will be using the product/service (e.g., at home on the kitchen table, in an office cubicle, at the company warehouse). Managers, analysts, and programmers with the primary responsibility for developing the product/service come to the customer's site as a team to conduct a demonstration, observe the customer's use of the prototype, and interview the customer about his/her experiences and preferences. Findings from each visit are documented and analyzed immediately—beginning in the vehicle used to travel to and from customer meetings—using structured templates designed to extract key information for improving the product and marketing strategy. Before the next wave of customer meetings are conducted, the product/service and strategy are tweaked according to customer feedback and the results of these discussions. The team goes through several waves of these customer meetings until it has reached saturation, finding little new information, and getting consistently high marks from the customers interviewed. At this point the findings are summarized into a marketing plan and the first production version of the product is developed.

MARKETING APPROACHES

The 1990s Internet boom fueled a proliferation of marketing trends as startups and established companies alike sought to build strategies that fully exploited e-commerce technologies. A small cottage industry grew up around the retooling and repackaging of traditional marketing approaches to take into account the opportunities and pitfalls afforded by the new technologies and the momentum they created. This section will present a sampling of these marketing approaches with the intention of providing readers with a basic toolkit of approaches to use in marketing their products/services. Hopefully, they will help spark some creative new approaches as well. The marketing approaches presented in this section are guerilla marketing, brand management, viral marketing, and permission marketing.

Guerilla Marketing

As its name implies, guerilla marketing brings together a set of aggressive, grassroots tactics for marketing a product or service. Generally, the guerilla marketing tactics assume that the marketing effort is being conducted on a shoestring budget. Therefore, the tactics will often take advantage of the free to almost free resources available on the Web and suggest ways of using any type of public forum as a setting for promoting products/services. Guerilla marketing has engendered an online community (http://www.gmarketing.com) that shares simple, low-cost, yet effective promotional strategies. Examples of such strategies are printing coupons on post-it notes and placing them on the front doors of potential customers, or intelligent participation in discussion groups.

Brand Management

About halfway through the Internet boom, many e-commerce startups realized the importance of building a brand and started filling traditional channels like radio and TV with advertisements designed to help customers build an emotional bond with their brands. Although many of these brands did not last long enough to sustain such bonds, brand management remains an important aspect of any marketing strategy. An established brand enhances the profitability of a company when customers' emotional ties to the brand keep them coming back to purchase the company's products or services. Furthermore, a strong brand also affords a company a measure of power over customers' choices among offerings from a variety of vendors. A strong brand enables a company to not only build products to match customer preferences but also the power to influence those preferences. Once a brand has positive characteristics associated with it, customers look to it as a signal that the products sporting the brand also possess these positive characteristics. For example, Gap, Inc.'s strong brand makes it possible for them to have strong influence over young consumers' fashion choices (Kirsner, 1999).

Every occasion for communication from a company is an opportunity to strengthen and promote the brand. A key part of brand management is making sure that a consistent brand is projected throughout all company materials and communications—from brochures, to advertisements, and the company Web site. A well-run e-mail communications program has also been found to positively impact branding (Saunders, 2002b). For Gap, Inc., entry into the e-commerce arena provided the opportunity to buttress the consistency of its brand globally. The company Web site mimicked the casual, stylish atmosphere of its stores. Pictures of Gap billboards around the world were featured on the Web site, promoting the global reach and consistency of the company's brand and image.

Viral Marketing

Viral marketing gained notoriety with the surprising and phenomenal success of Hotmail. Viral marketing aims to have a company's customers act as its marketers. As they use the company's products and services, customers simultaneously spread the word about the company's offerings. The marketing message then spreads like a "virus" throughout each customer's social network, bringing news of the product and service to a wide range of potential customers. For example, Hotmail was able to sign up 12 million users in 18 months by inserting the tagline "Get your free e-mail at Hotmail" at the bottom of every e-mail sent out by its customers (Jurvetson, 1998). Recipients of the e-mails were able to click through to the Hotmail Web site to set up their own e-mail accounts.

Permission Marketing

As e-commerce adoption grows, both business customers and consumers are getting more and more uncomfortable about the volume of e-mail they receive and possible threats to their privacy. The more information customers share with e-commerce companies, the greater their exposure to being victims of identity theft, fraudulent credit card activity, and rampant spamming by companies and persons seeking to market products and services to them. As technologies for ensuring privacy become easier to use, it is possible to envision a time when customers will have the ability to control access to their private information and preferences via a seamless interface. Until then, companies will be using the techniques of permission marketing to make sure that customers get messages that they are really interested in. Customers must opt-in to receive any communications about a company's products/services. That is, before a company can send materials to a customer, the customer will have to actually choose to receive the materials. This is generally done by having customers fill out a form to select their areas of interest and submit e-mail addresses to indicate their interest in receiving periodic e-mails closely matched to their interests. This is known as single opt-in. Unfortunately, single opt-in creates problems when a customer enters an e-mail address inadvertently or if someone submits an e-mail address without approval from its owner. Double opt-in (also known as safe opt-in) addresses this issue. With double opt-in, the company sends a letter to the submitted e-mail address requesting that the customer confirm interest in joining the mailing list. Only upon confirmation by the customer will the e-mail address be added to the list (Brownlow, 2001). Other permission marketing approaches provide additional incentives such as the opportunity to win attractive prizes in an online game (Taylor, 1998).

MARKETING CHANNELS

There is no shortage of channels available to get your message across to your potential customers. The difficulty lies in finding the one that best reaches the customers who will pay attention to the message and then purchase your product/service. As above, knowledge of your customers is key to selecting the best marketing channels and getting the most value for your marketing dollars. When conducting research on your customers, it would be helpful to gather as much information as possible about the marketing channels they pay attention to, what types of messages they respond to, what types of messages turn them off, and what would entice them to make a purchase. It is important to also try and determine the intentions behind customers' purchases. Are your target customers liable to form a relationship with your brand or product or do other intentions motivate their purchases—e.g., convenience, low price (Dowling, 2002)? Knowing your customers' intentions will help you place your promotions in the channels that maximize your chances at generating sales. This section provides a sampling of the marketing channels available for an e-commerce endeavor, be it B2C or B2B, on three platforms—Web based, traditional, and mobile computing.

Web-Based Marketing

The Web provides one of the least expensive ways of promoting products. Unfortunately, techniques for promoting products over the Web are in their infancy and, thus, do not have the same track record of success as other, more traditional channels. Still, the low cost and relative ease of implementation of Web promotions, and the more that 389 million online users worldwide (Neilsen/Netratings, 2002), recommend the Web as a channel to consider, at the very least in tandem with more traditional channels. Furthermore, Web-based channels have the potential to allow more exact targeting of the groups who might receive your marketing messages. Providers of Web-based promotions are getting more sophisticated in their ability to identify groups of individuals to send messages to, based on past online browsing and purchasing behavior of users. With more opt-in and personalization capabilities provided by Web sites, it is possible to target messages customized to the level of each individual receiving the messages.

The most common platforms for Web-based promotions are the company Web site, banners and pop-ups, search engine placements, and e-mail promotions. The company Web site is a key part of building a company's digital image. It is the place where customers will come to after they are enticed by banners, pop-ups, and other methods of promotions. Aside from achieving a look and feel consistent with your company's brand, it is imperative that the Web site be thoughtfully structured to ease navigation through its pages. Banners have come a long way from their lowly roots as easy to ignore minor annoyances on Web pages. Thanks to rich media technologies such as Flash and DHTML and innovative formats and placements, one can now find banners with attractive animation and messages that are difficult to ignore (Rewick, 2001). Similarly, pop-ups have started to dominate user desktops, much to everyone's consternation. Top search engines sell placements on their results pages so that company links appear on prominent areas of results pages. It is possible to purchase many of these placements centrally via services such as http://www.registereverywhere.com or http://www.GetTraffic.com. Lastly, e-mail promotions are some of the most effective and best value promotional methods afforded by the Web (Saunders, 2002a). With the advent of HTML-mail, companies have been able to create attractive advertising messages to be sent directly to inboxes of potential customers who have, ideally, opted-in to receive them. Alternatively, companies can purchase e-mail lists of individuals identified as potentially interested in receiving such messages. Companies such as Double-Click (http:///www.doubleclick.com) provide comprehensive, hosted, online advertising services that organizations can use to build campaigns that span all or several of the online channels listed here.

There is evidence that online advertising is gaining some respect among both B2C and B2B advertisers. Research by Double-Click found that rich media ads (almost a fourth of all ads served up in 3Q 2002) had click-through rates (CTR) of 2.7%. Traditional banner ads continued their decline with only a 0.27% CTR (Greenspan, 2002b). Seventy-eight percent of 1000 consumers surveyed by Double-Click stated that they made a purchase after clicking on a direct marketing e-mail message. The trick, however, is to get a consumer to open the e-mail despite the continued overfilling of inboxes and growing use of bulk mail folders. Direct mail marketers are becoming more savvy about placing attractive brands names on the "From" line of their e-mails. A combination of viral and direct marketing techniques is also used to address this issue as marketers are including "mail to a friend" buttons in their marketing pitches so that the e-mail gets sent with a familiar name on the "From" line (Saunders, 2002a).

Until recently B2B marketers were reluctant to use the online platform for advertising, accounting for only 11% of online ad dollars, according to a GartnerG2 study. However, the same study predicts that this figure could rise to 22% by 2005 as B2B marketers warm up to the possibilities brought on by rich media and improvements in ad targeting and accounting tools (Saunders, 2002c). Additionally, studies such as one done by Nielsen/NetRatings have shown that more than half of business decision makers say that the Web is the best way to reach them. Due to improvements in targeting, B2B marketers are looking to attract potential business buyers with rich media banners in general Web sites like http://www.ESPN.com and http://www.Weather.com. In keeping with business buyers' high information requirements when making purchases, once buyers click onto a seller's site, they are often provided information-rich materials such as white papers, editorial pages culling together information on a given vendor's products, and webcasts of marketing seminars or consultant/researcher analyses (Bialik, 2002).

In all Web-based promotions, companies must be cognizant of the risks of negative emotions being attached to their messages because, as of now, many of these ad placements are construed as intrusive. There is a growing industry around the elimination of pop-ups, e-mail spam, and banners, as online users have become inundated with unsolicited, poorly targeted, promotional messages. Eventually, we may get to a point where only messages that individuals opt into will be allowed to impact their desktops. This underscores the importance of considering permission-marketing methods as part of your strategy.

Traditional Methods of Promotion

As in the design of your product/service, taking a bricks-and-clicks approach to promotions can bring benefits. At the very least, one can use traditional media such as billboards, print media, television, and radio to broadcast the availability of your product/service and promote or create a brand. The techniques for using such media for brand promotion are well established and have proven to be successful for many companies. The audience for such media also dwarfs that available online. These traditional media can point viewers or listeners to the company Web site where more extensive and detailed materials and an online storefront are provided.

A bricks-and-clicks approach need not be limited to the use of traditional media. Any bricks-and-mortar setting can be used to complement Web promotions. The most common one used by companies with bricks-and-mortar stores, such as Barnes and Noble, is to provide incentives for individuals to provide e-mail addresses to be used by the company to send e-mail promotions. Rite-Aid promotes its partnership with http://www.drugstore.com by printing the Web site's address on its shopping bags and prescription labels. Recent studies indicate that Internet users often surf the Web and watch TV simultaneously. A significant finding is that for a majority of these individuals, the activities on both channels are unrelated. This presents an opportunity for marketers to introduce promotions that integrate the capabilities of both platforms (Cyberatlas Staff, 2002).

Aside from in-store tie-ins, companies can consider other, less common offline methods. Some companies have used company cars as moving promotions by covering them with the company logo—much the way Southwest Airlines paints its airplanes to resemble some of the attractions (e.g., killer sharks) in its destination cities. Building communities offline to complement those built online would be another approach to bricks and clicks promotion. Companies may consider sponsoring gatherings of customers to share their interests and knowledge of the company's products. Local user groups can be effective ways of getting the word out about a product/service and its updates, help new users learn about the product/service, and collect additional information to improve the product. They can also be a setting for selling the product or service, effectively, the e-commerce equivalent of the "Tupperware party" where customers get together to learn how to use a product and are able to purchase the product at the same time.

Promotions on the Mobile Computing Platform

An emerging platform for promotions is mobile communications—involving portable devices such as cell phones and personal digital assistants (PDAs) that are wirelessly connected via the cellular phone network, via some version of the 802.11 networks, or via a short-range networking technology such as Bluetooth or infrared. With the advent of widespread broadband wireless networks, such as T-mobile's Hotspots, advertisers have the capability to send rich data and build interactivity into their mobile computing promotions. On a mobile computing platform, advertisers can send messages to customers that are not only customized according to their needs but also targeted to be appropriate to their positions in time and space. Marketing messages can be targeted to arrive at users' devices as they are positioned close to a bricks-and-mortar store where they can make purchases. The message can also be timed to arrive when a purchase is most likely to be made. For example, an advertisement for a nearby restaurant and a listing of their prix fixe menu could arrive at a business executive's device at around noon. At a supermarket, coupons for soft drink discounts may arrive at users' devices as they are walking along the soft drinks aisle.

These examples assume that the mobile computing infrastructure has the ability to identify the location of the customer. This can be achieved in one of two ways. First, the customer could enter his/her location on the device—the zip code of the area he/she is in, or he/she may click on a section of a store map provided on the device, or enter a code for the section of the store. Second, the location of the mobile device itself may be tracked by a central service, such as the customer's telecommunications provider. The infrastructure for this second method is still being developed, most likely using either of two platforms or a combination of the two—the global positioning system (GPS) and triangulation based on the position of the device relative to the cells in the local cellular network.

An initial application of this is part of the mMode service from AT&T Wireless. The service allows subscribers to locate friends and get directions to nearby restaurants and other services. The service called Find Friends determines a user's location based on the most recent cellular tower that his/her phone contacted (AT&T Wireless, 2002).

Currently, cellular phones are—somewhat controversially—being used to broadcast advertising messages in text. Generally, these messages use the short messaging service (SMS), a text messaging service that is very popular in Asia and Europe and just gaining ground in the United States, where the traditional cellular phone standards are quite different and diverse (Cellular News, 2003). When traveling around the world with a phone that can patch onto all flavors of the most widespread cellular phone standard, Global Standard for Mobile Communications (GSM), one is usually greeted with a host of SMS marketing messages from the local telecommunications provider. As with pop-ups and e-mail advertising, there is the risk that the customer will find these messages intrusive. Thus, any company intending to use this platform must consider this risk and be cognizant of any developments in the regulation of this channel.

As cellular phones with large color screens, PDAs, and all manner of handhelds and wearable computing devices enter the mobile computing domain; this marketing channel will gain the ability to provide rich and attractive marketing messages. As of this writing, PDA users receive Web page ads formatted to fit the small screen when they synchronize their devices to information services such as AvantGo. The service then queries the customer to see whether he/she would be interested in opting-in to receive more information messages via e-mail. As the mobile computing platform matures, these messages can provide not only promotional messages but also value-added services appropriate to a customer's location. As described in the bricks-and-clicks section, these services have the potential for transforming the shopping experience and leveraging the strengths of both the online and offline retail settings.

PUTTING IT ALL TOGETHER—WRITING THE MARKETING PLAN

With all the research, brainstorming, and experimentation you have done to learn about your customers and develop a viable product and strategy, writing the marketing plan document itself can be a daunting task. Ideally, you will be working on versions and parts of it as you conduct your research—building comparison tables, drawing figures, and writing short notes that allow you to capture and organize your thoughts. Fortunately, there are many software products available on the market today to help with this organizing and document preparation task. Table 3 lists some of these products and their key features.

Most of the packages provide versions of the traditional marketing plan outline as templates that you can fill in with your ideas as you go along. You can view marketing plan samples to get some ideas on format, language, and content. Packages provide extensive lists of marketing research resources that you can use for your research. Some of the packages also help with the market research and analyses, providing sets of questions, templates, and spreadsheets to help you develop your marketing plan, schedule, and budget. Lastly, the packages support collaboration among your core team by providing version control capabilities and even Web-based editing—so that everyone on the team can work on the document wherever

Table 3 Sampling of Market Planning Software Packages

Package Name, Vendor, Web Site URL, and Price	Plan and Analysis Templates	Links to Market Research Sources	Sample Plans
MplanPro Palo Alto Software http://www.mplans.com/ Price: $99.95	Yes	Yes	Yes
Business Plan Writer Deluxe Nova Development http://www.novadevelopment.com/mainus/products/bqw/index.htm Price: $99.95	Yes	Yes	Yes
Microsoft Business Planner Microsoft Corporation http://www.microsoft.com Available as part of MS Office Small Business and Professional Editions	Yes	Yes	Yes
Smart Online Smart Online, Inc. http://www.smartonline.com/servlets/syngen/navigation/hometab_c/ Online Subscription: $24.95/month	Yes	Yes	Yes

he/she may be located. These packages and their capabilities are well worth the small investment they require not only because they help you to organize and think through your ideas but also because they will allow you to produce a document that has a structure familiar to potential partners and investors.

Once you have encapsulated your ideas into a well-researched manuscript, consider your marketing plan a living document. Every encounter with a potential investor, partner, or customer can provide additional data and ideas to hone your product/service and strategies. Use your evaluation criteria and methods to continually assess your progress toward your goal. Do not be hesitant to change course if your plans do not bring the results you expect. Your marketing plan will go a long way toward establishing your credibility with investors and partners. Going through the process of building the marketing plan ensures that you have studied your market thoroughly and your core development team has had enough occasion to balance its creativity with the realities of your target markets and industry.

GLOSSARY

B2B exchanges Online hubs that bring together suppliers and buyers for a given industry segment (e.g., Covisint for the auto industry) and that facilitate online information sharing and auctions of raw materials and/or maintenance, repair, and operating goods. Next-generation B2B exchanges have strengthened their offerings with services that facilitate negotiations, product design collaboration, and request for quote processes.

Bluetooth A wireless personal access network technology designed to connect devices within a local area spanning up to 100 m at speeds from 1 to 20 Mbit/s. Early applications include connecting an earpiece to a cellphone, and connecting printers to various computing devices.

Bricks and clicks An e-commerce strategy that aims to integrate activities from both the digital and bricks-and-mortar platforms. Integration can be achieved at various levels—sharing brands, operations, management, and/or equity—across traditional and online organizations.

Click-through rate (CTR) The percentage of visitors to a Web site that click on a banner ad displayed on the Web site to view more information on the advertised product or service.

Customer relationship management (CRM) A suite of practices and software toolkits that support all or some subset of the procedures associated with managing ongoing contact with customers—generally includes contact management, customer database, personalization, and recommendation engines.

E-mail marketing The use of e-mail technologies to conduct online promotions—generally to push advertisements to potential customers and to convert direct mail operations onto the digital platform.

Five-forces analysis A well-established model for analyzing an industry's competitive climate as a guide for building a strategy that puts an organization in a position of advantage in the industry. The analysis looks at the relative strength of five aspects of industries—power of buyers, power of suppliers, industry rivalry, substitute products, and barriers to entry.

Focus groups The primary data collection venue for a research method that asks a small group of individuals for their comments, concerns, and preferences regarding a given product or service.

Guerilla marketing A set of grassroots strategies for conducting product/service promotions at low cost. The set of strategies continues to grow via the contributions of an online community of practitioners, found at http://www.gmarketing.com.

Market validation An intensive process for developing products/services and marketing strategies closely tied to customer preferences; involves several waves of meetings where the core development team—including business and technical members—present versions of the product/service prototype to customers to get feedback-used as a basis for improving the product/service and market strategy. The team goes through several waves of these meetings until they have developed a product that consistently meets or exceeds customer expectations.

M-commerce The porting of e-commerce activities onto a mobile computing platform. Such platforms generally include a form of wireless networking linking all manner of portable devices—from laptops to personal digital assistants (PDAs), to cellular phones.

One-to-one marketing Marketing that aims to customize products and services to meet the preferences of every customer.

Online category manager A third-party distributor specializing in a product category—video, sporting goods, etc. These distributors enable retailers to broaden their online product mix while avoiding inventory and distribution costs.

Permission marketing A marketing technique that requests the approval of potential customers before promotional material is sent to them; designed to increase the attention paid by the customer to marketing messages, presumably because they have expressed an interest in receiving the messages. Customers may also be given incentives to opt-in to a promotional campaign.

Personalization engine A component of customer relationship management toolkits that allows a company to customize its marketing messages and, potentially, product/service attributes based on the preferences of its customers.

SWOT analysis A strategic analysis technique that involves an assessment of a company's internal strengths and weaknesses, and the opportunities and threats in an industry it is operating in or planning to enter.

Viral marketing A marketing technique that essentially has a company's customers acting as its promoters, whereby customers promote the company's products/services in the process of consuming or availing of them. Usually this involves a marketing message tacked onto some method of delivering the company's products. One of the most popular e-commerce successes for this technique is Hotmail.

CROSS REFERENCES

See *Business Plans for E-commerce Projects; Business-to-Business (B2B) Electronic Commerce; Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Click-and-Brick Electronic Commerce; Collaborative Commerce (C-commerce); Consumer Behavior; Consumer-Oriented Electronic Commerce; Customer Relationship Management on the Web; Electronic Commerce and Electronic Business; E-marketplaces; Marketing Communication Strategies; Mobile Commerce.*

REFERENCES

AT&T Wireless (2002). *mMode features—Find friends.* Retrieved January 13, 2003 from http://www.attws.com/mmode/features/findit/FindFriends/

Bialik, C. (2002, October 21). Sell first, advertise later. *Wall Street Journal,* p. R11.

Brownlow, M. (2001). Double opt-in. *ibizBasics.* Retrieved January 13, 2003, from http://www.ibizbasics.com/online030601.htm

Cellular News (2003, January 13). *GSM market doubles in size in the USA.* Retrieved January 13, 2003, from http://www.cellular-news.com/story/8118.shtml

CyberAtlas Staff (2002, September 17). Many surf Net and channels simultaneously. *CyberAtlas.* Retrieved January 12, 2003, from http://cyberatlas.internet.com/markets/advertising/article/0,,5941_1465021,00.html

Dowling, G. (2002). Customer relationship management: in B2C markets, often less is more. *California Management Review, 44,* 3.

Glazer, R. (1999). Winning in smart markets. *Journal of Interactive Marketing, 13,* 1.

Goldwag, W. (2002, September 5). Complaint sites can focus firms on issues to fix. *Marketing,* p. 16.

Greenspan, R. (2002a, October 23). Media mixing for the multicultural market. *CyberAtlas.* Retrieved October 25, 2002, from http://cyberatlas.internet.com/big_picture/demographics/article/0,,5901_1487151,00.html

Greenspan, R. (2002b, December 5). Rich media ads, CTRs up in 3Q. *CyberAtlas.* Retrieved January 12, 2003, from http://cyberatlas.internet.com/markets/advertising/article/0,,5941_1553231,00.html

Gulati, R., & Garino, J. (2000, May–June). Getting the right mix of bricks and clicks. *Harvard Business Review, 78,* 107.

Jap, S., & Mohr, J. (2002). Leveraging internet technologies in B2B relationships. *California Management Review, 44,* 4.

Jurvetson, S. (1998, November). Turning customers into a sales force. *Business 2.0.* Retrieved October 29, 2002, from http://www.business2.com/articles/mag/0,,12761,FF.html

Kirsner, S. (1999). Brand matters. *Fast Company,* Fall, 22. Retrieved October 29, 2002, from http://www.fastcompany.com/nc01/022.html

Lee, H. (2002). Aligning supply chain strategies with product uncertainties. *California Management Review, 44,* 3.

Moore, G. (1999). *Inside the tornado: Marketing strategies from Silicon Valley's cutting edge.* New York: Harper Collins.

Neilsen/Netratings (2002). *June 2002 global Internet index.* Retrieved October 29, 2002, from http://www.nielsennetratings.com/hot_off_the_net_i.jsp

Parmar, A. (2002, September 2). Second chances. *Marketing News, 36,* 18.

Peltier, J., Schibrowsky, J., Schultz, D., & Davis, J. (2002). Interactive psychographics: Cross-selling in the banking industry. *Journal of Advertising Research, 42,* 2.

Porter, M. (1975). *Note on the structural analysis of industries.* Boston: Harvard Business School Press.

Rewick, J. (2001, April 23). Choices, choices. *Wall Street Journal,* p. R12.

Saunders, C. (2002a, October 25). E-mail works for direct marketing. *CyberAtlas.* Retrieved October 29, 2002, from http://cyberatlas.internet.com/markets/advertising/article/0,,5941_1488681,00.html

Saunders, C. (2002b, October 16). E-mail efforts can impact brand. *CyberAtlas.* Retrieved January 12, 2003, from http://cyberatlas.internet.com/markets/advertising/article/0,,5941_1482921,00.html

Saunders, C. (2002c, July 29). Major sites missing out on B2B ad dollars. *CyberAtlas.* Retrieved January 12, 2003, from http://cyberatlas.internet.com/markets/b2b/article/0,,10091_1430501,00.html

Sechler, B. (2002, July 15). Behind the curtain. *Wall Street Journal,* p. R12.

Taylor, W. C. (1998). Permission marketing. *FastCompany, 14, 198.* Retrieved October 30, 2002, from http://www.fastcompany.com/online/14/permission.html

Thomas, D. (2002, June 6). Online exchanges offer competitive advantage. *Computer Weekly,* 10.

Winer, R. (2001, Summer). A framework for customer relationship management. *California Management Review, 43,* 4.

Medical Care Delivery

Steven D. Schwaitzberg, *Tufts-New England Medical Center*

INTRODUCTION

The Internet has inexorably changed the practice of medicine. In the "old days" the doctor–patient relationship was sacred. There was a direct line of communication of medical information from the physician to the patient. Patients were rarely found in medical libraries researching their own health-related issues because of the perception that those "hallowed halls" were places for doctors only. Magazine articles and popular press books were occasional sources of health information; however, these sources of information were of variable utility and were not highly regarded. The Internet has changed the nature of this once sacrosanct relationship. What was once a linear relationship is now a triangular one (Figure 1). The doctor–patient relationship is not all that is changed. The Internet and its predecessors have also changed forever the way physicians seek medical information. This chapter will explore the nature of the content available for medical care. The nature of how physicians utilize the Internet to conduct their daily business and medical education and research will be explored. Finally, how patients find and communicate with their doctors will be reviewed.

THE INTERNET AND THE PHYSICIAN

Searching the Reference Collections

Although the Internet dates back to origins in the 1970s, it had little impact on the daily life of the clinical physician until the vast expansion of the World Wide Web in the 1990s. Doctors communicated with each other by mail or telephone. They educated themselves by reading books or journal articles. Sifting through the vast array of journal citations in the giant reference collection entitled *The Indicus Medicus* was a chore that could take hours simply to find even a few references if the topic were obscure. Thus, when the National Library of Medicine (NLM) made MEDLINE available through dial-up interfaces such as Bibliographic Retrieval Service (BRS), Colleague, or Grateful Med to physicians for a monthly fee, it seemed like a godsend. Even at 2400 baud, which is a snail's pace by today's standards, the research process was greatly expedited as long as one did not need a reference that was indexed prior to 1966 (a limitation that still exists today). (There are references available from 1958 to 1965 on OLD MEDLINE, which is also available via the NLM Gateway.) These dial-up services interfaced through CompuServe, telnet connections, or their cousins, accessed front-end programs such as BRS, Grateful Med Ovid, or Paperchase, all of which searched through the same databases such as MEDLINE, Medlars, Cancer Lit, or Toxline. Searching was slow and the key to success was a smart search strategy. Too broad a search might reveal 1,000 articles in a variety of languages and too narrow a search might come up empty. In 1996 Ovid Technologies introduced the Ovid Web gateway that has enjoyed use by tens of thousands of people utilizing institutional subscriptions all over the world (Ovid Technologies, 2002). But the average physician in training could not spend what precious few dollars were available just to search the Web. Enter Physicians Online (POL). First a dial-up service (1992) and thereafter a Web-based service, POL was free to medical students and doctors (only) as long as one was willing to put up with a little advertising from your friendly neighborhood pharmaceutical company. POL provides a variety of services such as disease-based forums, continuing medical education (CME), and links to sponsors and health care organizations. POL provided to registered users an e-mail address so one did not have to buy into AOL, CompuServe, or Prodigy. Later in 1997, the NLM profoundly changed the nature of medical information and its dissemination when MEDLINE access was rendered free of charge to everyone. Physicians had gotten used to the idea that the medical databases were not for general consumption; after all, it took years of training to make any sense of the information. Clearly the genie is out of the bottle for good!

Now anyone with an Internet connection can peruse the vast array of medical knowledge from the comfort of his or her home or office. The abstracts of seemingly limitless journal articles can be provided in seconds when searches are conducted with high-speed access. In 2001, the NLM retired Internet Grateful Med in favor of PubMed as the primary search interface (http://www.

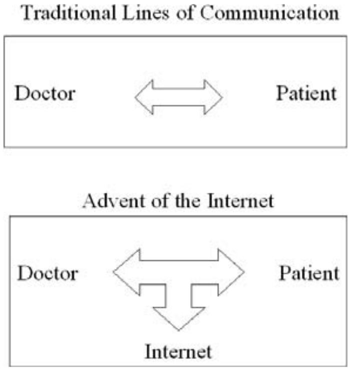
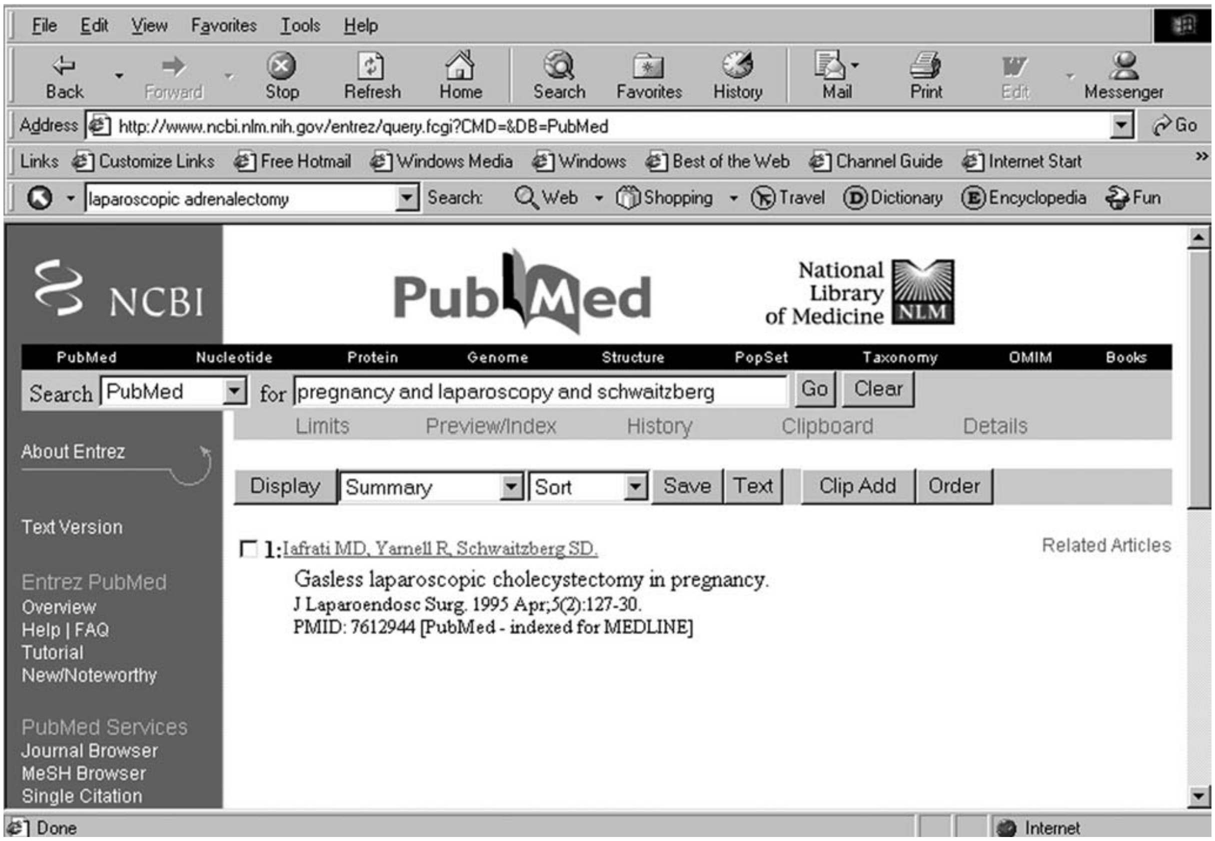


Figure 1: The lines of communicating health related information between doctors and patients have been transformed by the advent of the Internet from a linear to a triangular method of communication.

ncbi.nih.gov/entrez/query.fcgi or http://gateway.nlm.nih.gov/gw/cmd).

Searches via high-speed Internet access take only seconds providing abstracts of innumerable journal articles (Figure 2). PubMed is managed by the National Center for Biology Information (NCBI), which is a division of the National Library of Medicine at the National Institutes of Health. In addition to PubMed several other digital archives are available via the Internet such as PubMed Central (http://www.pubmedcentral.gov/), a digital archive of life science journal literature; Bookshelf (http://www.ncbi.nih.gov/entrez/query.fcgi?db = books), a growing collection of biomedical books that has been developed in collaboration with book publishers allowing Internet-based digital retrieval of text information; and Online Mendelian Inheritance in Man, which is a catalog of human genes and genetic disorders. There is a growing list of databases available free of charge to anyone such as the molecular modeling database that provides three-dimensional data and molecular reconstructions (Figure 3). Of all the offerings from the NLM perhaps the most intriguing is the images from the History of Medicine collection (http://www.ihm.nlm.nih.gov). The History of Medicine Division of the NLM has collected and digitized a large collection of images for ordering or downloading. Some of the material is copyrighted but much of it is not, available for use with a credit line "courtesy of the National Library of Medicine" (Figure 4).

Today most hospitals have high-speed access readily available. It is a common sight to see a physician in front of a terminal reviewing a bibliographic reference. But abstracts contain a very limited amount of data. The trip to



Address: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=&DB=PubMed

Search: laparoscopic adrenalectomy

Search PubMed for pregnancy and laparoscopy and schwaitzberg

Display Summary Sort

□ 1: Iafrati MD, Yarnell R, Schwaitzberg SD.
Gasless laparoscopic cholecystectomy in pregnancy.
J Laparoendosc Surg. 1995 Apr;5(2):127-30.
PMID: 7612944 [PubMed - indexed for MEDLINE]

Figure 2: A PubMed search can be as broad or as narrow as desired. Abstracts of the articles are often on hand. Links to full-text versions are increasingly available.

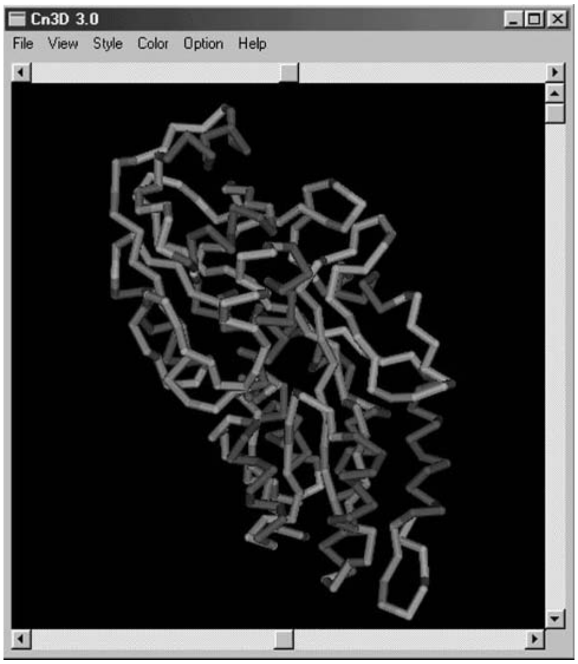


Figure 3: A search through the NCBI Web site will reveal an opportunity to download a free three-dimensional viewer that can display the molecular structure of a wide array of molecules available on the Web site.

the library (with references in hand) continues to be the mainstay of complete information acquisition, but this is changing as well. Full-text versions of journals are now appearing online. Often this requires a subscription to the print version of the journal, but not in all cases, especially after a period of time has passed since publication. If the experience learned from bibliographic citation access is an indicator then perhaps all journals will be free in the future; certainly once a critical mass of journals are available online for free, the remainder will have to follow or be left unread. Freemedicaljounals.com (http://www.freemedicaljournals.com) is leading the way in this regard. A wide number of journals can be searched in full-text format complete with the original illustrations in place. Will this replace the piles of journals in my office? Except for a few favorites I would be happy to maintain my literature collection online. Skeptics argue that the experience of reading print has yet to be reproduced or improved upon electronically, suggesting print journals will persist for some time to come. As high-speed Internet access continues to disseminate, it is only a matter of time before electronic versions of published journals begin to offer a multimedia experience not available in print. This will likely include peer-reviewed video submission or symposia. This will add substantial value to a journal subscription. "Internet only" journals have slowly appeared. ISPUB.com (http://www.ispub.com) has created a collection of 40 journals across the specialties of medicine and surgery that are available only on the Internet. Will

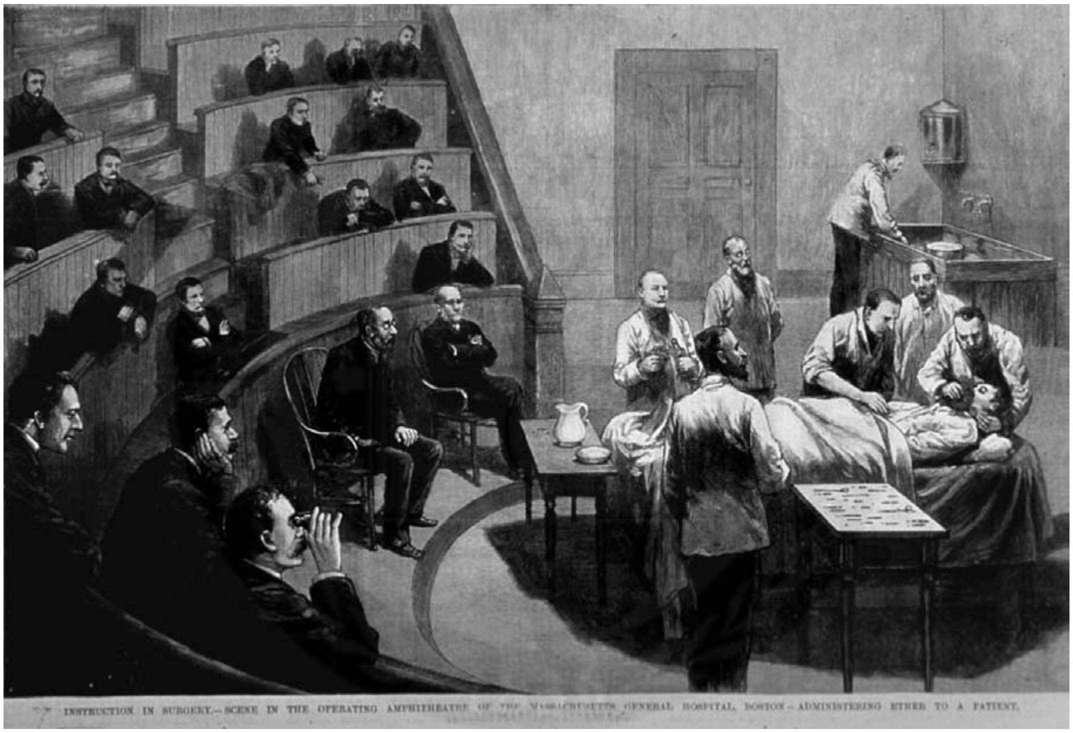

INSTRUCTION IN SURGERY.—SCENE IN THE OPERATING AMPHITHEATRE OF THE MASSACHUSETTS GENERAL HOSPITAL, BOSTON—ADMINISTERING ETHER TO A PATIENT.

Figure 4: This is an image sample from the History of Medicine Division Web sites of the National Library of Medicine. This image depicts a scene from the operatory of the Massachusetts General Hospital in the 1800s (courtesy of the National Library of Medicine).

physicians read these Internet-only journals? In general, medical journals are rated by "impact factor," a standard that may take years to reach a top quartile ranking. Ease of accessibility and low cost may facilitate the rise of journals such as these.

One could argue that there is too much information available to the health care practitioner today. With thousands of articles emerging every month, there is no way for anyone to keep up with the latest information concerning emerging treatments, popular therapies, and the like. The Cochrane Collaboration was established in 1993 to organize and review clinical trial information in order to provide an evidence basis for clinical decision making (Levin, 2001). The Cochrane Library contains over 1,400 reviews and is available on compact disk or via the Internet (http://www.update-software.com/cochrane). Thus the reader is treated to an unbiased distillation of large sections of the literature on a given topic. This is of enormous potential benefit now that we have entered the information (overload) era.

Virtual Meetings

Nearly every physician in America belongs to at least one professional society. Physicians in academic practice may belong to 10 or more although it is simply becoming too expensive to attend that many meetings. The Internet has begun to have profound impact on how medical professional societies and their physician (or nurse, therapist, etc.) members interact. Five years ago all of the dissemination of society information was done by mail; meeting notification, call for abstract submissions, and voting issues were all on paper. Today so much of that has changed with e-mail notification with embedded Web links that direct the reader to the society Web site for meeting registration, hotel reservations, and the like. Deadlines for sending abstract submissions for meeting presentation were a matter of working until the overnight carrier came to make the last pickup. Now these submissions are often worked on online, racing the clock until the Web site will not take any more offerings, alas, allowing for new refinements in procrastination. Web-based submission is maddeningly precise. When the instructions call for a 250-word limit—that is truly all you get! On the other hand an immediate receipt is generated providing a measure of relief for those last-minute submitters, who seem to be everywhere in the medical community. Medical society Web sites, when well constructed, provide a number of services to the physician as well as to the patient. Password-protected member databases provide for quick location of an out-of-town colleague and can be a great resource for a patient who is moving and needs his or her doctor to provide a contact in another city (Figure 5). Society guidelines for clinical care are often posted and downloadable. The cost of the Internet integration into the affairs of these organizations is offset by the decrease in postal costs and telephone time spent answering repetitive questions. Society matters can be voted on without expensive conference calls in many cases; meeting minutes are distributed electronically saving time and money. The evolution of the medical society continues with the appearance of the virtual meeting. Web-based meeting attendance in real time or thereafter for free or fee continues to increase. Even physicians who attend

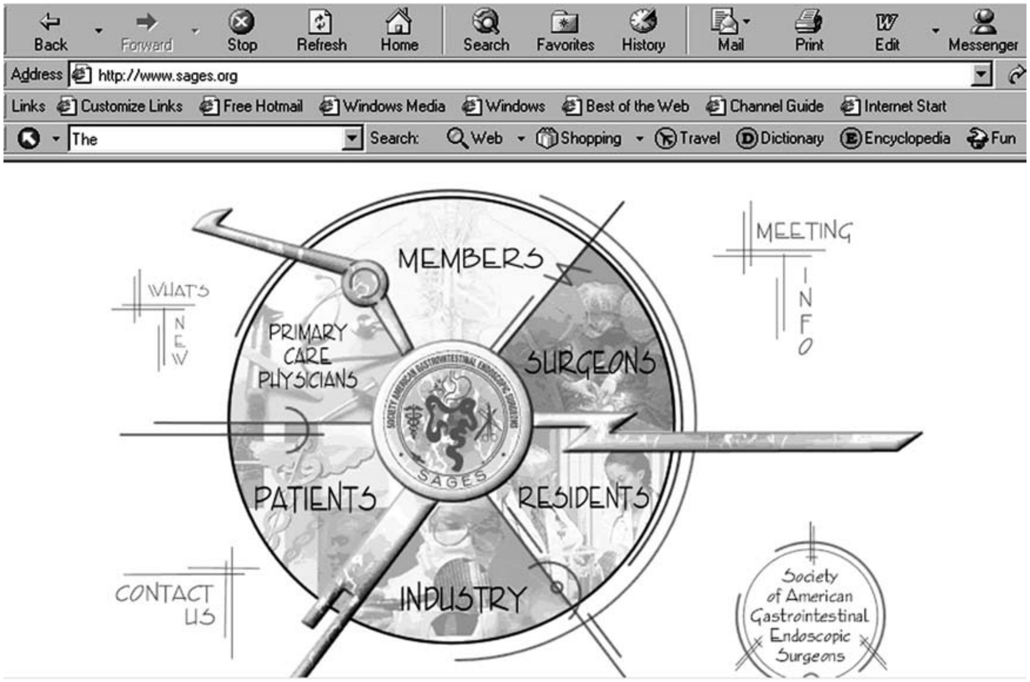
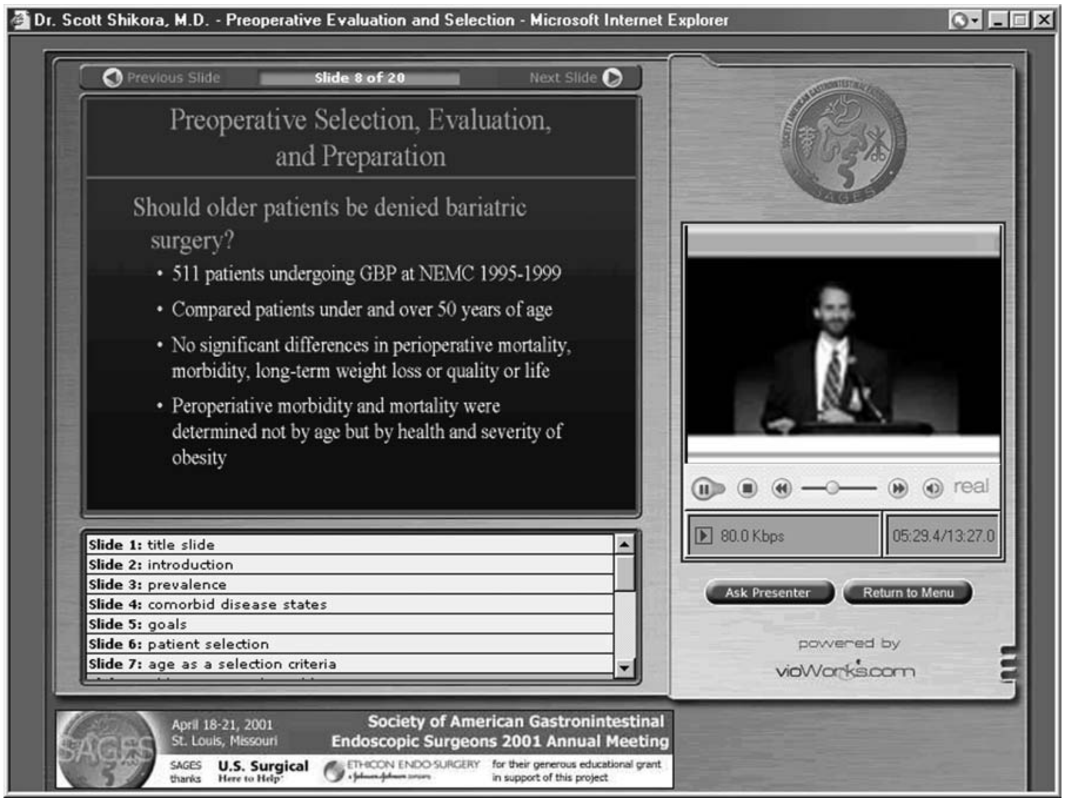

Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Print
Edit
Messenger
Address http://www.sages.org
Links Customize Links Free Hotmail Windows Media Windows Best of the Web Channel Guide Internet Start
The
Search: Web Shopping Travel Dictionary Encyclopedia Fun
WHAT'S NEW
MEMBERS
MEETING INFO
PRIMARY CARE PHYSICIANS
SURGEONS
PATIENTS
RESIDENTS
SAGES
CONTACT US
INDUSTRY
Society of American Gastrointestinal Endoscopic Surgeons

Figure 5: This is the home page of the Society of American Gastrointestinal Surgeons Web site. Options are immediately presented for surgeons, other physicians and patients directing the users to appropriate material.



Dr. Scott Shikora, M.D. - Preoperative Evaluation and Selection - Microsoft Internet Explorer

Slide 8 of 20

Preoperative Selection, Evaluation, and Preparation

Should older patients be denied bariatric surgery?

- 511 patients undergoing GBP at NEMC 1995-1999
- Compared patients under and over 50 years of age
- No significant differences in perioperative mortality, morbidity, long-term weight loss or quality or life
- Peroperiative morbidity and mortality were determined not by age but by health and severity of obesity

Slide 1: title slide
Slide 2: introduction
Slide 3: prevalence
Slide 4: comorbid disease states
Slide 5: goals
Slide 6: patient selection
Slide 7: age as a selection criteria

80.0 Kbps
05:29.4/13:27.0

April 18-21, 2001
St. Louis, Missouri

Society of American Gastronintestinal Endoscopic Surgeons 2001 Annual Meeting

SAGES thanks U.S. Surgical Here to Help ETHICON ENDO-SURGERY for their generous educational grant in support of this project

Figure 6: This is an example of a virtual lecture. The presentation is recorded on videotape at the actual meeting. The digital file of the Powerpoint presentation used at the time is then synchronized to the video that has been converted to a streaming format in a navigatable fashion.

a particular meeting cannot see all of the presentations and may choose to revisit the meeting virtually later. Will this affect meeting attendance numbers? This is unclear since physicians often come to a meeting for more than just the information. This is no small issue since most medical societies need the meeting registration income to survive. Will physicians pay to sit at their desks to watch a meeting even for the much-needed continuing medical education credits required to renew medical licenses in most states? That too has been debated. A recent survey indicated that less than 15% of physicians polled had used the Internet for CME activity (Lacher et al., 2000). Content is key but transmission quality is crucial too. While many physicians have high-speed access at work, attending these virtual meetings is more likely to be a home activity. Broadband access is needed in order to provide a good experience but residential penetration is limited. As of this writing only 15% of Internet users have parted with their dial-up modems. This is expected to increase to 30% by 2005, however (The YankeeGroup, 2002). The American Society of Clinical Oncology was among the first major societies to promote the virtual meeting dating back to 1999. The virtual meetings from 1999 to the present are available to physicians and the public alike with lectures, virtual tours, posters, and a virtual commercial exhibit floor. Lectures have generally appeared in two formats: Web-ready PowerPoint presentation synchronized to audio, and slide/audio presentations synchronized to a video of the actual lecture in a small window (Figure 6). A stream of 80 kbps produces reasonably smooth motion in a quarter screen window when viewed using a Real Player or Windows Media Player. Free of charge CME programs have been available through commercial sponsorship.

Drug Information

There are innumerable drugs that physicians might prescribe. The traditional source of information has been the Physicians Desk Reference (PDR), which usually is distributed free of charge to licensed physicians. As a large book it is quite inconvenient to carry around and must be replaced yearly, with intermittent supplements as needed. Drug information is always changing and the PDR is out of date on some topics even as it arrives. The Internet has changed this process for physicians able to take advantage of the adaptability of the Internet. Since 1996 the PDR has been available online (http://www.PDR.net) to medication prescribing health professionals (doctors, nurse practitioners, and physician assistants). Consumers are directed toward different sites specifically constructed for the layperson, such as http://www.gettingwell.com, which contain drug and disease information in lay terminology. The challenge to the PDR's leadership in the field of drug information began in 1998 when ePocrates Inc. was founded by Richard Fiedotin, MD, and Jeff Tangney. They created an Internet accessible database designed to be downloaded into a Palm OS-based personal digital assistant (PDA). This made a pocket-sized compendium of drug information not only available, but also updatable as frequently as needed by performing the "hot-synch operation," which updates the database in the PDA from the Web. Perhaps because it was free to physicians and medical students, it was an overnight sensation. By 2000 there were over 18,000 users with as many as 400 new users being added daily. This is important because medication errors may cause as many as 7,000 fatalities yearly (ePocrates, 2002). Systems such as these can only reduce this error rate. A recent study from the Harvard Medical School indicated that the physicians surveyed felt that one to two medication errors per week (of varying significance) were avoided using this system. Translated across the 200,000 physicians who had downloaded this database into their PDAs by 2002, a lot of mistakes have been avoided (ePocrates, 2002). This Web-based/PDA system has been so successful that PDR.net introduced a downloadable database for Palm OS and Pocket PC handheld devices in 2002. Both of these approaches offer a variety of links to other services, discussions, and information for the end users. Other less well-known (but seemingly unlimited in quantity) Web sites catering to the health care professional and consumers appear on search engines including drug information, Web-based medical consultation, lectures on audio, and the like.

Clinical Trials

Among the most discouraging tasks any physician can face is telling a patient that he/she has a cancer or any other serious disease. Patients often hear nothing else after the initial bad news. Later once the shock has worn off there is a thirst for knowledge. How bad is this cancer? What types of therapies are available? What about clinical trials? The National Cancer Institute (NCI) has simplified the task of getting accurate and up-to-date specific information in this regard for physicians (as well as patients.) The cancer.gov Web site provides the doctor with up-to-date information concerning clinical trials, research funding, and statistics never so readily available before the explosion of the World Wide Web.

Rural physicians especially can find out where a clinical trial for a specific malignancy is being conducted, make contact with the study coordinators, and establish care expeditiously for their patients. Similar information is maintained at the Web sites of cancer cooperative groups such as Eastern Cooperative Oncology Group, Children's Oncology Group, and Gynecologic Oncology Group, as well as others. These cooperative groups conduct a majority of the clinical cancer trials in cooperation with the NCI. Traditionally patients have been entered and randomized to one treatment or another over the phone or by fax. Recently there is a trend toward Web-based registration into cancer trials to facilitate tracking of cancer patients particularly in our mobile society. While there are privacy issues that need to be scrutinized, this trend clearly will grow in the future.

E-mail

Institutionally based physicians in particular have been consumed by e-mail. The author receives 30 or 40 hospital-based e-mail messages a day. The predecessor to e-mail was voicemail, which is a much slower process. Nonetheless time needs to be set aside daily to avoid a tragic e-mail overload, a task that did not exist 10 years ago. Within our institution there is a lot of patient care related e-mail traffic. If necessary the e-mail message can be printed for the paper chart. Replies can be made effortlessly to hard-to-reach physicians or surgeons who spend a lot time in the operating room or similar environs. It is now simply assumed in a hospital such as ours that everyone will read their e-mail daily if not several times a day. This is a vast improvement over the delays imparted by an institutional mailroom. Photos can be attached to e-mail messages within the institution. Medical charts today are still for the most part kept on paper. As we move, albeit slowly, toward electronic medical records (EMRs) the Internet/intranet will be a key component in this effort. The problem is that it is almost too easy now to move information, some of which is sensitive in nature, e.g., a serious disease or complication. It was this incredible ease that spawned the privacy regulations in the Health Insurance Portability and Accountability Act (HIPAA) of 1996. This legislation is quite encompassing and was signed into law in 1996, but the HIPAA Privacy Rule was not due to become operational until April 2003 for large institutions (April 2004 for small ones). It specifies to whom and what types of information can be distributed electronically (and on paper for that matter). Severe penalties are assessed on individuals and institutions that do not take adequate steps to protect the individually identifiable protected health information. As will be discussed later, this will potentially impact the electronic relationship between patients and their doctors.

THE INTERNET AND THE PATIENT

Why Does the Lay Public Search Medical Topics?

Almost half of the adult population in the U.S. alone has Internet access, according to a 1999 Harris Poll and three-quarters of these have looked for health information online (Taylor, 1999). In 2001, the breakdown of the activities of the 60 million users who went online to seek out health related information (*Smart Computing*, 2001) is as follows:

Look for general health care information:	69%
Look for mental health information:	34%
Purchase medicines, nutritional supplements, or other health care items:	21%
Seek advice from a doctor other than your own:	16%
Communicate with doctor:	6%
Participate in chat room discussions about health care topics:	7%
Participate in chat rooms on mental health topics:	6%.

A subsequent Harris Poll estimated that nearly 100 million people go online to get health-related information

(Taylor, 2001).Why don't people just talk to their doctors? That is a good question that does not need a complex survey to understand. First is the fact the appointments have to be scheduled. An Internet search can provide immediate gratification. Second is the fact that that time always seems to be limited in the office visit. Third is the pressure that physicians feel to see more patients in the short time allotted since reimbursements have been falling in the past few years. Last, doctors are hard to reach by phone. Thus it is not hard to understand why so many people have flocked to the Internet to seek health care information. The Harris Poll has termed these users "cyberchondriacs" surfing the Web for health-related information three times a month on average (Taylor, 2001). There are perhaps as many as 100,000 health related Web sites for them to visit (Eysenbach et al., 1999).

Many problems exist with this new paradigm of medical information transmission. The quality/accuracy of the information varies greatly from site to site. The information may not be presented at an appropriate level for the reading audience. The difference between marketing and information services may be hard to discern for many patients. Patients may be asked to divulge personnel health information that possibly could have unintended uses. Despite these limitations, the search for health care information and the purchase of medications and supplies (*eHealth* is the preferred term) (McLendon, 2000), etc., on the Internet is here to stay.

Where Patients Search for Information

Patients approach the Internet searching for health care information in a variety of different ways. The first separator is whether or not the searchers begin with a link on their default portal, a health portal, or a search engine. Obviously the major Web portals have defined relationships worked out in the background with preferred sites, but so might search engines. The best way to understand the differences is to look at primary "hits" to the same search term at the various entry points. We also find that exactly how a patient searches their health care issue will produce differing results as well. By way of example, suppose you have abdominal pain in the right upper quadrant of your abdomen. You are told that the cause of this pain is due to gallstones that can be removed surgically using some new techniques. You are given an appointment to see a surgeon in a week. How can you find information about your disease, the proposed procedure, or the surgeon recommended? Does it make a difference where or how you start? In April 2002, if one entered the designated "health" link on the AT&T Worldnet home page, you arrived at a site "powered by Health Scout." The same approach at MSN.com took you to a MSN site with a Web MD overlay. Prodigy took you to Prodigy health site. AOL took you to a different Web MD site but similar to the page you linked to from Netscape's Netcenter portal. However, Netscape used it own search engine. Yahoo and Alta Vista went to topical directories. For the patient described, the choice of what terms search terms to use is multifold. One could choose by organ, pathology, or surgery, i.e., gallbladder, gallstones, or laparoscopic cholecystectomy. The results of such a search via major Web portals are shown in Table 1.

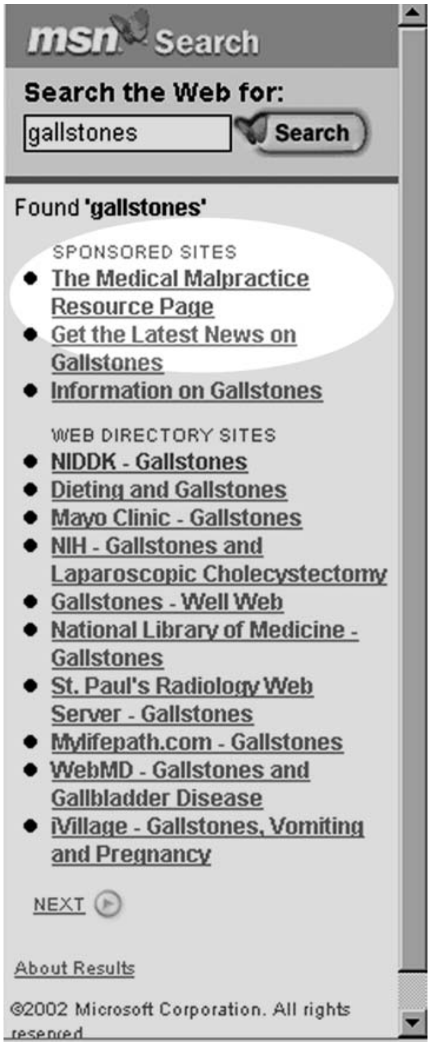
As you can see at the time of this writing, an initial search via the health links on the major portals, the Web MD search engine predominates. Between the health information search engines there is a wide variety of directed Web sites with varying degrees of usefulness and relevance. The direct use of search engines from the main portal sites produces different results. For instance, at the

Table 1 Portal-Based Internet Search Results (Article Title) for Health Information about Gallstones

		Search Term		
Portal	**Health Service**	**Gallbladder**	**Gallstones**	**Laparoscopic Cholecystectomy**
AT&T Worldnet	Health Scout	Can Vitamin C Shield Gallstones	*Can Vitamin C Shield Gallstones?*	*Study Faults Laparoscopic Surgery for Colon Cancer*
MSN AOL	Web MD	*Laparoscopic gallbladder surgery for gallstones—Surgery Overview*	Gallbladder cancer-Patient Information	*Laparoscopic gallbladder surgery for gallstones—Surgery Overview*
Prodigy	Healthology	Surgical Treatment for gallstones (Healthology article)	*Gallbladder Removal: Before and After the Surgery (webcast)*	-Surgical Treatment for gallstones (Healthology article)
Yahoo	Yahoo search	National Institutes of Diabetes and Digestive and Kidney Diseases information site	Laparoscopic gallbladder removal	Issue (from Gastonia Surgical Associates)
Netscape.com	WebMD	*Surgery Needed for Gallstones?*	*Surgery Needed for Gallstones?*	*Gallstones: Symptoms, Treatments, Prevention*

Source: Based on Internet search May 10th, 2002.

time of this writing, Web MD was used on the MSN.com health link, but a search could also be performed from the home page of the MSN portal. The search (search date: May 10, 2002) for gallstones (the same result as that for Yahoo, Google, Alta Vista, Lycos, and Excite) directed one to the National Institutes of Diabetes and Digestive and Kidney Diseases sites about gallstones. There was an interesting twist, however. The very first offering from all of these search engines (that was sometimes split into a "sponsored sites section") was a malpractice trial lawyer site (Figure 7)! It is doubtful that the medical community would have a warm feeling to the concept that on a vast array of general Internet search engines, the first site a prospective patient is directed to in a search about gallstones is a sponsored site about malpractice attorneys. This certainly raises the issue of what the public is "fed" from the Internet via the search process. Since anyone with enough money can afford to be a "sponsor," his or her site under the right circumstance can consistently be listed first.



Found 'gallstones'

SPONSORED SITES

- The Medical Malpractice Resource Page
- Get the Latest News on Gallstones
- Information on Gallstones

WEB DIRECTORY SITES

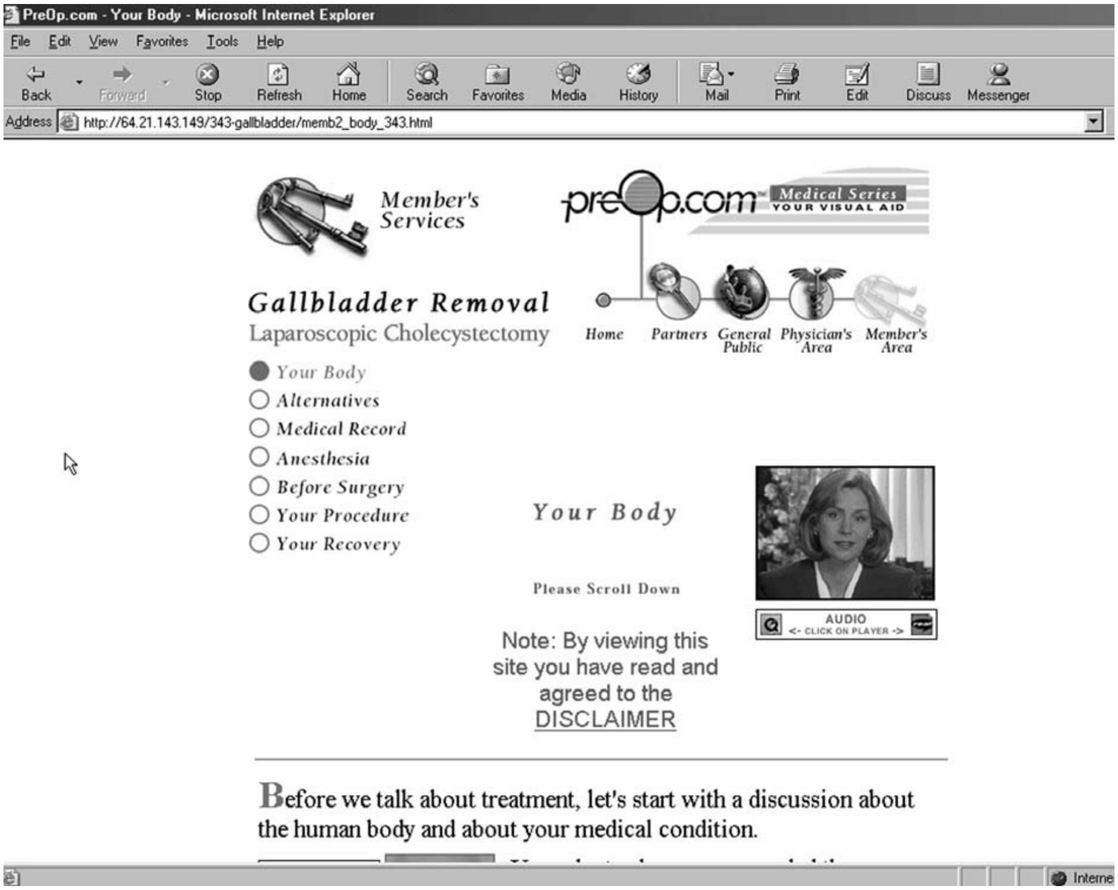
- NIDDK - Gallstones
- Dieting and Gallstones
- Mayo Clinic - Gallstones
- NIH - Gallstones and Laparoscopic Cholecystectomy
- Gallstones - Well Web
- National Library of Medicine - Gallstones
- St. Paul's Radiology Web Server - Gallstones
- Mylifepath.com - Gallstones
- WebMD - Gallstones and Gallbladder Disease
- iVillage - Gallstones, Vomiting and Pregnancy

©2002 Microsoft Corporation. All rights reserved

Figure 7: This figure depicts the results of a search under the term "gallstones." The sponsored sites are listed subtly ahead of the actual search results. When this search was performed a medical malpractice law firm owned the first sponsored site.

Understanding the limitations of the current system is central to the maturation of the Internet in this arena. The potential of the Internet to distribute information to patients and interested parties is limitless, but when the medical literature is searched for commentary on the quality and accuracy of medical information available on the Internet in 1999–2002 on any given medical topic, the results are disturbing. Medical information traditionally was transmitted via the doctor–patient relationship—that is no longer universally true as noted in the opening of this chapter. For the millions of Americans searching for health information on the Web, the physician is not a necessary part of the process (Hubbs et al., 1998). The literature is replete with articles that review the available Web-based information on a given topic, and clear trends are observed. Topical information at a given site is incomplete (Chen et al., 2000; Pandolfini et al., 2000; *The Quality Letter for Healthcare Leaders,* 2001; Gagliardi & Jadad, 2002). Topical information is often inaccurate or misleading (Appleby, 1999; Shon et al., 2000; Cline & Haynes, 2001; Kisely, 2002; McKinley et al., 1999; Soot et al., 1999; Hellawell et al., 2000; Libertiny et al., 2000; Mashiachet al., 2002). There are concerns about readability in the sense that the material is not often presented in a fashion appropriate for the reader (Berland et al., 2001; Murero et al., 2001). There are concerns about commercialism (Forsstrom & Rigby, 1999; Mabrey, 2000; Winker et al., 2000; Smithline & Christenson, 2001; Tatsumi et al., 2001). Finally, minority interests do not seem well served as seen in the paucity of high-quality translations into Spanish (Berland et al., 2001). To highlight this, the search for Web-based information on pediatric cough was reviewed. More than half of the 19 sites identified had more incorrect information than correct information (Pandolfini et al., 2000).

There are terrific medical information Web sites (if one searches persistently) currently available; however, the ideal Web site is not ideal for everyone. It is a simple fact that health care providers and the lay public have differing needs and ultimately need to be directed to different types of sites. (This is a major problem that will ultimately be overcome.) For example, it is well established that the information for obtaining consent for medical procedures or medical research trials needs to be promulgated on a fifth- to eighth-grade level in order to be understood even by relatively well-educated people (Davis et al., 1990; Hopper et al., 1998; Hochhauser, 1999). Yet again this is an arena where the Web should excel. Typical medical patient education is text based. Patient satisfaction and comprehension is relatively low with this type of material. Multimedia approaches can improve patient satisfaction (Agre et al., 1997). Especially with the march of the Web toward high-speed access, the Internet should become the penultimate tool in physicians' education armamentarium. Multimedia educational sites


PreOp.com - Your Body - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss Messenger
Member's Services
preOp.com Medical Series YOUR VISUAL AID
Home Partners General Public Physician's Area Member's Area
AUDIO <- CLICK ON PLAYER ->
Internet

Address http://64.21.143.149/343-gallbladder/memb2_body_343.html

Gallbladder Removal
Laparoscopic Cholecystectomy

● Your Body
○ Alternatives
○ Medical Record
○ Anesthesia
○ Before Surgery
○ Your Procedure
○ Your Recovery

Your Body

Please Scroll Down

Note: By viewing this site you have read and agreed to the DISCLAIMER

Before we talk about treatment, let's start with a discussion about the human body and about your medical condition.

Figure 8: Preop.com Web site (retrieved May 13, 2002). This is an example of a educational site that is not affiliated with medical devices, pharmaceutical firms, or health care systems.

have been arriving on the scene for a few years now with (medically) professionally edited content, artwork, and compressed video to educate patients on specific topics. Preop.com (http://www.preop.com) is an example of this concept. The material is prepared by multimedia specialists without a link to a specific health related industry (drug, health care system, device, etc.) in concert with health care professionals with a stated goal of providing balanced preoperative information targeted at the lay population. This approach solves many of the problems that occur when individual physicians or even hospitals provide Web-based material as just noted (Figure 8).

Is the Information Self-Serving?

The major health gateways and other sites have recognized the potential for commercialism, conflict of interest, and poor quality information on the Web in general and on their sites in particular. Some organizations have banded together to articulate a code of conduct to prevent or minimize this behavior called HONcode. The principles behind this code are eight-fold and are shown in Table 2. This code of conduct developed by a Swiss nonprofit organization, Health on the Net Foundation (HON), was introduced in 1996 and revised in April 1997. It has been translated into 19 languages (Boyer et al., 1998; Boyer et al., 1999). The HONcode seal (Figure 9) is an active hyperlink that, when authentic, links the user to the HON site and displays the site's unique PIN and HONcode status. This is not the only attempt to raise the reliability and credibility of the health information on the Web. Other organizations, such as the American Medical Association (AMA), which has developed Web site guidelines (http://www.ama-assn.org/ama/pub/category/ 1905.html); the Internet Healthcare Coalition (http://www.ihealthcoalition.org), which sponsors the eHealth Code of Ethics; URAC (http://www.websiteaccreditation.urac.org); and Hi-Ethics Inc. (http://www.hiethics.com/Principles/index.asp) are working toward similar goals. A logo or seal displaying the accreditation is placed on the Web site when reviewed by these organizations as well. Like the HONcode, included in most of these guidelines are comments about the confidentiality of the personal health information transmitted by the end user. This is critical because when a consumer truly becomes a patient due to a medical problem acute or chronic he/she clearly becomes more vulnerable and may

Table 2 HONcode Principles

Authority	Any medical or health advice provided and hosted on this site will only be given by medically trained and qualified professionals unless a clear statement is made that a piece of advice offered is from a nonmedically qualified individual or organization.
Complementarity	The information provided on this site is designed to support, not replace, the relationship that exists between a patient/site visitor and his/her existing physician.
Confidentiality	Confidentiality of data relating to individual patients and visitors to a medical/health Web site, including their identity, is respected by this Web site. The Web site owners undertake to honor or exceed the legal requirements of medical/health information privacy that apply in the country and state where the Web site and mirror sites are located.
Attribution	Where appropriate, information contained on this site will be supported by clear references to source data and, where possible, have specific HTML links to that data. The date when a clinical page was last modified will be clearly displayed (e.g., at the bottom of the page).
Justifiability	Any claims relating to the benefits/performance of a specific treatment, commercial product, or service will be supported by appropriate, balanced evidence in the manner just described in the attribution principle.
Transparency of authorship	The designers of this Web site will seek to provide information in the clearest possible manner and provide contact addresses for visitors that seek further information or support. The webmaster will display his/her e-mail address clearly throughout the Web site.
Transparency of sponsorship	Support for this Web site will be clearly identified, including the identities of commercial and noncommercial organizations that have contributed funding, services, or material for the site.
Honesty in advertising and editorial policy	If advertising is a source of funding it will be clearly stated. A brief description of the advertising policy adopted by the Web site owners will be displayed on the site. Advertising and other promotional material will be presented to viewers in a manner and context that facilitates differentiation between it and the original material created by the institution operating the site.

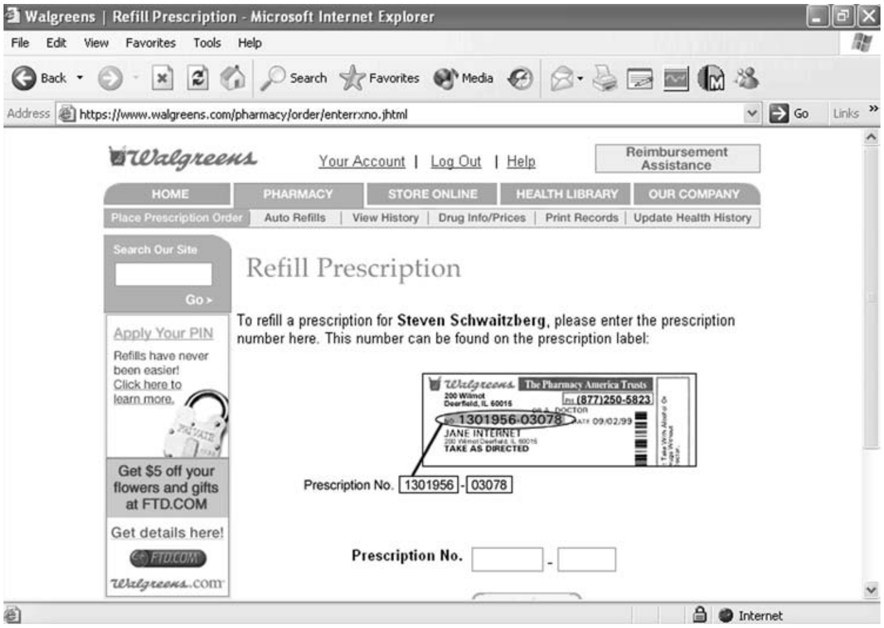
Source: Adapted from Boyer et al. (1998).

be more willing to divulge personal information otherwise unattainable. Protecting the patient is critical in this setting and yet we have no idea how well this is being accomplished.

How Do We Evaluate Web-Based Medical Information?

To make matters worse, in sorting out the overabundance of material on the Web, there is almost no consensus as to how to rate health care Web sites for quality. Even if the webmasters are well intentioned, is the information any good? This is a real problem since it has been said "trying to get information from the internet is like drinking from a fire hose, you don't even know what the source of the water is" (McLellen, 1998). In 1998, Jadad reviewed the rating systems available to evaluate health related information. He discovered 47 instruments for this purpose. Few provided criteria for the rating or instructions for use. None provided evidence of validity. His conclusion was that these instruments were incompletely developed and not useful in general (Jadad & Gagliardi, 1998). In 2002 the same group revisited the issue, 51 new instruments were available and were generally plagued with the same problems especially the lack of validity of the rating instruments that was seen four years earlier (Gagliardi & Jadad, 2002). What we are left with is information overload. With little to help the user sorts out the wheat from the chaff (Craigie et al., 2002), perhaps health care providers (physicians, nurse practitioners, physician assistants, and the like) should identify Web sites themselves that they feel have high quality information and direct patients in that fashion. A recent poll indicated that a vast majority of patients would desire this, but few receive this advice from their physicians (Ullrich & Vaccaro, 2002). Clearly there is a need to strive toward this since rating scales are unreliable at present and the "seals of approval" emerging appear as honest attempts to raise the quality and safety while minimizing the commercialism of the few thousand at best Web sites that display them.

Figure 9: HonCode logo. The seal serves as a hyperlink to the Honesty on the Net Foundation Web site that provides site verification and PINs to assure users that the Web site is a legitimate participant in the HONcode program.


Walgreens | Refill Prescription - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://www.walgreens.com/pharmacy/order/enterrxno.jhtml
Go Links
Walgreens
Your Account | Log Out | Help
Reimbursement Assistance
HOME
PHARMACY
STORE ONLINE
HEALTH LIBRARY
OUR COMPANY
Place Prescription Order | Auto Refills | View History | Drug Info/Prices | Print Records | Update Health History
Search Our Site
Go ›
Apply Your PIN
FTD.COM
Walgreens.com
Walgreens The Pharmacy America Trusts
200 Wilmot Deerfield, IL 60015
(877)250-5823
DOCTOR
1301956-03078
09/02/99
JANE INTERNET
TAKE AS DIRECTED
Internet

Refill Prescription

To refill a prescription for **Steven Schwaitzberg**, please enter the prescription number here. This number can be found on the prescription label:

Prescription No. 1301956 - 03078

Prescription No. □ - □

Refills have never been easier! Click here to learn more.

Get $5 off your flowers and gifts at FTD.COM

Get details here!

Figure 10: An example of a Web site that can be used for medication refill.

Prescription Drugs on the Web

As noted earlier, the activity that ranks after information retrieval is the acquisition of medical supplies especially drugs or supplements. Online prescribing represents an evolutionary step from mail order businesses in that minutes after an online pharmacy Web site is established, a potential wordwide audience is established as well.

Recently, the drug that has received the most attention from Internet users is Sildenafil, otherwise known as Viagra (Pfizer Inc). In theory, at least, men with a history of sexual dysfunction could answer related questions in the privacy of their on homes. In some cases this may result in a more thorough sexual history than what might be obtained in a doctor's office (Jones, 2001). The advantages of Internet-based prescribing probably stops there. The Internet does not provide a doctor–patient relationship within the context of a state issued licensed. The failure to achieve this and go on to sell these medications to patients is a breach of the ethical considerations (Kahan et al., 2000). Completing an online questionnaire and paying a fee for a medication have been termed the modern equivalent of "diagnosis by mail" and has long been condemned and in fact has resulted in physician disciplinary action in Wisconsin (Thexton, 2002).

Other states are following suit (Carnall, 1999). At least in the United States this practice may be risky business (Rice, 2001). The most thorough evaluation of Internet prescription practices concerning Viagra was performed by Gunther Eysenbach in 1999 when he masqueraded as a totally inappropriate patient for the use of the drug and attempted to purchase it. Of the 22 Web sites that were identified at the time (more are online as of this writing): two required a written prescription by a "real" physician, nine dispensed the drug without any prescription at all, and 11 issued an "online prescription" after an alleged physician reviewed the online order form containing medical questions (Eysenbach, 1999). He went on to attempt to complete the transaction with 10 of the latter group and found three companies who actually mailed the drug (from Europe). Several companies did decline to dispense the drug; nonetheless it was clear that a patient who listed several contraindications to the use of Viagra could easily obtain it from the Web. Minors with Web access could do the same. Other drugs that seem to lend themselves to this same practice include those for weight loss, hair loss, birth control, pain (nonnarcotic), excess hair, skin care, cessation of smoking, and antivirals. The American Medical Association has come out strongly against this practice to the point of singling out Web sites who prescribe in this fashion. However, the AMA does support the use of the Internet to prescribe medications with appropriate safeguards (American Medical Association, 1999). The obvious use for Internet-based prescribing is in the refilling of chronically used drugs with a legitimate prescription on file. This is already an acceptable snail mail and telephone-based practice made simpler on the Web (Figure 10). At least in the U.S. the use of the Internet for drug prescription will continue to evolve with increasing scrutiny with respect to ethical and medical standards of care.

PATIENT TO DOCTOR AND PATIENT TO PATIENT

Doctors and Patients

Electronic Communication

Patients reach physicians via the Internet in three ways: they can e-mail their doctors directly, they can send a "cold call" e-mail to a physician seeking advice, or they can go to chat sites specifically run by physicians to deal with health topics. Should a physician give his/her e-mail address to a patient? Clearly this is not a yes or no question, but one that requires a fair amount of thought before a physician does place this information on his/her business card. Currently it seems that less than 10% of physicians communicate with patients in this fashion (Lacher et al., 2000). The physician who places his/her e-mail address on a business/appointment card is saying by implication "you may contact me by using this method" analogous to the telephone or fax number also listed. As such once an inquiry is received in this fashion from an established patient, there is an obligation to respond. Does this imply that physicians are now obligated to routinely check their e-mail messages on a daily basis or more frequently? What about vacations or professional travel? There is certainly incomplete adoption of computer technology in the medical community at this time and it is not to be assumed that all physicians check e-mail on a daily or more frequent basis. Like all medical communications the answers can be short or long. Short answers are simple for the physician to respond to via e-mail and probably more convenient than telephone. Long ones are problematic. The typing can take much longer than a conversation and the clarity of the communication dwindles, as many physicians perceive themselves as overburdened from a time management standpoint, reducing the attractiveness of communicating by e-mail (Kleiner et al., 2002). On the positive side there is a record of the interaction, but on the negative side there is a record of the interaction that travels through a cyberspace environment that is potentially open to prying eyes. E-mail within a major medical institution is generally secure as a major effort to protect patient confidentially is a major thrust of most information technology groups, recognizing that there is a huge interphysician volume of traffic much of which is patient related. On the other hand once e-mail leaves the medical center or a private physician's account through an ISP, security cannot be guaranteed. This is especially true for unencrypted wireless users. As such communicating in this fashion carries risk, one that should be pointed out to each patient every time a medically related e-mail transmission occurs. While it is possible to establish encrypted-only communication as a basis for e-mail communication with patients, this is not currently a practical solution or common practice. Continuing a dialog by e-mail that may have been recently initiated in the office may prove useful to both patient and physician, but receiving e-mail from a patient not recently treated with a list a symptoms with or without a request for medication is far more problematic. This is more similar to the diagnosis by mail trap that is distinctly frowned upon as mentioned and carries a risk of an inappropriate medical action on the part of the physician. However, is this really so different from the millions of phone calls taken by physicians "on call" each year? (Probably not.) Certainly it is also a matter of degree and situation in so far as the physician who has already been plagued with reimbursement reduction by Medicare and other agencies of late may construe chronic e-mailing for health advice as an attempt to avoid an office visit charge. On the other hand, for the patient who has to wait unpredictable amounts of time in the office for even a short amount of face-to-face time with a physician, e-mail is a godsend. This is even more relevant for the disabled patient who may have real issues just getting to the doctor's office. In 1998 Kane and Sands summarized the effort of an American Medical Informatics Association task force creating guidelines concerning the use of e-mail between patients and physicians in established relationships (Kane & Sands, 1998). These principles are shown in Tables 3 and 4.

The situation gets stickier for the physician who receives an unsolicited e-mail message. In this setting there is no established doctor–patient relationship. Is there a duty of a physician to reply? Would advice rendered in this circumstance be covered under "good samaritan" provisions given that there is no financial relationship? A few studies have shown that about half of the physicians who receive unsolicited e-mail will respond (Eysenbach & Diepgen, 1998; Oyston, 2000; Sing et al., 2001). In many cases the information received from the patient is accurate but incomplete. In most cases the patient is

Table 3 Communication Guidelines for E-mail between Health Care Providers and Patients

Establish turnaround time for messages. Do not use e-mail for urgent matters.
Inform patients about privacy issues. Patients should know who besides addressee processes messages during addressee's usual business hours, during addressee's vacation, or illness and that the message is to be included as part of the medical record.
Establish types of transactions (prescription refill, appointment scheduling, etc.) and sensitivity of subject matter (HIV, mental health, etc.) permitted over e-mail.
Instruct patients to put category of transaction in subject line of message for filtering: "prescription," "appointment," "medical advice," or "billing question."
Request that patients put their names and identification numbers in the body of the message.
Configure automatic reply to acknowledge receipt of messages.
Print all messages, with replies and confirmation of receipt, and place in patient's paper chart.
Send a new message to inform patient of completion of request.
Request that patients use autoreply feature to acknowledge reading provider's message.
Maintain a mailing list of patients, but do not send group mailings where recipients are visible to each other. Use blind copy feature in software.
Avoid anger, sarcasm, harsh criticism, and libelous references to third parties in messages.

Source: Adapted from Kane and Sands (1998).

Table 4 Medico-Legal and Administrative Guidelines for the Use of E-mail in the Health Care Setting

Consider obtaining patient's informed consent for use of e-mail. Written forms should itemize terms in Communication Guidelines, provide instructions for when and how to escalate to phone calls and office visits, describe security mechanisms in place, indemnify the health care institution for information loss due to technical failures, and waive encryption requirement, if any, at patient's insistence.

Use password-protected screen savers for all desktop workstations in the office, hospital, and at home.

Never forward patient-identifiable information to a third party without the patient's express permission.

Never use patient's e-mail address in a marketing scheme.

Do not share professional e-mail accounts with family members.

Use encryption for all messages when encryption technology becomes widely available, user-friendly, and practical.

Do not use unencrypted wireless communications with patient-identifiable information.

Double-check all "To:" fields prior to sending messages.

Perform at least weekly backups of mail onto long-term storage. Define "long-term" as the term applicable to paper records.

Commit policy decisions to writing and electronic form.

Source: Adapted from Kane and Sands (1998).

advised to see a physician. Clearly not every patient–physician interaction demands an in-person visit and physical examination. General health information in the form of premade e-pamphlets could be distributed in response to unsolicited e-mail messages with minimal risk. Others have suggested using the automatic reply format indicating general instructions of how to obtain care or a simple statement that a response is not forthcoming (Oliver, 1999). Eysenbach has suggested that patients have to be educated that it is unethical to diagnose and treat their medical conditions over the Internet via e-mail alone in the absence of a preexisting patient–physician relationship (Eysenbach, 2000). More advanced telemedical technology may make the virtual house call a reality in the future. Other groups including ours have performed teleconsultation to specialized remote locations including foreign countries using ISDN telecommunication lines. This allows for static x-ray review, cardiac and pulmonary auscultation, and direct patient interview. Current Internet protocol bandwidth for medical purposes will be supplanted by Internet2, ultimately replacing ISDN for this purpose (see later). Although the medical provider cannot quite examine patients in their own homes as yet, Web-based protocols have been used to augment the management of a variety of chronic diseases. Patients can fill out Web-based forms about their current status which trigger case manager responses when criteria are met (*Disease Management Advisor,* 2000). Telemonitoring equipment for diseases such as diabetes or asthma could be adapted for Internet transmission to providers alerting providers to situations that need attention (Cai et al., 2000). Seniors have been willing to embrace Web-based strategies for follow-up (*Disease Management Advisor,* 2001a, 2001b).

As in any shift away from the traditional office-based medical interaction, there are challenges to overcome (*Disease Management Advisor,* 2001a, 2001b). How will providers be reimbursed not only for the interaction, but also for the investment into the e-health infrastructure? Also like all electronic medical communications issues of security and liability require solutions before widespread diffusion. Despite these obstacles there is an undeniable appeal to utilize Web-based interactions in the management of chronic illness.

A variation in the unsolicited e-mail is the Web site that offers advice in response to a question. These sites can be general or disease specific. Some are free, but others offer "consultations for as low as. . . ." Each of these "ask the doctor" sites offers a disclaimer of some sort that varies with the site warning patients that emergent needs are best served by seeking acute medical care or that the progress of medical research may render the advice on site obsolete. Whether the advice or information is rendered for free or fee the existence of these sites raises an interesting question concerning the practice of medicine across state lines. What happens when a patient experiences an adverse outcome as a result of the advice from such a Web site? Is this different from a physician's health column in newspapers or magazines that respond to specific questions? The answer to the latter is probably yes in the fact that the patient is asking a question in an obvious public forum and the free of charge public response is subjected to editorial review that provides a measure of protection against dispensing risky advice. Most Web sites have no oversight in this regard especially when provided as a fee for service. Will we see a new type of malpractice action in the future? Time will tell.

Electronic Medical Records

The medical record has grown from skimpy hand and typewritten notes decades ago to a complex repository of medical action, interaction, and outcome. Despite extraordinary theoretical capability, the infiltration of the truly electronic medical record is small. There are compelling reasons to employ EMR (Grimson, 2001). EMR is likely to be more legible, thus more accurate than conventional records. The record would be continuously available (unlike a medical records room). The codification of data would lend itself to research, audit, and review with far greater efficiency than currently possible. Electronically available laboratory data, radiology reports, operative notes, and discharge summaries are common, but paper medical records in offices and hospitals chronicling day-to-day actions and patient–practitioner interaction are the norm. Progressive centers exist where the electronic medical records not only thrive, but thrive in a Web-based milieu (Dreyer et al., 2000; Safran & Goldberg, 2000). This is an integration of two potentially separate concepts since the Internet and EMR could coexist in total isolation. A successful EMR needs to become more than just a repository for laboratory data and reports. The inclusion of electrocardiographic, telemetric, photographic, and radiologic (via Picture Archiving and Communications System utilizing the Digital Imaging and Communications in Medicine standard) data is the future of the EMR. The EMR also must evolve to chronicle all interactions between a wide variety of practitioners (not only physicians) in disparate geographic locations in order to be truly comprehensive and reach the ultimate potential possible. This is the tricky part. It is a relatively straightforward process to build a closed, proprietary hospital network based system. Records, data, and image integrated EMR have been constructed. The INFODOM system prototyped in Italy utilizes an extensible markup language (XML) based medical record. The electronic patient record is delivered over the Web via secure socket layer as an XML document and viewed in a browser. Clinical images within the browser are seen on the JavaBeans image viewer, which can display static and animated sequences (Boyer et al., 2001). However, no one really receives care in a single health care system in the United States. People move, travel, and often seek care from a number of health care systems even within a single city. All of this information sits in isolated pockets, frequently duplicated needlessly. In the author's clinical experience, many patients do not remember who performed a prior surgery or even which hospital it took place in. The impact of this can be dramatic. As surgical procedures become less invasive; in fact performed through tiny "keyhole" incisions, the ability to predict the nature of a prior surgery (e.g., appendix, gallbladder, or uterus) based on incision location is lost. This can lead to great inefficiencies and misdiagnosis. The ability to access "cradle to grave" information would be a tremendous boon for the care of patients. Consolidated health care systems found in somewhat smaller countries will provide far greater efficiencies than what seems possible in the U.S. at the present time. For instance, the United Kingdom's National Health Service has outlined a strategy for accomplishing this task (National Health Service, 1998). This system known as NHSnet is not deployed over the Internet (although there is a secure Internet gateway for worldwide SMTP users), rather it is available through service providers. Adoption has been slow as reservations concerning cost and confidentiality over this proprietary network have been raised (Chadwick et al., 2000). The Internet provides an opportunity for global health care information management. The realization of this can only be enacted through the development of health information standards, which would allow for data transmission globally in a secure fashion. In fact, Section 262 of the Health Insurance Portability and Accountability Act requires the Secretary of Health and Human Services to adopt standards for health care transactions to enable electronic exchange of health information (Department of Health and Human Services, 2000). One of the most far-reaching standards is Health Level Seven (HL7). Health Level Seven is one of several ANSI-accredited standards developing organizations operating in the health care arena. It was created in the 1980s to address the need to communicate between disparate and often-proprietary health care information systems. The HL7 domain is clinical and administrative data (Health Level Seven, 2002). The standard corresponds conceptually (does not actually conform) to the highest level defined by the International Standards Organization's open systems model for application-to-application communication. This standard supports the functions needed to conduct secure health information management.

On the other hand it is not clear that everyone wants such an all encompassing, globally accessible, longitudinal health information management system. This is particularly true when it is not clear to the individual patients where the information actually resides. HIPAA legislation has brought privacy issues in health care to the foreground. Should a patient worry about the consequences of an employer finding out about past medical or psychiatric care? As we unravel the mysteries of the genetic sequence, would it be unreasonable for a woman to be concerned about her insurability if she were carrying mutations in the BRCA1 gene that appears to have an association with an increased risk of breast and ovarian cancer? Of course it is reasonable to be concerned. Already many insurance companies require a medical records review prior to issuing insurance. But this concerns actual medical events. Does a commercial organization have a right to know about a "predisposition" that may never materialize? How safe is this private information? The news has been replete with stories of computer security failure. The public faith in cyberspace security is certainly not resolute as manifested by the unwillingness of many to complete financial transactions online. In the U.S., the public will need to be convinced that their health information is truly secure and that there will be some measure of personal control over access to the information. Until then health care networks will continue to develop proprietary health information management systems accessible internally via intranet or Internet hopefully utilizing established open standards that will someday allow for their interconnection.

Patient to Patient

The disease-specific sites for neurologic or hepatic diseases for instance (just to name a few) offer forums where patients with common problems can communicate with one another creating the virtual support group. The ability to communicate long distances cheaply and instantly brings groups of patients together with common problems. They can get recommendations or simply commiserate. This can be a tremendous asset for patients with uncommon problems where insufficient critical mass for local support groups leads to isolation and often depression. There are data to indicate that supervised e-mail chat groups for patients with chronic illnesses can have a positive impact on medical care utilization (Lorig et al., 2002). Sadly, like all potentially useful tools, there is potential for abuse. Case reports of intentionally false stories propagated within the Internet support group forums have been demonstrated demonstrating pathology of another kind (Feldman, 2000). Patients with long-standing illnesses are a vulnerable population to abuse, and the aforementioned safeguards should be sought out whenever possible.

FUTURE DIRECTIONS

The next phase of growth for medical applications on the Internet will require increased bandwidth for the transmission of video for telemedical applications, real-time radiology, and remote patient examination. This will most likely be accomplished on Internet2, which began in 1996 as a result of interest from corporations, universities, and nonprofits in doing advanced research on a faster electronic network than the one offered by the Internet. The project acts as a laboratory for companies and researchers developing new technologies and also serves as a blueprint of what the Internet could look like in the future. Internet2 is deployed largely over a pair of high-performance backbone networks called Abilene and vBNS+. Signals are transmitted at speeds far greater than what is currently available on the Internet down fiber-optic lines to the (currently 190) universities and other places that are tied to the high-capacity networking centers. This project will bring the kind of multimodal data transmission that will make a whole range of new dreams possible such as Internet based robotically controlled telesurgery.

Summary

The Internet's potential in health care delivery will continue to grow. Unlimited multimedia information can be made available to health care provider and consumers. Our immediate task will be to sort out the wheat from the chaff from this bounty of data thus increasing the viability of eHealth itself. Educating patients about the diagnostic and privacy limitations of the Internet in its current form will also serve to improve overall satisfaction. Will the lack of Internet access for some widen the health care divide? In time we will be able to answer the question as to whether or not the absolute quality of health care has been improved by the advent of the Internet.

GLOSSARY

DICOM/PACS The American College of Radiology and the National Electrical Manufacturers Association formed a joint committee in 1983 to develop a standard to promote communication of digital medical image information, regardless of device manufacturer. This has facilitated the development of picture archiving and communication systems that can also interface with other systems of hospital information.

Electronic medical record (EMR) The EMR is the digitized representation of the health information for a given patient. Ideally this would be an all-encompassing cradle to grave accounting that is universally accessible under the proper circumstances. Realization of this standard will require many years.

Health Insurance Portability and Accountability Act of 1996 The major provision of this law was to ensure the continued health care coverage for workers when they lose or change jobs. However, there are broad implications concerning the protection of a patient's personal health information (PHI) in this law as well in 2003. These provisions govern to whom PHI can be disclosed whether electronically or on paper.

MEDLINE MEDLINE is compiled by the U.S. National Library of Medicine. It is the world's most comprehensive source of life sciences and biomedical bibliographic information. It contains nearly 11 million records that are indexed from more than 7,300 different publications dating from 1965–present. There are additional related libraries that index specialized information as well.

Virtual meeting In this context the virtual meeting refers to a health care practitioner attending a professional conference and its educational sessions from an Internet connection in real time or at a later date. Continuing medical education credits would be issued for this activity.

CROSS REFERENCES

See *Distance Learning (Virtual Learning); Electronic Data Interchange (EDI); Health Insurance and Managed Care; Health Issues; Internet Navigation (Basics, Services, and Portals); Legal, Social and Ethical Issues; Web Search Fundamentals.*

REFERENCES

Agre, P., et al. (1997). Patient satisfaction with an informed consent process. *Cancer Practice, 5*(3), 162–167.

American Medical Association. (1999). Retrieved May 13, 2002, from http://www.ama-assn.org/sci-pubs/amnews/pick_99/biza0719.htm

Appleby, C. (1999). Net gain or net loss? Health care consumers become Internet savvy. *Trustee, 52*(2), 20–23.

Berland, G. K., Elliott, M. N., Morales, L. S., Algazy, J. I., Kravitz, R. L., Broder, M. S., et al. (2001). Health information on the Internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association, 285*(20), 2612–2621.

Boyer, C., et al. (1998). The health on the net code of conduct for medical and health Web sites. *Computers in Biology and Medicine, 28*(5), 603–610.

Boyer, C., Appel, R. D., Griesser, V., Scherrer, J. R., et al. (1999). Internet for physicians: A tool for today and tomorrow. *Revue Medicale de la Suisse Romande, 119*(2), 134–144.

Boyer, C., Selby, M., Scherrer, J. R., Appel, R. D. Brelstaff, G., Moehrs, S., et al. (2001). Internet patient records: New techniques. *Journal of Medical Internet Research, 3*(1), E8.

Cai, J., Johnson, S., & Hripcsak, G., et al. (2000). Generic data modeling for home telemonitoring of chronically ill patients. *Proceedings of the AMIA Symposium,* (pp. 116–120). Bethesda, MD: American Medical Informatics Association.

Carnall, D. (1999). Missouri fines internet pharmacy. *British Medical Journal, 319,* 1324.

Chadwick, D. W., Crook, P. J., Young, A. J., McDowell, D. M., Dornan, T. L., New, J. P., et al. (2000). Using the Internet to access confidential patient records: A case study. *British Medical Journal, 321*(7261), 612–614.

Chen, L. E., Minkes, R. K., Langer, J. C., et al. (2000). Pediatric surgery on the Internet: Is the truth out there? *Journal of Pediatric Surgery, 35*(8), 1179–1182.

Cline, R. J., & Haynes, K. M. (2001). Consumer health information seeking on the Internet: The state of the art. *Health Education Research, 16*(6), 671–692.

Craigie, M., Loader, B., Burrows, R., Muncer, S., et al. (2002). Reliability of health information on the Internet: An examination of experts' ratings. *Journal of Medical Internet Research, 4*(1), E2.

Davis, T. C., Crouch, M. A., Wills, G., Miller, S., Abdehou, D. M., et al. (1990). The gap between patient reading comprehension and the readability of patient education materials. *The Journal of Family Practice, 31*(5), 533–538.

Department of Health and Human Services. (2000). Health insurance reform: Announcement of designated standard maintenance organizations. Retrieved October 2, 2002, from http://aspe.hhs.gov/admnsimp/final/dsmo.htm

Disease Management Advisor (2000). Use your data to efficiently monitor chronic illnesses. *Disease Management Advisor, 6*(12), 181, 188–191.

Disease Management Advisor (2001a). Pioneering organizations expect financial and clinical benefits from "online encounters." *Disease Management Advisor,* 7(4), 49, 60–63.

Disease Management Advisor (2001b). Solid outcomes show e-health and chronically ill senior populations are compatible. *Disease Management Advisor, 7*(7), 97, 103–106.

Dreyer, K. J., Mehta, A., Thrall, J., et al. (2000). Can radiologic images be incorporated into the electronic patient record? *Journal of Digital Imaging, 13*(2 Suppl. 1), 138–141.

Epocrates (2002). Palm and epocrates team to deliver custom handheld solution for healthcare professionals. Retrieved April 20,2002, from http://www.epocrates.com/headlines/story.cfmstory = 10013

Eysenbach, G. (1999). Online prescribing of sildanefil (Viagra) on the World Wide Web. *Journal of Medical Internet Research, 1*(2), E10.

Eysenbach, G. (2000). Towards ethical guidelines for dealing with unsolicited patient emails and giving teleadvice in the absence of a pre-existing patient–physician relationship systematic review and expert survey. *Journal of Medical Internet Research, 2*(1), E1.

Eysenbach, G., & Diepgen, T. L. (1998). Responses to unsolicited patient e-mail requests for medical advice on the World Wide Web. *Journal of the American Medical Association, 280*(15), 1333–1335.

Eysenbach, G., Sa, E. R., & Diepgen, T. L., et al. (1999). Shopping around the Internet today and tomorrow: Towards the millennium of cybermedicine. *British Medical Journal, 319,* 1294–1296.

Feldman, M. D. (2000). Munchausen by Internet: Detecting factitious illness and crisis on the Internet. *Southern Medical Journal, 93*(7), 669–672.

Forsstrom, J. J., & Rigby, M. (1999). Considerations on the quality of medical software and information services. *International Journal of Medical Informatics, 56*(1–3), 169–176.

Gagliardi, A., & Jadad, A. R. (2002). Examination of instruments used to rate quality of health information on the Internet: Chronicle of a voyage with an unclear destination. *British Medical Journal, 324*(7337), 569–573.

Grimson, J. (2001). Delivering the electronic healthcare record for the 21st century.*International Journal of Medical Informatics, 64*(2–3), 111–127.

Health Level Seven. (2002). *About HL7.* Retrieved October 4, 2002, from http://www.hl7.org/about/hl7about.htm

Hellawell, G. O., Turner, K. J., Le Monnier, K. J., Brewster, S. F., et al. (2000). Urology and the Internet: An evaluation of Internet use by urology patients and of information available on urological topics. *British Journal of Urology International, 86*(3), 191–194.

Hochhauser, M. (1999). Informed consent and patient's rights documents: A right, a rite, or a rewrite? *Ethics & Behavior, 9*(1), 1–20.

Hopper, K. D., TenHave, T. R., Tully, D. A., Hall, T. E., et al. (1998). The readability of currently used surgical/procedure consent forms in the United States. *Surgery, 123*(5), 496–503.

Hubbs, P. R., Rindfleisch, T. C., Godin, P., Melmon, K. L., et al. (1998). Medical information on the Internet. *Journal of the American Medical Association, 280*(15), 1363.

Jadad, A. R., & Gagliardi, A. (1998). Rating health information on the Internet: navigating to knowledge or to Babel? *Journal of the American Medical Association, 279*(8), 611–614.

Jones, M. J. (2001). Internet-based prescription of sildenafil: A 2104-patient series. *Journal of Medical Internet Research, 3*(1), E2.

Kahan, S. E., Seftel, A. D., Resnick, M. I., et al. (2000). Sildenafil and the Internet. *Journal of Urology, 163*(3), 919–923.

Kane, B., & Sands, D. Z. (1998). Guidelines for the clinical use of electronic mail with patients. The AMIA Internet Working Group, Task Force on guidelines for the use of clinic–patient electronic mail. *Journal of the American Medical Informatics Association, 5*(1), 104–111.

Kisely, S. (2002). Treatments for chronic fatigue syndrome and the Internet: A systematic survey of what your patients are reading. *The Australian and New Zealand Journal of Psychiatry, 36*(2), 240–245.

Kleiner, K. D., Akers, R., Burke, B. L., Werner, E. J., et al. (2002). Parent and physician attitudes regarding electronic communication in pediatric practices. *Pediatrics, 109*(5), 740–744.

Lacher, D., Nelson, E., Bylsma, W., Spena, R., et al. (2000). Computer use and needs of internists: A survey of members of the American College of Physicians–American Society of Internal Medicine. In *Proceedings of the*

AMIA Symposium 2000 (pp. 453–456). *Bethesda, MD: American Medical Informatics Association.*

Levin, A. (2001). The Cochrane collaboration. *Annals of Internal Medicine, 135*(4), 309–312.

Libertiny, G., et al. (2000). Varicose veins on the Internet. *European Journal of Vascular and Endovascular Surgery, 20*(4), 386–389.

Lorig, K. R., et al. (2002). Can a back pain e-mail discussion group improve health status and lower health care costs? A randomized study. *Archives of Internal Medicine, 162*(7), 792–796.

Mabrey, J. D. (2000). Survey of patient-oriented total hip replacement information on the World Wide Web. *Clinical Orthopedics,* (381), 106–113.

Mashiach, R., Seidman, G. I., Seidman, D. S., et al. (2002). Use of mifepristone as an example of conflicting and misleading medical information on the Internet. *British Journal of Obstetrics Gynecology, 109*(4), 437–442.

McKinley, J., Cattermole, H., Oliver, C. W., et al. (1999). The quality of surgical information on the Internet. *Journal of the Royal College of Surgeons of Edinburgh, 44*(4), 265–268.

McLellen, F. (1998). Like hunger, like thirst: Patients, journals and the Internet. *Lancet, 352* (Suppl. II), 39–43S.

McLendon, K. (2000). E-commerce and HIM: Ready or not, here it comes. *Journal of American Health Management Association, 71*(1), 22–23.

Murero, M., D'Ancona, G., Karamanoukian, H., et al. (2001). Use of the Internet by patients before and after cardiac surgery: Telephone survey. *Journal of Medical Internet Research.*

National Health Service. (1998). NHS executive. Information for health, an information strategy for the modern NHS 1998–2005. Retrieved October 1,2002, from http://www.nhsia.nhs.uk/strategy/full/contents.htm

Oliver, C. (1999). Automatic replies can be sent to unsolicited email from general public. *British Medical Journal, 319*(7222), 1433.

Ovid Technologies. (2002). The history of Ovid Technologies. Retrieved May 3,2002, from http://www.ovid.com/company/history.htm

Oyston, J. (2000). Anesthesiologists' responses to an email request for advice from an unknown patient. *Journal of Medical Internet Research, 2*(3), E16.

Pandolfini, C., Impicciatore, P., Bonati, M., et al. (2000). Parents on the Web: Risks for quality management of cough in children. *Pediatrics, 105*(1), e1.

Rice, B. (2001). Online prescribing. The growing problem of online pharmacies. *Medical Economics, 78*(11), 40, 42, 45.

Safran, C., & Goldberg, H. (2000). Electronic patient records and the impact of the Internet. *International Journal of Medical Informatics, 60*(2), 77–83.

Shon, J., Marshall, J., Musen, M. A., et al. (2000). The impact of displayed awards on the credibility and retention of Web site information. In *Proceedings of the AMIA Symposium* (pp. 794–798). *Bethesda, MD: American Medical Informatics Association.*

Sing, A., Salzman, J., Sing, D., et al. (2001). Problems and risks of unsolicited e-mails in patient–physician encounters in travel medicine settings. *Journal of Travel Medicine, 8*(3), 109–112.

Smart Computing. (2001). Technology news and notes. *Smart Computing, 112*(6), 7.

Smithline, N., & Christenson, E. (2001). Physicians and the Internet: Understanding where we are and where we are going. *Journal of Ambulatory Care Management, 24*(4), 39–53.

Soot, L. C., Moneta, G. L., Edwards, J. M., et al. (1999). Vascular surgery and the Internet: A poor source of patient-oriented information. *Journal of Vascular Surgery, 30*(1), 84–91.

Tatsumi, H., Mitani, H., Haruki, Y., Ogushi, Y., et al. (2001). Internet medical usage in Japan: Current situation and issues. *Journal of Medical Internet Research, 3*(1), E12.

Taylor, H. (1999, August 5). *Cyberchondriacs: The Harris poll.* Retrieved June 13, 2003, from http://www.harrisinteractive.com/harris_poll/index.asp?PID=117

Taylor, H. (2001, April 18). *Cyberchondriacs update: The Harris poll 19.* Retrieved June 13, 2003, from http://www.harrisinteractive.com/harris_poll/index.asp?PID=229

The Quality Letter for Healthcare Leaders. (2001). Information on the Internet found to be usually accurate but also incomplete. *The Quality Letter for Healthcare Leaders, 13*(7), 1, 11–12.

The Yankee Group. (2002). Cable modem providers continue to lead the high speed Internet charge: Yankee group predictions on consumer broadband service. Retrieved May 5, 2002, from http://www.yankeegroup.com/custom/research/report_overview.jsp?D = 4121

Thexton, A. (2002). Internet prescribing. *World Medical Journal, 101*(1), 28–30.

Ullrich, P. F., Jr., & Vaccaro, A. R. (2002). Patient education on the Internet: Opportunities and pitfalls. *Spine, 27*(7), E185–E188.

Winker, M. A., Flanagin, A., Chi-Lum, B., White, J., Andrews, K., Kennett, R. L., et al. (2000). Guidelines for medical and health information sites on the Internet: Principles governing AMA Web sites. American Medical Association. *Journal of the American Medical Association, 283*(12), 1600–1606.

Middleware

Robert Simon, *George Mason University*

INTRODUCTION

Middleware is a software layer that allows multiple processes running on multiple machines to interact across a network (Bernstein, 1996; Vinoski, 2002). Although middleware can run on stand-alone computers that are not connected to a network, the subject of middleware is interesting because it represents an extremely powerful way to develop fully distributed software systems and applications. As distributed systems evolve into essential information utilities with global reach, the role of middleware becomes even more important, because it is the software glue that allows business, government, and personal systems to communicate and interoperate. In the context of this chapter, the word "interoperate" means that two or more computer systems running different software or operating system code can exchange information or applications.

The development of large-scale middleware systems was motivated by many technological and economic factors. I identify three here: the growth of client–server computing (Adler, 1995), the need for enterprise-wide integrated computing solutions (Hasselbring, 2000; Nezlek, Jain, & Nazareth, 1999) and the birth of e-commerce (Bolin, 1998; Storey, Straub, Stewart, & Welke, 2000). This chapter describes how these developments motivated the design and introduction of increasingly powerful middleware models and implementations.

Although the term client–server computing initially referred to the use of personal computers (PCs) on a local area network (LAN), a more general model emerged during the 1980s. This development was spurred by the migration of mainframe applications to a more flexible and cost-effective client–server environment. A client is a computer or software process that requests services, and a server is a computer or software processes that provides those services. An example of a service is a database server, a file sharing system, or a Web server. The client–server model represented a break from the older mainframe model, where dumb terminals directly connected to intelligent central servers or mainframes.

Client–server architectures require the use of heterogeneous hardware, network, and operating system platforms. Application programmers do not want to rewrite their code for each type of platform. A middleware layer hides the complexity of the underlying platform by providing a standard programming interface and abstraction of platform-dependent system services. Examples of a system service include *open* a file, *close* a file, *read* data from a file, *write* data into a file, and so on (Stallings, 1998). The ability to access a common set of system services allows programmers to focus more of their attention on application development. These concepts have also led to the development of network operating systems (NOS), which allow users to access remote OS-level resources directly, and distributed operating systems, which provide the same level of functionality in a fully transparent way (Orfali, Harkey, & Edwards, 1998; Stallings, 1998).

Another factor that led to the development of middleware systems was the need for integrated, enterprise-wide computing (Nezlek et al., 1999). This is the enterprise integration problem. An example of how this problem came into being is from the history of PC-based file and printer sharing. A company or organization could buy several PCs, connect them together with a LAN, and then allow files and printers to be shared electronically, perhaps through a local file server. As the popularity of these systems grew it became desirable for users to access files, printers, and programs on different servers supported by different LANs. In practice, even within a single organization or department, the servers and the supporting LANs ran different hardware or software. Sometimes the LANs themselves were isolated from a wide area network (WAN) that could be used to provide cross LAN connectivity. This made sharing files and resources difficult to do, complex to program, and expensive to develop. To address these problems of heterogeneity, developers needed software to solve interoperability and portability issues that was scalable and that was independent from specific network and operating system technologies. Specifically, programs running on one type of system must be able to access programs and data on another type of system in a way that maintains reasonably good performance levels, thus providing the illusion of locality.

One technical response to the enterprise integration problem is to use *application programming interfaces* (API). Each API provides a standard set of programming calls that access functions in a library (Stallings, 1998). APIs have associated with them a set of commonly understood rules and procedures for using the calls in meaningful ways. Because the same API can run on different platforms, it is possible to achieve portability. One example of a successful and frequently used API is the *socket* library

(Stevens, 1999). Socket calls are functions that provide standard interfaces for accessing different transport and network level protocols, including the transmission control protocol (TCP), the user datagram protocol (UDP), and the Internet protocol (IP). The socket API, or one of its descendents, is available on all commonly used operating systems. Notice that protocols and APIs are independent of one another. In other words, the rules for using a protocol like TCP are independent of the specific calls and functions for developing TCP code on a particular machine.

APIs are powerful and commonly used tools for achieving interoperability and portability for specific architectural layers. Furthermore, the existence of standard protocols, such as TCP and IP, greatly simplifies the development of distributed computing in general, at scales beyond that of a single enterprise. Despite this, APIs and standard, layer-specific protocols by themselves do not address many of the problems of software development for enterprise and distributed computing systems. This is because current distributed applications require the use and integration of multiple types of programs and services, such as databases, multimedia players, Web servers, message servers, common graphical user interfaces (GUIs), and so on. What is required is the composition of multiple types of services in ways that are semantically meaningful to the application developer. This is evident when considering a third factor in the evolution of middleware: the explosive growth and importance of e-commerce systems.

E-commerce is the activity of buying and selling goods and services over a network, typically the Internet. A *business-to-consumer* (B2C) transaction occurs when a business allows customers to purchase products or services electronically, using a communication platform such as the World Wide Web. Examples of B2C activities include purchasing books, airline tickets, or gourmet food. There are many advantages to this business model. From the consumers' point of view, B2C offers more choices for product type, product customization, and delivery options and reduces the time and expense for finding products and services. From a business perspective, B2C greatly expands the geographic scope of a target market and eases the difficulty in finding customers.

A *business-to-business* (B2B) transaction occurs when a producer of goods or services communicates electronically with one or more suppliers (Bussler, 2002). Businesses selling to other businesses use the Internet to cut transaction costs and increase efficiency. Examples of B2B activities include electronically managing supply chains for producing items such as computers and automobiles. B2B offers the same potential business advantages as B2C, including more choices for product type, customization, and "just-in-time" delivery, as well as expanded market reach.

As e-commerce matures into a more general "e-business" model, where entire business processes such as marketing, banking, supply chain management, and knowledge management are performed electronically, there is a growing desire to develop fully customizable and intelligent software (Pinker, Seidmann, & Foster, 2002). This would enable customer demand to be translated and reacted to instantly in a customizable and cost-effective fashion. For this to occur, distributed systems must be able to share raw data, knowledge and processing capabilities, and business logic know-how. In some sense, this is the holy grail of distributed systems for e-commerce and has given rise to complex and highly interrelated types of middleware that occupy the so-called middle-tier of a distributed system. The middle-tier is sandwiched between the client layer and the server layer.

The remainder of this chapter first discusses middleware from a distributed system architectural point of view and amplifies on the description given here. I next turn to specific middleware technologies, focusing more on core technical issues rather than either standards or vendor products. This is appropriate both because products change rapidly and because there are a multitude of standards.

MIDDLEWARE ARCHITECTURES

One way to classify and understand middleware technology is to identify the architectural location in which it runs within a distributed system. The push for middleware solutions came about as distributed systems migrated from a mainframe architecture to a file-sharing approach and finally to true client–server systems. From the point of view of supporting a business in a distributed environment, what is required is high-performance, persistent storage, and database functionality (Menascé, Barbará, & Dodge, 2001). The response to this need fueled the development of one of the first widely recognized client–server architectures, the *two-tier model* (Adler, 1995). In this approach, a business could use a commercially available database management system to answer user queries rapidly and directly. The client–server architecture reduced network traffic by providing a query response rather than total file transfer. Some of these earlier systems made the first widespread use of middleware, in the form of remote procedure calls or the structured query language. One drawback to the original client–server architecture is that it was computationally intensive for the clients. One way to address this problem is to offload computational processes to the servers. This led to the development of a two-tier architecture.

Two-tier architectures have two layers, the client layer and the server layer. Usually, however, three distinct components exist within this architecture. The first is the user system interface, which runs at the client. The second is database and file management, which runs at the server. The third component consists of program management, including stored procedures to process data before it is sent to the client, resource monitoring routines, and program initialization and control routines. This third component is split between the client and the server. A two-tier architecture is depicted in Figure 1.

The two-tier approach has definite performance advantages over a simple, client–server pure file-sharing approach. However, certain drawbacks exist as well. For instance, consider the activities of the third component just described. Many of its activities derive from business decisions, such as what type of processing to present, how to monitor activities, etc. These business logic decisions

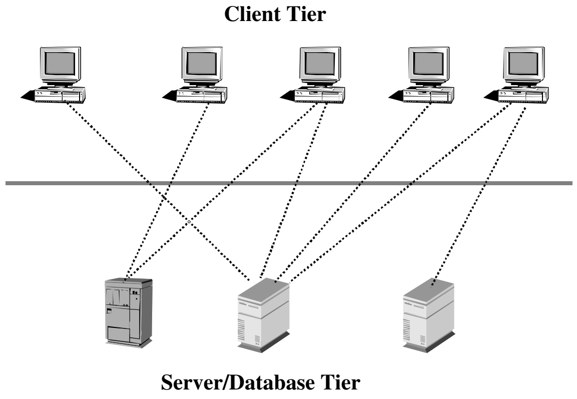


Figure 1: Two-tier architecture.

should be independent of the role of the database and should not be tied to specific vendor policies. This realization led to the introduction of a middle-tier architectural approach, which is currently the environment toward which most middleware architectures are geared. The middle tier itself is sometimes split into multiple tiers, so these architectures are also called N-tier systems (Eckerson, 1995; Helal, Hammer, Zhang, & Khushraj, 2001). In theory, business functions and business logic are placed in the middle tier to improve performance by decoupling processing decisions from either the client or the server, to make these functions reusable, and to make them scalable. Furthermore, administration, management, and changes and code enhancements became easier because system function is placed into a single architectural component, rather than scattered throughout the system. Note that these tiers may be virtual, and it is possible to have several different tiers reside on the same machine.

A frequently seen N-tier system is a Web-based application, consisting of four tiers, as shown in Figure 2. Notice the separation of the Web server functions from the processing logic that is run on a series of application servers. The Web server is responsible for taking requests directly from clients and communicating with the application servers. A typical Web-based processing request is supported by the use of scripting languages such as the *common gateway interface* (Guelich, 2000). A more recent development is to use one of a number of technologies referred to as *dynamic HTML* (hypertext markup language), such as *PHP* (PHP: Hypertext Preprocessor), *Java Server Pages* (JSP), or *Active Server Pages* (ASP), the Microsoft version of PHP and JSP (Tanenbaum, 2003).

Middle-tier activities are typically highly distributed, occur concurrently, and are highly interrelated. The function of middleware within this environment is to simplify management, ensure interoperability between different platforms and even vendors, and achieve high performance. Several types and generations of middleware have been designed to address these needs, including transactional processing (TP) monitors, application servers, and message brokers.

MIDDLEWARE TECHNOLOGY OVERVIEW

As already noted, the fundamental purpose of middleware is to reduce the complexity of developing a distributed system by shielding the application developer from differences in operating system services and network platforms. To be useful, a middleware solution must also provide a set of high-level system services to the developer. This statement is true regardless of whether the architecture of the system is two-tier or N-tier or whether the middleware under consideration is a simple remote procedure call or a complex transaction server, which itself controls many other types of middleware layers and software. Programmers use middleware through APIs, so one way of characterizing a middleware system is by its API and the services and protocols that it supports. From this point of view, one considers the application program itself as the application that uses the middleware system, as illustrated in Figure 3.

Before exploring specific middleware technologies, it is necessary to have a basic understanding of some of the principles of modern software design for distributed

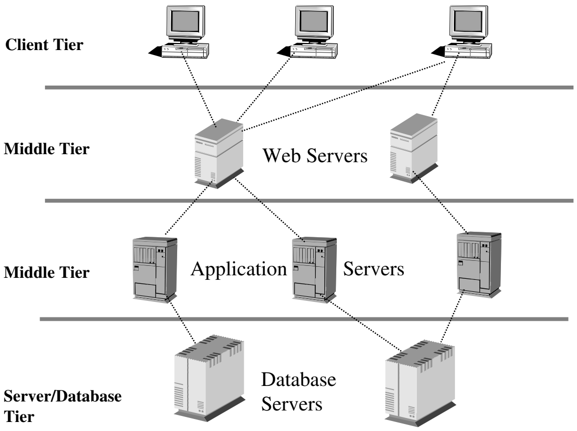


Figure 2: N-tier architecture.

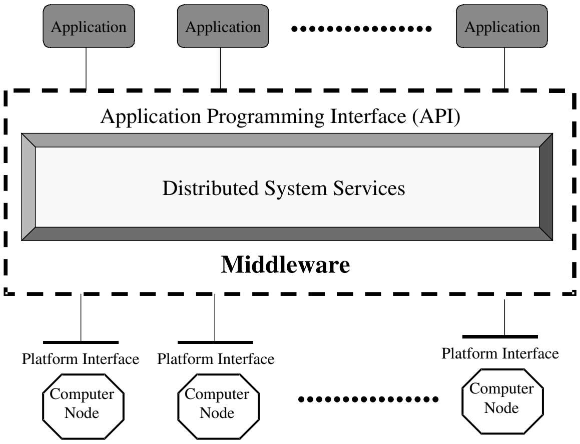


Figure 3: Generic middleware architecture.

middleware systems (Gamma, Helm, Johnson, & Vlissides, 1995; Sommerville, 1998). Software engineers typically use an *object-oriented* approach to developing code and increasingly develop large-scale systems using *component-based* programming (Gomaa, Menascé, & Shin, 2001). Because component-based programming is a direct descendent of the object-oriented approach and incorporates many of the same ideas, I first describe object-oriented programming and then present some of the basics of the component-based approach.

An *object* is a software engineering mechanism used to simplify software development through *encapsulation* and *reuse*. Encapsulation is the ability to hide and change implementation details without affecting working code. Reuse is achieved through techniques such as inheritance. Inheritance allows other programs and programmers to make rapid use of previously coded objects, and to extend their abilities. A *distributed object system* is an object running on one or more machines that can be accessed by programs running on entirely different machines. Object-oriented design and object-oriented programming languages have been one of the cornerstones of modern software development, including the development of object request broker middleware.

In practice, an object is a piece of a program that contains data and that operates on the data. Programmers who use objects don't use them directly but rather make use of *interfaces*. An interface is a programmatic representation of functions (sometimes called methods) that can be performed on the object. For instance, suppose an electronic brokerage house wants to provide a service that allows online traders to buy and sell stock after they register with the firm. The brokerage house specifies a set of interfaces to use with the object, which it names *STOCK_BROKER*. An online trader, or more precisely the application program developer who writes the software to support the online trader, uses the interfaces to access the object. The programmer would add lines of computer code into the program to get access to the object. This code looks like this:

```
SB = STOCK_BROKER.register(Trader_ID)
```

Here, the value of `Trader_ID` is a parameter that can uniquely identify the trader, or the program operating on behalf of the trader, and `SB` is a program variable that can be used to access the object. If the trader wants to execute a buy request for a particular stock called `Stock_name` and wants to pay no more than `max_buy_price` dollars, the code to do this is

```
SB.buy(Stock_name, max_buy_price)
```

Finally, if the trader wants to sell `Stock_name` and wants to get no less than `min_sell_price` dollars, the code to do this is

```
SB.sell(Stock_name, min_sell_price)
```

By making use of the interface code, the application developer need not worry about how the `STOCK_BROKER` object is implemented. Furthermore, the software vendor can change the internal details of the implementation without forcing the online trading software to be modified.

A programming approach based on *components* shifts the emphasis into developing code that can be easily "plugged" into larger software systems. There is also a strong emphasis on the reuse of existing and legacy code. Component pluggability is achieved by encapsulation, although this flavor of encapsulation is potentially much stronger than an object-oriented approach. The implementation of a component is completely hidden, to the point that a component user does not necessarily know in which language the component was programmed. Only the public interface is exposed.

Pluggability, reuse, and extreme encapsulation are achieved in a component environment by the use of *containers*. Components are shielded from a specific runtime environment and are managed within containers. The role of a container is to control the life cycle of a component, to register and deploy the component within an environment, and to provide the component access to other system services such as security. I revisit this technology later, in the context of transaction-oriented middleware.

Broadly speaking, there are four important established technological forms of middleware, with several more in the early stages of development, all of which are based on object-oriented or component-based software engineering principles. The four technologies are *remote procedure call*, *message-oriented middleware* (MOM), *object request brokers* (ORBs), and a class of *transaction-oriented* middleware. I discuss each of these in turn and conclude with some current trends. For each of the following sections, in addition to discussing the core enabling technologies, I also discuss some of the middleware services associated with the particular approach.

Remote Procedure Call (RPC)

The concepts underlying the RPC provide the building block for distributed system programming, and hence for modern middleware (Birrell & Nelson, 1984). Dating from the early 1980s, an RPC enables programmers to access functions on remote systems as easily, and in the same manner, as invoking a function on the local system. RPCs can be used to support client–server computing directly and thus have led to the widespread use of a two-tier architecture. In this case, the server, running on a remote machine, provides the function or service that the client needs.

A programmer uses an RPC as follows: when the program is compiled, the compiler creates a local *stub* for the client portion and another stub for the server portion of the application. (Compiling is the process of translating computer-programming code as written by a programmer into binary machine code that can be run directly on the machine.) A stub is a small piece of code that is used to mediate communication between local and remote machines. Stubs have the necessary logic and functionality to manage network communication and to account for differences among the platforms that the client and the server run on. These stubs are invoked when the application or client requires a remote function.

An important distinction is whether the RPC is *synchronous* or *asynchronous*. A synchronous call is one in which the caller, in this case the client, must wait until the server replies. In this case, I say that the client is "blocked" and cannot do any useful work until the server sends the reply. An asynchronous call is one in which the caller does not have to wait but can go ahead and perform useful work. One drawback to using asynchronous calls is that they can be more difficult to debug. Most RPCs use synchronous communication. By themselves, RPCs represent an older, more primitive technology, one that is not powerful enough to support modern middleware. The basic idea underlying an RPC—the ability to perform a remote invocation—is the basis for all high-level programming in a distributed computing. A descendent of the RPC is the Java Remote Method Invocation (RMI; http://developer.java.sun.com/developer/onlineTraining/rmi/). The Java RMI is used to perform network programming of distributed objects. Object-oriented RPCs such as the RMI are embedded into most middleware systems and form the backbone of distributed object control, even when developers do not use them directly.

Message-Oriented Middleware (MOM)

MOM is a second technological class of middleware (Rao, 1995). As with RPCs, MOMs run on all parts of a client–server system. Unlike RPCs, however, MOMs are designed to support completely asynchronous communication calls. Logically, this is done by the use of a separate *message queue* that provides temporary storage for client or server messages. The message can be the data involved in a specific function call, such as "please multiply these two numbers for me," or can be a higher level message, such as "sell 300 shares of GlobeWorldCo stock at no less than $45 a share."

One advantage to MOMs is that the queuing nature of the system ensures that messages are not lost. Another advantage is that the client is not blocked, even when the destination program, server, or network is busy or entirely unavailable. Within MOM systems clients and servers do not communicate directly, as in the RPC approach. This means that other styles of distributed computing are possible, including *peer-to-peer* computing. In a peer-to-peer system, there is no central server. Rather, computers play the role of both clients and servers.

MOM systems are capable of supporting the needs of large-scale, enterprise-wide computing in a heterogeneous environment, in which different parts of the enterprise need to communicate information in a fully asynchronous way. For instance, MOMs are often used in large-scale N-tier systems to provide communication support between different system components. MOMs are well suited for general purpose environments. Consider the example of an online stock trader. The trader wants to be able to initiate buy and sell orders at any time of the day or night. The trader also wants to be able to place multiple orders at once. Each message that the trader sends could at a minimum contain the trader's ID, the name of the stock, whether it is a buy or a sell order, and the target price. The trader puts as many orders in the form of messages as desired into the message queue. Notice that the trader cannot predict when the order will be fulfilled, or even if the order can be fulfilled. The MOM style supports this type of transaction. Later, a brokerage house can pull the messages off the queue in an entirely asynchronous fashion and send a response message to the trader. Furthermore, it may be possible to have multiple brokerage houses pull off messages.

For building middleware, MOMs provide clear advantages over basic RPCs. There are also several practical drawbacks to using these systems, however. Like RPCs, MOM systems must be run on all computers that participate in the system and therefore consume computing and communicating resources. Furthermore, many MOMs are proprietary, so different vendors messaging systems cannot necessarily interoperate.

Vendors sell MOMs with a rich set of middleware services. It is instructive to examine some of these in detail, because they illustrate the type of services that programmers require in order to build robust distributed systems:

- *Guaranteed message delivery:* Because MOMs operate asynchronously, it may be the case that the sender puts a message in the queue and then disconnects itself from the network before the receiver is even connected. In this case, the message is stored in the message queue. In practice, this means that the message is stored somewhere in the system, in either main memory or persistent storage, such as a disk or a database-like server. When a possible receiver is present then the message is removed from the queue and delivered. Thus, message delivery is guaranteed.
- *Recoverable message service:* Storing messages in persistent storage also means that a MOM can offer a recoverable service, meaning that messages will not be lost in the case of a system failure.
- *Concurrent execution:* Because MOMs are asynchronous, message queues are implemented to allow concurrent execution, meaning that messaging can be processed in parallel.
- *Prioritization:* MOMs also can provide for prioritization. Without a priority service, messages are put in the queue in a "first-in, first-out" order, so when a receiver retrieves the next message, it pulls out the message that was placed in the queue before any of the other messages currently in the queue. By allowing a priority service, messages within the queue are ordered by priority, rather than the time they were placed inside the queue. For instance, the online stock broker may want to place a higher priority on the "sell" orders over the "buy" orders.
- *Notification:* Another useful service is notification. One advantage of a synchronous communication mechanism is that the sender knows when the receiver got the message. A notification service allows the sender to request notification explicitly when the receiver got the message, while maintaining asynchronous calling semantics.
- *Journaling:* This is an auditing service that allows a MOM to keep track of all messages sent, perhaps by making a copy. Journaling helps for different logging and auditing functions.

Although MOMs are used directly to build complex e-commerce applications, they are increasingly seen as one component within an integrated business processing system.

Object Request Brokers

Object request brokers (ORBs) are a type of middleware that manages communicating distributed object systems (Henning & Vinoski, 1999). The ORBs initial development represents one of the first truly comprehensive attempts to apply software engineering principles to distributed system middleware. They support interoperability and portability because they enable programmers to build distributed systems by putting together objects through the ORB. This connectivity is possible even if the objects run on different platforms or are made by different vendors. When programmers use objects through a well-designed ORB they only need care about the details of the *object interface* of the objects in question. This form of encapsulation reduces the complexity of system development and greatly increases system maintainability and extensibility, because object implementation and object communication details are hidden from programmers.

There are several basic functions that all ORBs provide (Henning & Vinoski, 1999). One is the ability to express, easily and in a standardized fashion, the interface for a particular object. Another is advertising the existence of a new object and to find the object regardless of where it is in the system. This is known as location transparency. A third critical function is to invoke an object remotely by sending it commands, and possibly data, and to have a result returned. Another function of the ORB is to enable communication across and through different network and platform domains. Finally, to be useful, all ORBs should be able to add middleware services.

To support this functionality, ORBs have several basic architectural features. Similar to a RPC, one feature is the need to have object *proxies* or stubs. Objects that are used through an ORB need well-defined interfaces. Specific interface details, such as the type of functions the object can offer, the data types the object expects and can deliver, and the type of network communication that the object can support, are encapsulated in a separate piece of code called a proxy or a stub. There is normally one stub associated with each object, although the terminology is not standardized. For instance, one of the most important standards and technology bodies in the distributed object world is called the Object Management Group (http://www.omg.org). OMG is responsible for the *Common Object Oriented Request Broker Architecture* (CORBA) standard, one of the best known ORBs. In the CORBA world, the stub associated with a client object is called a stub, and the stub or proxy associated with a server object is called a skeleton. In COM and COM+, which are Microsoft's ORB products, the client side stub is called a proxy, and the server side stub is called a skeleton (Microsoft COM Homepage, n.d.; Pritchard, 2000). For the purposes of this chapter, I simply refer to these entities as "stubs," regardless of whether they are on the client or the server side.

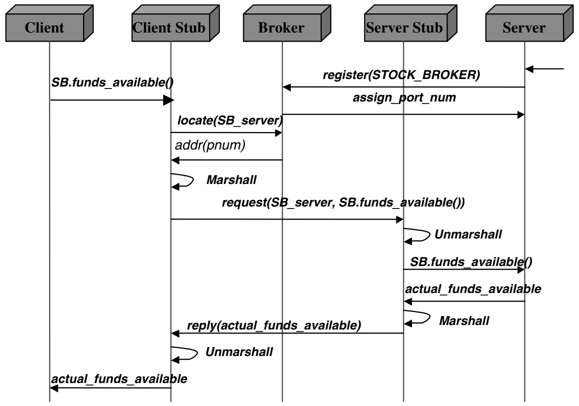


Figure 4: Object request brokers interaction diagram. See the text for an explanation.

Stubs also are the system elements response for all communication to the ORB and to the other objects. As such, client and server objects do not communicate directly with each other, but with their local stub. The core of an ORB is a *broker* object, which functions as a kind of switching device, connecting objects to other objects. Stubs talk to the broker before they talk to each other.

ORB Usage

To illustrate an ORBs basic functioning, I describe the steps of a client communicating with a server. Assume that the client object represents an online stock trader, and the server object is an electronic trading company. The program calls are similar to the ones shown for the `STOCK_BROKER` object described earlier in Middleware Technology. The first step is for developers to program the client and server side of the `STOCK_BROKER` object. This entails writing and compiling code for the functions of the object itself, as well as the interface and communication code contained in the respective stubs. Usually most, if not all, of the stub code is automatically generated by a compiler.

Next come the steps necessary for the client to invoke remote actions on the server. ORBs require that applications know the location (i.e., the network address) of the broker. Broker addresses can either be statically known to a particular system or dynamically determined. The advantage of this is that by simply knowing where the broker is, it is not necessary to have address information for all objects within the system, because the broker can supply them. The sequence of actions is illustrated in Figure 4. To understand this diagram, note that the flow of time goes from top to bottom. There are five objects shown, the client, the client stub, the broker, the server, and the server stub. Events are labeled by the name of the event and the direction of the event. Furthermore, events can happen concurrently.

On the server side, the first action is to start up the server. This can be done in response to a client request, by a system administrator, or through some other mechanism. Continuing on the server side, the next step is to call register(STOCK_BROKER) with the ORB broker. (The ORB broker should not be confused with the STOCK_BROKER object). As the name implies, this function registers information about the server and the STOCK_BROKER object with the broker. In response, the broker sends a assign_port_number(pnum) message. The purpose of this message is to tell the server what transport level address to listen to for remote requests from clients to invoke the STOCK_BROKER object. The networking software delivers information to a specific program or process on a specific machine by using a transport level address. Once the STOCK_BROKER object has been registered, the server can start to satisfy remote requests.

Starting at the top side of the client portion of the diagram, we assume that the application program needs to invoke a function or method on the STOCK_BROKER object. For simplicity, we'll also assume that the STOCK_BROKER object has been initialized and has some funds available for the trader. Suppose that on the client side the programming variable used to access the object is given by SB and that the name of the function from the STOCK_BROKER object that returns the amount of money available for the trader is funds_available. Then the client program uses the computer code SB.funds_available(). This function call is actually sent to the client stub.

The first thing the client stub does is contact the ORB broker to obtain the address of the server. This is shown as the locate(SB_server) call. The ORB broker returns to the stub this address, shown in the addr(pnum) call. Following that, the stub takes the contents of the function call SB.funds_available() and puts it into a format that can be understood by the server stub. This activity is called *marshaling.* After this is done the client stub calls request(SB_server, SB.funds_available()), which sends the function call to the server stub. The server stub unpacks the parameters in the call, and puts the request into a format that the server can understand. This activity is called *unmarshaling.* The server stub then calls the function SB.funds_available() on the server object. After the result is calculated, the server returns the answer actual_funds_available to the stub. The server stub then produces a reply to the client stub by marshaling the result and then making a reply(actual_funds_available) call to the client stub. When the client stub gets this, it unmarshals the result and returns the answer actual_funds_available to the client. The major advantage of the complexity shown in this example is that most of these intermediate steps are hidden from the application developer. Furthermore, using some ORB technologies it is possible to discover and use object interfaces dynamically, thus reducing the need for the programmer of the client side to have complete a priori knowledge of server side capabilities.

When a vendor provides an ORB with this type of functionality, it is typically necessary to provide other services as well. For instance, the CORBA standard describes different messaging models to support synchronous and asynchronous communication. Within CORBA, the *general inter-orb protocol* (GIOP) supports interoperation between ORBs from two vendors (Henning & Vinoski, 1999). By following the GIOP standards, ORBs from different manufactures can (at least in principle) have "wire-line" interoperability, meaning that messages sent between ORBs do not required special translation but rather can be sent directly between the systems.

It is instructive to examine the GIOP briefly because it demonstrates some of the features necessary to make middleware interoperate on the wire level. The GIOP has three basic elements. The first is the *common data representation*, which defines a way to go from one low-level machine specific interface definition to another. The second element is a set of standard message formats, including request, reply, cancel_request, message_error, and several others. Finally, there is a message transport standard, to account for different network and transport layer types. GIOPs themselves are often specialized for specific networking environments. For instance, the *Internet IOP* message transport protocol is a specific instance of a GIOP designed to operate in the TCP/IP environment, one of the most common combinations in the current Internet.

The previous discussion on ORBs focused exclusively on remote object invocation and control. To be useful in practice, however, all middleware systems must have additional services. As in the discussion of the services available for MOMs, one can list some of the commonly available ORB services. For concreteness, I describe some of the services in the CORBA standard.

- *Naming Service:* This is where clients can get object references. In this way, a naming service is like a file system, associating files with filenames. A name binding always associates a name with an object. This is similar to a white pages telephone directory.
- *Trader Service:* This service provides a more dynamic mechanism for obtaining object references then the naming service. In a trader service, rather than storing a name for each object we store a service description. This is similar to a yellow pages telephone directory.
- *Event Service:* This service is similar to the message queue in a MOM. It decouples senders and receivers and allows asynchronous communication and notification to occur.

Component-Based ORBs

A successful example of the component approach to ORB development is Microsoft's Component Object Model (COM), Component Object Model+ (COM+), and Distributed COM (DCOM). COM, COM+, and DCOM (which is designed to support network-based component interaction) are essentially middleware technologies supporting component interaction (Limprecht, 1997; Microsoft COM Homepage, n.d.). The COM environment defines the framework for building programs that are run within the Microsoft Windows environment.

Program objects running on Microsoft operating systems can be written as COM objects and can then have their interfaces available to other programs. For instance, the popular spreadsheet program Excel is written as a COM object and can therefore make many of its functions accessible to other programs. One way to make use of this function is through a process called "Automation." This allows external programs to control application functions, by, for instance, performing tasks that are normally controlled by graphical user interface–menu choices to be automatically and externally executed. An example of this is to write a program that would open up an Excel spreadsheet, read in an Excel file, sum a column of numbers, and return the result.

Transaction-Oriented Middleware

Transaction-oriented middleware encompasses several generations of middleware that are quite distinct from each other. They can be labeled "transaction-oriented" because they are often used to support the activities of large business enterprises that must integrate multiple software systems (Liebig & Tai, 2001). The word "transaction" arises from the database world (Silberschatz, Korth, & Sudarshan, 2002). An example of a transaction is transferring money between two checking accounts. Success in this case means that the money is subtracted from one account and added into the other account in such a way that no matter what type of temporary system failures occur (e.g., the database crashing, the network going down, etc.) the state of both accounts is consistent. For instance, a system failure in this scenario could be that the system crashed after the money was subtracted from the first account but before it was added to the second account. Transaction-oriented middleware prevents this from happening.

Formally, a transaction defines a set of events that are guaranteed to be *committed* or *rolled back* as a unit. Here the word "committed" means that all of the events within the transaction are executed, and any system state that needs to be changed is changed. A rollback means that none of the events are in effect executed, and no part of the system state is changed. For transaction-oriented middleware to work, it is necessary to define the beginning of the transaction, all of the events within that transaction and the end of the transaction. This can be done in a number of ways. One common way is for a client to call a *transaction coordinator* to start the transaction. The client then makes all of the necessary events happen. When this is done, the client tells the transaction coordinator that the transaction is completed. It is then up to the coordinator either to commit the transaction or to rollback the transaction.

Transaction-oriented middleware typically provides several services. One is to guarantee the integrity of a particular transaction, that is, to make sure the transaction either completely succeeds or has no effect on the system. This is generally done by keeping backup copies of earlier states of different system elements, such as database files or Web servers, and not releasing them until the transaction has been successfully reflected the system. If the transaction fails, then the appropriate system elements are restored to their original state. For instance, the state of the checking account before money is added to it is simply the amount of money before the transaction began. In effect, transactions define atomic units of work and the boundaries between those units.

One type of transaction-oriented middleware is the *transaction processing* (TP) monitor. The main function of a TP monitor is to coordinate the flow of user requests into the system by managing the applications that handle those requests in a transactionally secure way and to provide a location for application logic. This involves the coordination of a number of system functions, including database management, distributed communication, system level message queuing mechanisms, user interface management, and so on.

Like all other middleware environments, TP monitors provide a single uniform API to access their services. For instance, most TP monitors allow applications to start or finish a transaction. Other types of services available through a TP monitor API include using a system-level message queue for communication in a way that the message is considered part of the transaction.

TP monitors are designed to support large numbers of users in a large-scale environment and are capable of supporting hundreds or thousands of users simultaneously. As such, they typically are heavily optimized to perform extremely well, meaning that the average response time for a particular user does not change noticeably even when there are many users. A TP monitor accomplishes this by careful system resource management and allocation. One technique is to perform *load balancing,* so that multiple processors can run the same application. When a new user request comes in the TP monitor can assign that request to a processor that is lightly loaded. Another technique is to assign different priorities to different user requests. For instance, a request that involves repeated user input receives a higher priority in the system. This means that interactive users have better system response then users that are not expecting an immediate system response. Vendor-offerings of TP monitors include IBM's CICS (http://www-3.ibm.com/software/ts/cics/), the Microsoft Transaction Server (http://www.microsoft.com/com/), and BEA Systems TUXEDO (http://www.bea.com/products/tuxedo/index.shtml; Andrade, Carges, Dwyer, & Felts, 1996).

As the importance of middle-tier and Web-based business environments continues to grow, new generations of transaction-oriented middleware called *application servers* are becoming extremely popular (Leander, 2000). Application server software is typically Web enabled and is designed to operate in a component-container fashion in the middle tier. There is no single complete definition for what encompasses an application server; rather, it is useful to understand this middleware environment in terms of both component-container software engineering and enabling middleware platform services.

The idea behind the middleware technology of an application server is that the middle tier can be almost arbitrarily complex, must be flexible enough to respond rapidly to new customer demand, must be able to incorporate new business logic and processing methods rapidly, and must allow true interoperability between different vendor products, legacy code, and platforms. Current software engineering practice suggests that the way to address these needs is through a container-component approach that supports pluggability and simplified integration with a broad set of middleware platform services. There are several vendors and proposed standards that provide this, including Microsoft COM+ and .NET, the OMG CORBA Component Model (CCM), and Sun's Enterprise Java Beans (EJB; Monson-Haefel, 2000) and Java 2

Platform, Enterprise Edition (J2EE; http://java.sun.com/j2ee/overview.html).

I first revisit and elaborate on the component-container model. To be useful in an application server, a typical component requires several sets of interfaces. The first is the public interface, the set of function calls that clients can use to do useful work. The second set of interfaces defines the components life cycle actions, such as creation, deletion, and a finder service. A component also needs the actual binary machine code that implements the components business methods, as expressed by the public interface. Finally, most components have an associated index or *primary key*, so that they can be stored in a central name server or component broker. This description is actually the bare minimum of what components require, and some models provide for a much richer set of publicly available interfaces to allow environmental interaction, such as generic event interfaces.

Containers are used in application servers to plug components into larger systems and to add middleware services such as transaction management, event management, security, and persistence management. For the component model to work, it must be possible for a programmer to develop component code without any knowledge of the actual environment that the component will be used. The way that this can be achieved is through the use of containers, and special *configuration files* that specify the execution needs of a particular component. When a client wants to execute some of the functions of the component, the container intercepts the client call. The container then does some preprocessing by reading the component configuration file. This approach enables vendors to "package" containers for use within application servers. Each package consists of a set of resources that the container needs, the runtime binary code, and the configuration file.

The component-container approach provides the technological core for Application Server middleware platforms. Like the other middleware platforms I have described, application server middleware platforms provide a multitude of necessary middleware services to offer client and middle-tier business logic. For concreteness, I briefly examine the services provided by J2EE.

J2EE defines a platform for developing Java applications on components. It is actually a specification along with a reference implementation. The key technology within J2EE is the Java version of the component-container model, the EJB. J2EE defines a large variety of interfaces for linking and accessing other middleware services. These include the *Java Transaction Service* (JTS) and *Java Transaction API* (JTA) for transaction management, *Java Database Connectivity* (JDBC) for databases, the *Java Naming and Directory Interface* (JNDI) for directory and naming services, and the *Java Messaging Service* (JMS) for message management. It should be pointed out that other vendor application server products offer similar, if not identical, sets of middleware services.

CURRENT TRENDS

I conclude this chapter with a brief discussion of several open issues and current trends in middleware design for distributed systems and e-commerce. One important area is the development of a standard approach for the exchange of business information. Notice that this is a *semantic* question as much as a technology question; in other words, what type of information is important, and how should we interpret the meaning of that information? One possible approach to exchange business information is to use the *Electronic Data Interchange* (EDI) standard (Bussler, 2002). EDI is a set of protocols for conducting highly structured interorganization exchanges, such as describing or buying products. EDI defines a set of messages types, which contain a string of data elements, each of which represents a specific item of business data, such as the price of a product. Although the need for a set of EDI-like functions is clear, it is not certain whether this standard will survive intact.

Another type of middleware in the early stages of development is the *message broker* (Sommer, Gulledge, & Bailey, 2002). Messages brokers are essentially asynchronous, store-and-forward messaging systems, somewhat like MOMs. The major difference is that a message broker is in effect independent of the messaging system and the middleware system used for communication control and remote object invocation. In this way, a message broker functions as "middleware for middleware." Message brokers typically provide message *translation engines*, which allow the message broker to adapt to differences in the ways that applications represent data. Message brokers also generally provide for *intelligent routing*. This type of routing has nothing to do with network level routing. Rather, the message broker has the ability to figure out—intelligently, by using a set of dynamically configurable rules—the correct destination to send a message to. Similar to some of the features in current MOMs, message broker systems also provides message warehousing and message repositories. The repository contains information about message sources and destinations, include owner information, security, and processing rules.

In a message broker system, *adapters* are used to connect applications and middleware elements to the message broker environment. There are currently two styles of adapters. A *thin adapter* uses a simple API to map the interface of the application or component to the services supported by the message broker. A *thick adapter* is more sophisticated; it uses information available in the repository about the targeted system and is capable of providing a significant level of functionality.

An evolving trend in middleware implementation is the idea of *reflective middleware* (Leander, 2000). A system is "reflective" when it is able to manipulate and reason about itself the same way it does about its application domain. Reflective middleware has the ability to reason about its environment and function and then change itself during the course of its lifetime. The advantage of this type of design is that it is possible to load customizable and individualized components dynamically, including application-specific task schedulers, file managers, and so on.

Finally, in a fully distributed and Web-oriented business environment, the role of middleware will continue to provide, as much as possible, seamless integration and support for business functions and interoperability between software and hardware components (World Wide Web Consortium). To provide this level of support, a new set of middleware services, called *Web services*, is being proposed. Following is a very brief overview of this interesting development; see the chapter by Akhil Sahai, Sven Graupner, and Wooyoung Kim in this encyclopedia for additional information. Web service middleware should be able to support the construction of highly agile and composable networked services. Supporting this goal requires several types of technology. One is traditional middleware that can connect applications with services, allow for pluggable components and support legacy systems. Another type of technology is a simple and effective way to encode information and business knowledge. Currently it appears that an attractive way to do this is to use an XML-like structure.

Because the new economy and supporting infrastructure is Web driven, it appears desirable to base middleware communication on Web protocols. There are several efforts underway in this regard, including the *simple object access protocol* (SOAP; Chester, 2001). SOAP is essentially a type of RPC over the Web. It uses the Web transport protocol HTTP, and the messages that it sends are XML (extensible markup language) based. The advantage of this is that most distributed systems, regardless of platform, use HTTP (hypertext transfer protocol). Because XML is completely self-defining and extensible, it should be able to accommodate future standards and technologies.

GLOSSARY

Application programming interface (API) A standard set of programming calls for use in developing applications and protocols.

Application server Transaction-oriented middleware that exists in the middle of an N-tier system.

Business to business (B2B) E-commerce between two businesses.

Business to consumer (B2C) E-commerce between a business and a consumer.

Client-server model A model of distributed computing where local host computers (clients) requests services, programs, or data from remote computers (servers).

COM+ A component-based object request broker made by the Microsoft Corporation.

Common object-oriented request broker architecture (CORBA) One of the most popular and successful object request broker standards.

Component-based programming A software engineering technique placing emphasis on developing programming modules that can be rapidly "plugged" into different larger applications.

Distributed object computing A style of computing that supports programs and application development by using objects that are distributed on computers throughout a network.

E-commerce Buying and selling goods and services over the Internet.

Electronic data interchange (EDI) A set of electronic protocols used for business transactions.

Message-oriented middleware (MOM) A type of middleware that stores data and information in a distributed queue.

Middleware A software layer present in distributed systems designed to simplify application development, provide a common set of services to software engineers, and provide a single programming abstraction for a set of heterogeneous computer and communication platforms.

N-tier A distributed computing model containing a client layer with a user interface, a server layer with data management services, and one or more middle layers, consisting of multiple components required for e-commerce, such as Web servers, business logic functions, and network management functions. The middle layers are also called the middle tier.

Object-oriented programming A software engineering technique that uses abstraction and reuse, via objects, to simplify the development of new programs.

Object request brokers (ORBs) A type of middleware that manages and supports distributed object computing over heterogeneous systems.

Remote procedure call (RPC) A distributed programming technique enabling functions invoked on a local machine to be executed on a remote machine, with the results sent back to the local machine.

Transaction processing (TP) A type of transaction-oriented middleware that monitors the flow of users requests and coordinates a set of system functions, such as database access.

Transaction-oriented middleware A collection of middleware services designed to support complex business applications reliably.

Two-tier model A distributed client-server consisting of a client layer with a user interface, a server layer with database and file services, and a program management layer split between the client and the server.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Client/Server Computing; Consumer-Oriented Electronic Commerce; Web Services.*

REFERENCES

Adler, R. M. (1995). Distributed coordination models for client/sever computing. *Computer, 28,* 14–22.

Andrade, J., Carges, M., Dwyer, T., & Felts, S. (1996). *The Tuexdo sysem.* Boston: Addison Wesley.

BEA Tuxedo Web site (n.d.). Retrieved April 11, 2003, from http://www.bea.com/products/tuxedo/index.shtml

Bernstein, P. A. (1996). Middleware: A model for distributed services. *Communications of the ACM, 39,* 86–97.

Birrell, A. D., & Nelson, B. J. (1984). Implementing remote procedure calls. *ACM Transactions on Computer Systems,* 39–59.

Bolin, S. (1998). E-commerce a market analysis and prognostication. *StandardView, 6,* 97–105.

Bussler, C. (2002). Data management issues in electronic commerce: The role of B2B engines in B2B integration architectures. *ACM SIGMOD Record, 31,* 67–72.

Chester, T. M. (2001). Cross-platform integration with XML and SOAP. *IT Professional, 3,* 26–34.

Eckerson, W. W. (1995). Three tier client/server architecture: Achieving scalability, performance, and efficiency in client server applications. *Open Information Systems, 3*(20), 1–12.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* Boston, MA: Addison-Wesley.

Gomaa, H., Menascé, D. A., & Shin, M. E. (2001). Reusable component interconnection patterns for distributed software architectures. ACM SIGSOFT Software Engineering Notes. In *Proceedings of the 2001 Symposium on Software Reusability: Putting software reuse in context* (pp. 69–77). New York: ACM.

Guelich, S., Gundavaram, S., & Birznieks, G. (2000). *CGI Programming with Perl* (2nd ed.). Cambridge, MA: O'Reilly.

Hasselbring, W. (2000). Information systems integration. *Communications of the ACM, 43*(6), 32–38.

Helal, S., Hammer, J., Zhang, J., & Khushraj, A. (2001). A three-tier architecture for ubiquitous data access. In *ACS/IEEE International Conference on Computer Systems and Applications* (pp. 177–180). Los Alamitos, CA: IEEE Computer Society.

Henning, M., & Vinoski, S. (1999). *Advanced CORBA programming with C++.* Boston: Addision-Wesley.

Leander, R. (2000). *Building application servers.* Cambridge, UK: Cambridge University Press.

Liebig, C., & Tai, S. (2001). Middleware mediated transactions. In *Proceedings of the 3rd International Symposium on Distributed Objects and Applications* (pp. 340–350). Los Alamitos, CA: IEEE Computer Society.

Limprecht, R. (1997). Microsoft transaction server. In *Proceedings of IEEE Compcon '97* (pp. 14–18). Los Alamitos, CA: IEEE Computer Society.

Microsoft COM Homepage (n.d.). Retrieved April 11, 2003, from http://www.microsoft.com/com/

Monson-Haefel, R. (2000). *Enterprise JavaBeans.* Cambridge, MA: O'Reilly.

Nezlek, G. S., Jain, H. K., & Nazareth, D. L. (1999). An integrated approach to enterprise computing architectures. *Communications of the ACM, 42*(11), 82–90.

Orfali, R., Harkey, D., & Edwards, J. (1999). *Client/server survival guide* (3rd ed.). New York: Wiley.

Pinker, E. J., Seidmann, A., & Foster, R. C. (2002). Strategies for transitioning "old economy" firms to e-business. *Communications of the ACM, 45*(5), 76–83.

Pritchard, J. (2000). *COM and CORBA side by side.* Boston, MA: Addison-Wesley.

Rao, B. R. (1995). Making the most of middleware. *Data Communications International 24,* 89–96.

Silberschatz, A., Korth. H., & Sudarshan, S. (2002). *Database system concepts* (4th ed.). Columbus, OH: McGraw-Hill.

Sommer, R., Gulledge, T., & Bailey, D. (2002, March). The n-tier rub technology. *ACM SIGMOD Record, 31*(1), 18–23.

Sommerville, I. (2001). *Software engineering* (6th ed.) Boston, MA: Addison-Wesley.

Stallings, W. (1998). *Operating systems: Internals and design principles* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Stevens, R. (1999). *UNIX network programming, Volume 2: Interprocess Communications* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Storey, Veda C., Straub, D. W., Stewart, K. A., & Welke, R. J. (2000). A conceptual investigation of the e-commerce industry. *Communications of the ACM, 43*(7), 117–123.

Tanenbaum, A. (2003). *Computer networks* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Vinoski, S. (2002). Where is middleware? *IEEE Internet Computing, 6,* 83–85.

World Wide Web Consortium. Retrieved April 11, 2003, from http://www.w3.org

FURTHER READING

Altendorf, E., Hohman, M., & Zabcki, R. (2002). Using J2EE on a large, Web-based project. *IEEE Software, 19,* 81–89.

Ball, M. O., Meng, M., Louiqa, R., & Zhao, Z. (2002). Data management issues in electronic commerce: Supply chain infrastructures. *ACM SIGMOD Record 3*(1), 61–66.

Bernstein, P., & Newcomer, E. (1997). *Principles of transaction processing.* San Francisco: Morgan Kaufman.

Briot, J.-P., Guerraoui, R., & Lohr, K.-P. (1998). Concurrency and distribution in object-oriented programming. *ACM Computing Surveys (CSUR), 30*(3), 291–329.

Gang, X., Gang-Yu, X., Litokorpi, A., & Nyberg, T. R. (2001). Middleware-based solution for enterprise information integration. *Proceedings of the 8th IEEE International Conference on Emerging Technologies and Factory Automation, 2,* 687–690.

Kon, F., Costa, F., Blair, G., & Campbell, R. H. (2002). The case for reflective middleware. *Communications of the ACM, 45*(6), 33–38.

Lewandowski, S. M. (1998). Frameworks for component-based client/server computing. *ACM Computing Surveys (CSUR), 30*(1), 3–27.

Menascé, D. A., Barbará, D., & Dodge, R. (2001). Preserving QoS of e-commerce sites through self-tuning. In *Proceedings of the 3rd ACM conference on Electronic Commerce* (pp. 224–234). New York: ACM.

Slama, D., Garbis, J., & Russell, P. (1999). *Enterprise CORBA.* Upper Saddle River, NJ: Prentice-Hall.

Tai, S., & Rouvellou, I. (2000, April). Strategies for integrating messaging and distributed object transactions. Paper presented at IFIP/ACM International Conference on Distributed Systems Platforms, Hudson River Valley, New York.

Ullman, J., & Widom, J. (1997). *A first course in database systems.* Upper Saddle River, NJ: Prentice-Hall.

Wang, N., Parameswaran, K., Schmidt, D., & Othman, O. (2001). Evaluating meta-programming mechanisms for ORB middleware. *IEEE Communications Magazine, 39,* 102–113.

Mobile Commerce

Mary J. Cronin, *Boston College*

INTRODUCTION

By the end of 2003, over a billion people around the world will own a mobile phone or wireless personal digital assistant (PDA). In many countries, mobile devices outnumber traditional wired telephones, and the growth rate of new wireless users far outpaces that of landline customers. Projections for wireless device ownership by 2005 dwarf the projected growth of personal computer users during the same period according to the International Telecommunications Union. Daily life and work in any metropolitan area anywhere on the globe already reflect the ubiquitous availability of mobile communication. On busy city streets, in crowded commuter trains, coffee shops, business conferences, and college campuses, millions of mobile users are talking, messaging, accessing data, making payments, and educating and entertaining themselves without missing a beat.

It is not surprising that projections about the growth and value of mobile commerce have risen along with the rapid worldwide adoption of wireless communications. Many analysts, including Jupiter, Merrill Lynch, and Ovum, forecast that by 2004 worldwide mobile commerce will generate more than $20 billion in revenues annually. A report published in 2001 by Deutsche Bank Alex, Brown projects that by 2004 the wireless data industry in the United States alone will generate more than $33 billion in revenue. However, these projections are based on the assumption that wireless subscribers will embrace mobile commerce applications on their phones as eagerly as Internet users flocked to the World Wide Web. In fact, this assumption is still very much open to debate.

In most parts of the world, mobile customers still use their phones primarily for talking, not for transactions or data-driven applications. Even the most optimistic surveys report that fewer than half of today's wireless subscribers plan to use their phones for mobile commerce in the coming year. There is even some evidence that subscribers who were early adopters of mobile commerce using WAP (wireless application protocol) consider themselves unlikely to continue doing business on the phone. AT Kearney and the Judge Institute of Management, Cambridge University, conduct an annual survey of mobile users in Asia, Europe, and the United States about their current and planned use of mobile devices. The 2001 survey revealed that the percentage of mobile subscribers who intend to use their WAP-enabled phones for any type of transaction has dropped from 32% in 1999 to only 13% in 2001.

On the other hand, some simple mobile capabilities are generating significant revenue and attracting millions of new users every month. One of these, the very basic ability to generate and transmit plain text messages using mobile devices, has created a boom in SMS (short message service) applications. The number of SMS messages rose from an impressive 4 billion per month at the beginning of 2000 to a massive 25 billion per month by the end of 2001. In Europe, the typical mobile subscriber sends about 40 SMS messages each month according to the GSM Association. The ability to personalize the sound of a mobile phone's ring by downloading a customized ring tone onto the device is another early application with strong subscriber appeal, generating approximately $7 billion in revenue during 2002.

While the surge in SMS and ring tone popularity argue for the most widely available formats as a platform for mobile commerce applications, the profitability of NTT DoCoMo's iMode applications makes a case for highly graphic, content-rich services as a driver for mobile commerce adoption. According to the Telecommunications Carriers Association of Japan, more than 60% of Japan's 70 million wireless subscribers to one or more of these content services, making iMode an attractive model for wireless carriers and content providers in other regions.

The world of mobile commerce today encompasses a largely unmapped landscape between the poles of basic of SMS text messaging and multimedia content applications. Its brief history includes rapid technical advances and some frustrating delays in functionality, optimistic revenue projections and expensive failures. This chapter presents an overview of mobile commerce today, including its technical foundations, its major players and issues, and its still emerging value proposition for consumer and enterprise customers.

Topics in Mobile Commerce

Even the definition of what comprises mobile commerce is still in flux, so the first section of this chapter considers the relationship of mobile commerce to electronic commerce and how best to draw the line between these two modes of doing business. The section concludes with a review of the most common definitions of mobile commerce and a summary of the unique value proposition that mobile commerce offers to individual and enterprise users along with a typology of mobile commerce applications for consumers and enterprise users.

With a clear definition in hand, the next section addresses the technical foundations of mobile commerce and the evolution of wireless networks, including the core international wireless standards and the challenges of moving from second- to third-generation mobile networking speeds and infrastructure. The third section analyzes regional differences in mobile commerce adoption rates and discusses the drivers and barriers to adoption in different countries, including a detailed comparison of Japan and the United States. Because concern about the security of wireless transactions is one of the major barriers to adoption, the section concludes with a brief review of wireless security issues and solutions including encryption, mobile public key infrastructure (PKI) and biometrics. The chapter concludes with a review of the core success factors for mobile commerce growth, a preview of the next stage of wireless technology and mobile commerce, and suggestions for further reading.

DEFINING MOBILE COMMERCE

Before discussing the technical and commercial context of mobile commerce, it is important to be clear about its definition. Some analysts regard mobile commerce as a form of electronic commerce, whereas others see it as a distinct and potentially even larger marketplace. A sampling of the definitions that analysts have proposed highlights this difference of opinion. A MobileInfo.com article titled "What Is M-Commerce?" reports the following frequently used definitions (Abbott, n.d.):

"Business-to-consumer transactions conducted from a mobile device." (J. P. Morgan)

"E-Commerce over mobile devices." (Robinson-Humphreys)

"The use of mobile handheld devices to communicate, interact via an always-on high-speed connection to the Internet." (Forrester)

"The use of wireless technologies to provide convenient personalized and location-based services to your customers, employees and partners." (Mobilocity)

Taking a broader view, an Ernst & Young white paper published in 2001 places m-commerce within the framework of m-business and defines it as a subset of the this larger universe of wireless interactions in the consumer and enterprise markets:

> M-business is defined as the use of mobile technology in the exchange of goods, services, information and knowledge. M-commerce is a subset of m-business and defined as the mobile exchange of transactions. Therefore, m-business includes not only consumer applications but also enterprise solutions that allow companies to operate more efficiently, serve customers better, and expand enterprise connectivity. (Ernst & Young, n.d.)

To synthesize these varied views of mobile commerce into a useful working definition, this chapter recommends a focused and inclusive description for mobile commerce that does not assume a framework of e-commerce: Mobile commerce refers to all data-driven business transactions and exchanges of value by users of mobile devices via wireless telecommunications networks.

Comparison to Electronic Commerce

There are certainly a number of similarities and overlapping characteristics of e-commerce and m-commerce. Both support transactions using a digital connection between remote servers and personal devices to bridge time and distance and both offer different types of applications and value for consumer and enterprise users. In addition, just as e-commerce had to overcome concerns about the security and privacy protection of the Internet before it could achieve a critical mass of adopters, mobile commerce is now faced with questions about the end-to-end security of wireless networks and the potential risks of transmitting sensitive information through unknown intermediaries. Finally, just as e-commerce business models evolved at different speeds and with unique characteristics in different parts of the world despite the global penetration of the Internet, the adoption rate and leading applications of mobile commerce have been influenced by technology environments, network access speeds, mobile device capabilities, and subscriber expectations that vary widely from region to region.

It can be argued that the attempts of early wireless application stakeholders to characterize the mobile commerce environment as a wireless extension of the World Wide Web actually delayed consumer acceptance of mobile commerce. Certainly the hype that preceded the launch of WAP applications in Europe in the late 1990s was followed by widespread criticism of WAP shortcomings, early adopter disillusionment, and numerous failures of WAP-based startups. The design and implementation of WAP, including its core model of "wireless browsing" of online content using a hierarchical menu structure did not match the reality of 9.6-Kbps data transmission speeds, dial-in data connection, and small screen, monochrome, text-only handsets. Although a number of companies undertook to translate the contents of their graphics-rich Web sites into bits of text that could be fed to WAP-enabled phones or to create dedicated mobile portals designed specifically to offer WAP content, mobile subscribers who tried WAP services were not motivated to become regular customers. The consensus of the market was that WAP suffered from the obvious comparison to the higher speed and graphics capabilities of Web content on the Internet and that mass adoption of mobile commerce applications would require both higher data transmission speeds and more sophisticated, graphics-capable mobile devices. Wireless operators lowered their revenue expectations for WAP and set their sights on the launch of high-speed third-generation (3G) networks worldwide. (Issues that have delayed the transition from second-generation [2G] to 3G networks are discussed in detail in the Technology Foundations section.)

Meanwhile, the legacy of the WAP model has had a significant impact on the design of early m-commerce applications. Despite widespread acknowledgment that "wireless browsing" of online content is not the best model for mobile commerce, many of the applications that subscribers can access on their mobile phones today are like miniaturized versions of the services and transactions available on the Internet. Among the applications common to both e-commerce and m-commerce are the following:

- Financial activities such as banking, bill paying, and stock trading;
- Accessing pay-per-view information;
- Subscription services such as stock quotes, news, horoscopes, and weather reports;
- Managing customer information, taking orders, checking inventory, and other sales and customer relationship management functions;
- Accessing pay-per-play games and entertainment;
- Making reservations and purchasing tickets in advance of an event; and
- Participating in an online dating service and searching a database for a compatible personality profile.

Nevertheless, it is an oversimplification to define m-commerce as just a wireless form of electronic commerce. At this early stage of mobile commerce implementation, some application overlap is inevitable. It is easier (and perceived as less risky) for carriers and developers to adopt ideas that have worked on the Internet instead of experimenting with completely new types of services designed specifically for a mobile environment. Assuming that m-commerce will follow the same adoption trajectory and business models as e-commerce, however, is likely to lead to overestimates of short-term m-commerce revenue potential and misguided attempts to port Internet content and applications directly to mobile devices.

There are a number of key differences between the Internet as a development and distribution platform and the wireless world:

- Mobile commerce creates a very different user experience because it takes place on smaller and less capable devices that typically do not display graphics and rich content adequately.
- Conversely, mobile commerce can be more easily personalized to match individual situations, leveraging the ability to track the location of subscribers as they move around and being always available to transmit urgent information or support transactions.
- Instead of the global, open standards that underpin the Internet and Web, mobile commerce standards are fragmented by region and platform.
- Mobile content providers and application developers have less ability than their e-commerce counterparts to distribute products and services directly to customers. Since wireless carriers own the mobile networks and control access to subscribers and billing services, application developers must work within a complex multi-player distribution system to roll out applications and collect revenues from customers.
- Internet high-speed bandwidth and access became less costly as applications and users increased, but establishing high-speed wireless network infrastructure has required billions of dollars in investment by wireless carriers, creating a strong barrier to unmetered or free access to mobile commerce applications.
- E-commerce growth was driven by the rapid adoption of the Web as a killer app, but the killer app for mobile commerce has yet to emerge.

These differences mean that the distribution and business models that resulted in rapid global adoption of Web applications are not likely to be successful in the more constrained wireless environment. On the other hand, the widespread expectations that Web content should always be free have not yet taken hold on the mobile device. According to Jupiter, revenues in 2001 from European customers downloading mobile content such as ring tones, sports scores, and stock prices to phones was 590 million EUR compared with only 252 EUR for online content purchased using a personal computer. And by 2006, Jupiter estimates that Europeans will spend more than 3 billion EUR for mobile content compared with 1.7 billion for online content. This willingness to pay for mobile content and services is a significant advantage for mobile application providers, because it means that a relatively small number of mobile commerce customers can potentially generate revenues earlier in the adoption process than has been possible with Web-based content and services.

Mobile Value Proposition

Even more important than defining mobile commerce is understanding what differentiates it from other modes of doing business. The more applications that take advantage of the unique value provided by mobile transactions, the more motivation there will be for users to integrate mobile commerce into their daily lives.

The most fundamental benefit mobile commerce has to offer is freedom. Freed from the wires and the weight of other computing devices, a mobile subscriber can do business anywhere and everywhere. The device he or she carries is not just portable, it also knows a great deal about his or her whereabouts and habits. It has the potential for using this information to make the user's life more efficient, safer, more entertaining, and even more profitable.

Every mobile exchange of data can be sensitive to the user's current location. This location awareness has value for both enterprise and consumer users. It enables an enterprise mobile commerce application to direct urgent sales or support queries to the field representative who is closest to that particular customer and prompts a travel alert application to notify a long-distance truck driver immediately about the dangerous blizzard bearing down on his route. A location-based mobile dating service can notify subscribers when an attractive prospect with a compatible personality profile is in the vicinity or help locate a nearby romantic restaurant when the first meeting goes well and the two parties decide on impulse to extend their date over dinner.

A mobile device that is always available and always connected is the perfect platform for urgent notifications and impulse purchases. A mobile purse application takes the place of coins at vending machines and parking meters; a pocket-size screen displays and redeems the discount coupons at the grocery store. Using biometrics or mobile PKI, the mobile user has an identity token that can admit him to secure locations on the spot or legally sign contract agreements from afar.

Typology of Mobile Commerce Applications

A billion plus mobile subscribers provide strong evidence that even the most basic wireless communication features are powerful market drivers, but most of those subscribers are not regular mobile commerce users. As network speed increases and mobile devices become capable of more sophisticated services, there is general agreement that the missing ingredient in mobile commerce mass adoption is applications that create a unique mobile value proposition for the customer.

To understand the current and future prospects for mobile commerce, it is helpful to have a framework for considering the mobile value proposition different types of applications have to offer. In creating a useful typology for mobile applications, it is also important to clarify the different objectives and expectations that consumer users and enterprise users will apply to that value proposition. For example, consumers may value a mobile game or location-based service designed to keep them entertained during a long commute on the train. Consumers are also more likely to try out new mobile applications just to see if they are worthwhile. In contrast, enterprise users will typically look for applications that add value to the corporation by improving efficiency or reducing the cost of doing business. Unless there is a strong case for the return on investment of implementing the mobile application, an enterprise user is unlikely to make the effort to try it out.

Table 1 identifies five major categories of consumer-oriented mobile commerce applications. The first, Secure Transactions, includes mobile banking and using the mobile device for credit card purchases as well as using a mobile wallet on the device that serves as a debit cash card. These mobile payment applications were originally expected to account for a large percentage of overall mobile commerce transactions but in fact mobile shopping and banking have not been popular with many consumers. Rather than use the mobile phone as a credit card, subscribers express more interest in the convenience of mobile micropayments when purchasing from vending machines or making other very small expenditures. Mobile gambling, with secure transactions to validate bets and winnings has gained momentum and is now expected to become one of the leading drivers of secure mobile transactions.

A number of consumers experimented with mobile content and signed up with mobile portals during the launch of WAP services, but as noted earlier, this experimentation frequently ended in frustration about slow access speeds and unrewarding content choices, thus motivating subscribers to "pull" content by connecting to mobile portals remains a challenge. There are, however, a number of success stories based on subscribers signing up for content push in the "Infotainment" categories of horoscopes, jokes, and other items that can be pushed to the phone via SMS on a regular basis. To drive SMS traffic and generate messaging revenues, wireless operators often provide such content free of charge to their subscribers.

In the entertainment category, downloading customized ring tones remains one of the most popular and profitable mobile applications, accompanied more recently by downloading mobile screen savers and personalized graphics. Advances in mobile phone graphic and display capabilities have also encouraged an explosion of gaming applications, including mobile versions of games that have won an online and personal computer following and original mobile-only games that are designed to work well within the still constrained mobile phone environment. Like mobile gambling, subscriber adoption of mobile games and entertainment are projected to account for a significant increase in mobile commerce revenue over the next several years.

Location-based services that leverage the tracking capability of the wireless network infrastructure to pinpoint a subscriber's location at any moment provide an opportunity to create unique mobile value. To date, as with mobile shopping, the offer of location-based marketing and discounts from nearby stores has not met with widespread consumer adoption. Nevertheless, as with all the categories on the following table, the numbers of applications and adopters continue to grow worldwide. Instead of waiting for a single "killer application" to drive mobile commerce revenues, operators and developers alike have embraced the model of a value-added portfolio of applications and services that consumers can mix and match to meet their own personal objectives.

With an emerging consensus about how to design and build mobile commerce applications that will meet the differing expectations and goals of consumer and enterprise users, the pervasive adoption and revenue growth of mobile commerce should be assured. There are other barriers that impede the distribution and adoption of value-added applications for mobile users, however. It is important to understand that the implementation of applications and services in the wireless world takes place in a far more constrained technology, standards, and

Table 1 Typology of Consumer-Oriented Mobile-Commerce Applications

MOBILE APPLICATION FEATURE	EXAMPLE APPLICATIONS
Secure Transactions	Banking
	Coupon and discount redemption
	Gambling
	Micropayments
	Mobile shopping
	Mobile ticketing
	Stock trading, auction bidding
Content Push and Pull	Mobile portals
	News, weather, sports, etc
	Financial information
	Horoscopes, jokes
	"Infotainment" content
	Mobile learning
	Guides, directories, maps
	Personalized alerts based on preset profiles
	Permission marketing
Entertainment and Personalization on the Mobile Device	Games
	Rich media (audio, graphics)
	Ring tones
	Screensavers and personalized screen graphics
Interactive Services	Instant messaging
	Multiplayer games
	Mobile self-service or help desk interactions
Location-Based Services	Weather, maps, etc. based on current location of subscriber
	Traffic alerts
	Proximity of known or profiled subscribers to a matching service for dating
	Discount, coupon offers from nearby stores

development environment than the fixed Internet. The next several sections examine each of these factors in more depth.

TECHNOLOGY FOUNDATIONS

To complete a mobile commerce transaction, or even to place a wireless voice call, requires that a complex combination of technologies work together seamlessly. The wireless network infrastructure of cell towers, servers, and subscriber and roaming agreements must closely track the movements of each individual subscriber (or at least his mobile device). It must be prepared to transmit immediately, to receive, to encrypt, and to manage a massive stream of data traffic to and from hundreds of millions of such devices as users move from place to place or even from country to country. To make this task even more complicated, unlike the Internet there is no unified wireless standard that defines the connectivity protocols for every connected device regardless of geographic location. Because the complexity of this wireless infrastructure has a major impact on the adoption of mobile commerce, this section provides an overview of the most important technical foundations for mobile commerce. These components include the wireless networks themselves, the mobile equipment (ME) used by the subscriber, and the subscriber identity module (SIM), a tamper-proof chip inside the ME that connects the equipment to the network by authenticating the subscriber's identity and encrypts the traffic between the device and the network cells.

Wireless Networks

First generation mobile communication depended on analog wireless networks called AMPS (advanced mobile phone system). There are still some millions of subscribers to analog wireless services, particularly in the United States, but this first-generation of networks does not fit the scope of this discussion because it cannot be upgraded to handle digital transactions and therefore does not provide a basis for mobile commerce

The second generation of wireless networks, commonly called 2G, marked a major breakthrough in mobile communications. For the first time, data as well as voice could be accessed via a wireless device. Early devices were large and expensive, and transmission speeds were (and still are) frustratingly slow, but the diffusion of 2G networks worldwide paved the way for mass adoption of mobile devices and introduced the possibility of true mobile commerce.

Table 2 Summary of the Generations of Wireless Networks

	1G	2G	2.5 G	3G
Standards	AMPS (Advanced Mobile Phone Service)	CDMA (code division multiple access) GSM (global system for mobile communications) PDC (personal digital cellular) TDMA (time division multiple access)	GPRS (general packet radio service) EDGE (enhanced data rates for gsm evolution)	W-CDMA (wide-band cdma) (Migration path for GSM and TDMA) CDMA2000 (Migration path for CDMA)
Characteristics	Analog Voice transmission only	Digital, Circuit-switched, dial in connection to data (except for PDC) Voice or data transmission Speeds from 9.6 Kpbs to 14.4 Kpbs	Digital, packet switched, always on connection Speeds up to 384 Kpbs Simultaneous voice and data	Enhanced voice and data quality Packet-switched, Simultaneous voice and data Speeds from 384 K to 2 M per sec Support for multimedia and video applications

The slow data transmission speeds and other limitations intrinsic to 2G, along with the anticipation of billions in revenue from mobile commerce deployment, fueled a drive for faster and more capable wireless networks over the past several years. Just as various wireless carriers and geographic regions selected different paths for the migration from analog to digital transmissions, there is no universally accepted strategy or technology for the next step in wireless evolution, the launch of higher speed 3G networks. In fact, some analysts now question whether full 3G implementation is even necessary for mobile commerce adoption. A technology called general packet radio services (GPRS) provides an evolutionary step based on the more efficient model of packet switching rather than the dedicated circuits used in 2G networks. GPRS and the other technologies that increase the speed of 2G network throughput are often referred to as 2.5-generation networks. Because of speed and throughput limitations even the 2.5 enhancements will not support true wireless multimedia and content rich applications, however. These high-bandwidth applications require an entirely new infrastructure for wireless transmissions that will only become available with the rollout of 3G infrastructure.

Rollout of 3G was expected to coincide with the beginning of the new millennium and to be almost completed by 2003, but the launch of 3G networks has been delayed by a combination of technology and business issues. The cost of the wireless spectrum for 3G skyrocketed when many national governments decided to issue 3G licenses on the basis of competitive bidding among large carriers. At the same time that they took on heavy debt loads to buy up licenses, carriers were faced with unexpectedly expensive network infrastructure and equipment costs for new towers and gateways. As the world economy struggles, the revenue projections that had justified all this expenditure were called into question, forcing many carriers to delay implementation and reduce their debt. Even those carriers who did move ahead with the launch of 3G networks faced frequent problems with the availability and performance of essential components such as 3G capable handsets. As a result, 3G networks are still in early stages of implementation in most regions.

The adoption of mobile commerce around the world is closely related to wireless network performance and speed, so the next section presents a brief technical overview of the largest 2G networks worldwide and discusses the path that each network is taking to 3G. Table 2 summarizes the generations of wireless technology.

2G Networks

GSM and TDMA. With more than 750 million subscribers, GSM (the global system for mobile communication) is the world's largest wireless network infrastructure. GSM carriers have a worldwide agreement that supports roaming among all GSM networks and billing calls from any other GSM carrier. GSM compatible handsets operate on the 900- or 1800-MHz frequency band in most of the world, and at 1900 MHz in the United States. Triband handsets that handle all frequencies are available for customers who travel between regions. The GSM infrastructure includes the SIM, a smart chip that contains the carrier's encryption algorithms and other core network information. A separate section provides more detail about SIM features and functionality.

All European countries adopted the GSM standard to facilitate roaming across national boarders and create economies of scale in the design and manufacturer of wireless infrastructure and equipment in the early 1990s. As a result, European customers were early adopters of wireless technology and mobile commerce applications. GSM coverage and usage is also strong in Asia Pacific. It has been least popular in North and South America, but GSM coverage is increasing in the Americas as TDMA (time division multiple access) operators migrate to higher speed networks such as GPRS that are based on GSM/TDMA technology.

Standard circuit-switched TDMA technology underlies 2G GSM networks. This technology digitizes and compresses data, then sends it down a channel with two other streams of user data, each in its own time slot. TDMA networks are popular in their own right, with more than 100 million TDMA subscribers as of June 2002. TDMA has proven itself in the 2G environment, but it does not ramp up to the higher speed and capacity demands of 3G transmissions. The common basis for 3G networks will be CDMA (code division multiple access), as described in the next section, but CDMA itself is branching into two variants of 3G. The first, known as WCDMA is the migration path selected by most GSM carriers. The second, CDMA2000, is the choice of current 2G CDMA carriers.

CDMA. Code division multiple access is the second most prevalent wireless network standard, with more than 120 million subscribers worldwide. CDMA operates in radio-frequency bands at either 800 MHz or 1.9 GHz and supports data speeds of up to 14.4 Kbps. CDMA was developed by Qualcomm, which still holds many of the underlying patents for this technology. It is a "spread spectrum" technology, which means that it spreads the information contained in a particular signal across a greater bandwidth than the original signal. Because the CDMA technology does not assign wireless data traffic to a specific frequency, every channel uses the complete available spectrum to move data as efficiently as possible.

North and South America have the largest concentration of CDMA subscribers, accounting for more than half of the worldwide subscriber base for 2G CDMA. The United States alone has more than 50 million CDMA users. The Asia Pacific region has also adopted CDMA technology, with Korea and China providing rapid subscriber growth.

Although CDMA is currently far behind GSM in the total number of subscribers worldwide, CDMA2000 technology is leading in the race to 3G network implementation because planned WCDMA rollouts have suffered from numerous setbacks in the past two years. According to January 2003 statistics from the CDMA Web site, the 35 million CDMA2000 subscribers account for more than 98% of the total number of 3G subscribers worldwide.

PDC. With more than 70 million wireless subscribers and a population that has eagerly embraced mobile data applications, Japan is a center of advanced wireless technology and a bright spot in demonstrating the profit potential for mobile commerce. Japan has developed and deployed its own wireless network standards called PDC (personal digital cellular), which are based on TDMA technology. At 9.6 Kbps, the data transmission speeds supported by PDC are no better than 2G GSM and CDMA networks. However, PDC has the advantage of using packet-switched transmission instead of the less data-efficient circuit-switching design of the other 2G networks. Packet-switching is the basis of the familiar Internet protocol (IP) and in the case of PDC, it supports an always-on wireless connection to Web- and Internet-based services and content. This eliminates the need to dial in and wait for a connection, one of the major complaints of 2G customers in other parts of the world who are interested in accessing mobile content and applications.

Japanese carriers such as NTT DoCoMo, KDDI, and J-Phone have taken a proactive role in the deployment of application-friendly mobile handsets and in the distribution of mobile content. The typical Japanese mobile phone is lighter but has a larger display screen than 2G counterparts in other countries. Color screens and support for animated displays are common. According to NTT DoCoMo, more than 75% of subscribers have a phone with built-in browser software to facilitate Internet access. DoCoMo also supports independent content and application developers by providing billing services that cover thousands of third-party applications. All of these factors have contributed to a high adoption rate for mobile commerce in Japan.

2.5G Networks

GPRS and EDGE. Positioned in between full 3G deployment and standard 2G networks are upgrades and enhancements to 2G networks referred to collectively as 2.5G. The most popular of these, GPRS (general packet radio services), is a packet-switched wireless protocol that supports data transmission speeds up to 115 Kbps compared to the 14.4- or 9.6-Kbps connections of 2G. In addition to higher speeds of transmissions, GPRS can handle simultaneous voice and data transmission. It conserves network resources because the higher speeds are only invoked when data is actually transmitted.

With delays and cost considerations plaguing the launch of full 3G networks in many regions, some analysts speculated that GPRS may emerge as more than a transition technology and become a longer term solution for many carriers. More than 100 GSM and TDMA operators in 46 countries have implemented GPRS networks since June 2000 and made them available to subscribers as a higher speed, data-capable alternative. The adoption rate of GPRS in its first two years of commercial availability has been relatively limited, however, with just over 2 million subscribers actively using GPRS networks by June 2002 according to the EMC World Cellular Database.

EDGE (enhanced data rates for GSM evolution) is another half step toward 3G network speed and performance that builds on GPRS and GSM technology. EDGE upgrades key components of the hardware and software of current TDMA infrastructure, including the encoding algorithm used to encrypt data traffic, to achieve speeds of up to 384 kbps with EDGE-compatible handsets. As with GPRS, the 2.5 EDGE solution is attractive to GSM and TDMA carriers who are looking for a way to launch enhanced applications and services without taking on the full cost of a 3G network launch. It remains to be seen whether these 2.5G networks will become a longer term option or will be superseded once full 3G services become more widely available.

3G Networks

W-CDMA and CDMA2000. When the specifications for high-speed, third-generation wireless networks were still on the drawing board, carriers, mobile equipment makers, standards bodies, and other stakeholders talked optimistically about the goal of a single, global standard for third-generation wireless communication. As stakeholders aligned themselves behind competitive technical options, however, and different paths to 3G implementation emerged, it became more and more difficult to develop a uniform migration strategy that all the players were willing to adopt. Although the CDMA spread spectrum technology won out over other alternatives, the GSM and TDMA network carriers were reluctant to start from scratch with a completely new infrastructure. So CDMA itself split into two approaches to 3G implementation, with WCDMA tailored for the needs of GSM and TDMA network upgrades and CDMA2000 tuned to the needs of current CDMA carriers.

Phase One of the implementation of CDMA2000, already underway in a number of countries, supports packet-switched data speeds of up to 144kbps. Phase 2 of CDMA2000, also known as 3XRTT (Radio Transmission Technology) will achieve the long-awaited speeds of 384kbps for a user who is in motion and a top speed of 2Mbps for a stationary user. These Phase 2 speeds will be matched by the WCDMA implementations.

The technical differences between these approaches to 3G relate to the process of dividing the wireless spectrum and sending signals through different radio frequencies, a topic that is not within the scope of this article. From the perspective of mobile commerce adoption, the important question is whether the two standards can somehow be unified to allow wireless subscribers to take advantage of mobile applications anywhere in the world. The section titled Wireless Standards describes the attempts that international standards bodies are making to address this issue.

Drivers and Barriers to 3G Implementation

G networks promise some important benefits to both wireless carriers and mobile customers. For carriers, 3G offers more efficient use of the limited radio spectrum, allowing carriers to process more simultaneous calls and transmit more data at higher speeds while improving voice quality and reducing or eliminating dropped calls. For mobile customers, 3G will support more data intensive and graphic applications as well as multimedia-enabled handsets and, eventually, streaming video transmission.

These improvements, however, come with an extremely high price tag for the carriers. The sheer cost of implementation has become a major barrier to upgrading networks in many regions of the world. In Europe where the 3G radio spectrum was auctioned off in competitive bidding, wireless carriers collectively bid up the licenses to more than $95 billion. Few carriers have had the resources to justify the additional expense of build out, estimated to be more than $125 billion for carriers worldwide, especially in the absence of business models that demonstrate a strong return on their investment. A related barrier to rapid deployment of 3G is networks is doubt about customer demand and willingness to pay for enhanced speeds in the absence of compelling applications. As the success of mobile commerce on 2G networks in Japan and the worldwide popularity of SMS-based transactions demonstrates, many applications do not require full 3G capacity.

Mobile Equipment (ME)

Mobile equipment (ME) features and capabilities have improved in parallel with the advance of wireless network generations. In the AMPS era wireless handsets were large, heavy, expensive and limited in range. 2G digital data transmission and application capability stimulated the development of smaller digital handsets with display screens, caller ID, customizable ring tones, and other features. As handset makers competed for the rapidly growing wireless subscriber market, mobile phones appeared in designer colors and shapes, some with color screens. In many markets, the choice of a wireless phone model became a personal fashion statement and many users upgraded their phones to keep up with the latest fads.

The rapid global expansion of wireless subscriptions fueled high demand for all types of mobile equipment. This growth wave transformed Nokia from a small Scandinavian mobile phone manufacturer to become one of the largest multinational suppliers of handsets and network infrastructure. Nokia's strategy of aggressive innovation and introduction of many new models each year won market share compared with competitors such as Motorola and Ericsson. Some handset technology gambles, however, have not paid off. Customer demand for Internet and WAP-enabled handsets, for example, has been far below the ME manufacturers' optimistic projections.

Asia, on the other hand, has established a well-deserved reputation as the most advanced markets for ME innovation and widespread subscriber adoption of high-tech features. Japanese and Korean manufacturers launched the first widely adopted phones with color screen and graphics capabilities, the first popular phone with integrated digital camera, and, more recently, a full-motion video camera phone. These devices support Internet connections, multimedia applications, and may also run Java applications. As noted earlier, the early availability of full-featured devices in Japan and Korea provides an ideal platform for mobile commerce and has stimulated a faster adoption rate of mobile applications than in other parts of the world, as I discuss in more detail in the section Mobile Commerce Landscape. The introduction of equally advanced devices into the European and American markets may prove to be an important catalyst for mobile commerce adoption there as well.

One area of ME innovation in which the United States has a leading role in the convergence of PDAs and cell phones. Leading PDA manufacturers such as Palm, Compaq, and Handspring have introduced integrated wireless devices in an attempt to expand the market demand for PDAs and to reach new customers. These devices have much larger display screens than typical mobile phones and provide considerably more computing power to handle calendars, contact lists, games, and other applications. Although sales of these converged devices are still tiny compared with the dedicated wireless phone, they may become the device of choice for business-to-business mobile commerce applications. Microsoft's entry into this market, with a reference design for a phone–PDA combination that runs Windows and a series of announcements by carriers and manufacturers planning to build devices on the Microsoft platform indicate that this could become an important growth area.

As ME becomes more capable and wins more brand recognition and customer mind share, wireless carriers are looking for ways to maintain a strong presence in the mind and in front of the eyes of the subscriber. One carrier-owned and controlled component of the ME is a smart chip inside the phone or wireless PDA called the subscriber identity module. Although this component is not highly visible to the customer, it does control the network capabilities of the device and has some important features that support transaction security and mobile commerce applications.

Subscriber Identity Module (SIM)

The SIM was designed as a tiny computing platform to store securely the cryptographic keys that GSM carriers use to authenticate the individual subscriber to the network connection and track subscribers activity once they are on the air. The SIM identifies a particular mobile user to the network in a secure and consistent manner, through a private digital key that is unique to each subscriber and is known only to the wireless carrier. The key is used to encrypt the traffic to and from the handset. Because smart cards were designed to be highly tamper resistant, the smart card's core electronics and design architecture were adopted as the base of the SIM.

The SIM maintains a constant connection to the network as long as the ME remains on. This location-aware, authenticated connection is what allows subscribers to "roam" from network to network around the world. The SIM also provides support for location-based services. In addition, the SIM keeps track of and reports back on the subscriber's network usage and roaming activity so that the carrier can bill customers accurately.

The only way to ensure that the SIM can accomplish its handoff of subscribers from one network to another without interrupting communication is to base all of its functions on detailed international standards. Every GSM carrier adheres to these standards, which cover everything from the physical size and characteristics of the chip to the way it handles and stores incoming information. As carriers move to higher speed 2.5G and 3G networks, the role of the SIM becomes even more important. Because these networks will not all roll out at the same time around the world or even within a particular region, the SIM will manage the roaming of traffic between generations of networks as well as between geographic locations. The features of the SIM are of value to carriers beyond the GSM networks, so the SIM has been adopted in the standards of the CDMA, TDMA, and PDC networks to support 3G implementations and to provide added security for mobile commerce transactions.

Standards and Interoperability

There is no shortage of standards and standard-setting bodies in the wireless world. In fact, it can be argued that problems with network and ME interoperability stem from having too many competing interests claiming a role in establishing different types of wireless standards. As this chapter describes, the 2G network carriers have not managed to agree on a unified global approach to 3G network design and operation. Similarly, ME manufacturers do not embrace a single standard that will ensure that applications and user interfaces behave the same on all handsets. Among other drawbacks of this division of network and ME standards is the need for equipment manufacturers to provide different devices and infrastructure equipment for each network, adding to the expense and creating barriers to interoperability among different parts of the world.

Rather than review the long and complicated history of all the standards related to wireless equipment and transmissions, the following section focuses on the 3G standards development process and the steps that have been taken to improve 3G network interoperability. The International Telecommunication Union (ITU) started the process of defining the standards for third-generation systems by forming the Third Generation Partnership Project (3GPP) in 1998. As it became clear that the CDMA networks would not follow the same path toward 3G as GSM and TDMA carriers, the Third Generation Partnership Project 2 (3GPP2) was formed to carry out a parallel process to the GSM 3G standard setting. 3GPP2 has focused on developing standards and specifications for the technical development of CDMA2000. The top-level working groups within 3GPP and 3GPP2 are charged with communicating with each other and working to coordinate the efforts of both groups to minimize interoperability problems in the various 3G network implementations. Nonetheless, the complexity and proliferation of details in the specifications almost ensures that there will be discrepancies between the two types of 3G implementation strategies and the standards that underpin the implementations. Unfortunately, this will cause ongoing problems in moving mobile commerce applications between networks and regions.

MOBILE COMMERCE LANDSCAPE

Regional Variations

Access to wireless voice services spans the globe, but when it comes to mobile commerce, geography clearly matters. Whereas mobile transactions generate significant revenue in Japan and the majority of wireless subscribers regularly access mobile content and services, the average wireless customers in the United States have barely registered that their phone offers more than basic voice services. In fact, with just 40% wireless penetration rate as of June 2002, mobile phone ownership in the United States lags behind that in Europe and Asia where a large number of countries have penetration rates well over 70%. According to the ITU, the top 10 countries ranked in order of wireless penetration range from Luxembourg at number one, with a 96.7% penetration of wireless, to Portugal at number ten with 77.4% of its population using a wireless phone (International Telecommunications Union, 2002). These numbers put the U.S. wireless market significantly behind Europe and Asia in terms of market penetration and adoption curve for mobile commerce. Many analysts predict that it will take as long as 5 years for the United States to catch up. How did the country that led in Internet connectivity and e-commerce implementation fall so far behind in the wireless arena?

There are a number of factors that contribute to the striking difference in adoption of wireless and mobile commerce in the United States and the rest of the world, including availability of advanced handsets, always-on wireless connections, and a large selection of interesting

Table 3 Mobile Commerce Environment Comparison: Japan and United States

FACTOR	JAPAN	UNITED STATES
Network	PDC with packet-switched, always-on connectivity at 9.6 Kbps	CDMA, TDMA, GSM, and others, all with circuit-switched, dial-in requirement for data access
Mobile Equipment	Features: Advanced, with high-resolution color screens, graphics, and animation capabilities in a small and very light package	Features: Limited, with poor display quality, text only, variable size and weight
	Cost: Moderate to expensive	Cost: Free or low cost with calling plan (subsidized by carriers); otherwise high
	Market: Open, with a large number of independent vendors and multiple models available	Market: Limited, with majority of phones bundled with calling plans and preselected by carrier
Wireless Penetration	High	Moderate
Fixed Internet Penetration	Low	High
Mobile Commerce Business Model	Subscriber billed by carriers for downloaded content and value-added services as well as transactions; carrier retains small commission while content and application developers receive bulk of the revenue generated	Varies from carrier to carrier; billing for third-party applications not typically supported
Range of Applications Available	Extensive, with tens of thousands of wireless content and service providers of all sizes	Limited, with a handful of applications provided by carriers and other large institutions
Barriers to Application Development	Low: Technical difficulty is similar to developing Web applications; standardization around cHTML and iMode ensure interoperability among platforms	High: Requires understanding of device or SIM application interfaces. Multiple networks and protocols make it difficult to move application from one platform to another
Incentives for Application Development	High: Large subscriber base with advanced phones and interest in new services provides critical mass of potential users while revenue sharing business model by carriers provides significant returns for the most popular applications	Low: Few subscribers with appropriate devices or interest in wireless application and no clear model for revenue sharing with independent content and application providers

Note. CDMA = code division multiple access; cHTML = compact hypertext markup language; GSM = global system for mobile communications; PDC = personal digital cellular network; SIM = subscriber identity module; TDMA = Time division multiple access.

mobile applications. It is instructive to compare the wireless environment in Japan to that in the United States on these and other factors. Table 3 summarizes the results.

One positive indicator for the future expansion of U.S. mobile commerce penetration is a survey of the interest of various age groups in using mobile applications such as downloading music, accessing maps and directions, and making purchases at 3G speeds. Although interest levels were 25% or less across all the U.S. age groups surveyed, in the "Under 25" category, 45% of users said they were interested in mobile commerce applications (Taylor Nelson Sofres, 2001). Customer interest is just one driver for mobile applications and mobile commerce, however. Until the U.S. addresses some of the other factors outlined in Table 3, other regions of the world will continue to have the edge in wireless applications and mobile commerce adoption.

In Europe, for example, mobile phones are becoming the primary phone for many subscribers, especially in the important "Under 25" market segment. European countries accounted for 7 out of the top 10 wireless penetration rates in 2001, and despite a lukewarm consumer response to the launch of WAP applications in Europe, mobile commerce activities such as banking, paid mobile content and entertainment, and location-based services are generating increasing customer interest and expenditure.

In Asia, the largest numbers of mobile commerce users and applications are currently in Japan and Korea, whereas the highest wireless penetration rates are in Taiwan at 96.6% and Hong Kong at 84.4% (ITU, 2002). Most eyes are focused on China as an emerging wireless giant with a population large enough to fuel a mobile commerce growth spurt. At the end of June 2002, China had more than 175 million wireless subscribers, making it the largest single mobile market in the world with many more millions of potential customers ready to go wireless over the next several years.

Rapid growth in wireless subscribers is not just a phenomenon of the industrialized countries. Delays and deficiencies in landline availability and traditional telecommunications infrastructure have stimulated high growth in developing economies as well. According to the ITU, 22 of 49 developing countries had more mobile than fixed phone customers by the end of 2001. Wireless customers in these countries are typically interested in the availability of mobile content and mobile commerce, because access to in-person banking and other services is often problematic. Recognizing the advantages that wireless communications can provide to individuals and to entire economies, the ITU's 2002 World Telecommunication Development Report recommends that most low-income economies strive to achieve 90% wireless coverage by 2006.

Security Issues and Solutions

A perception that mobile commerce harbors even more security risks than business on the Internet is a major factor inhibiting the adoption of mobile commerce adoption around the world. Consumers who are accustomed to entrusting their credit card numbers and personal information to online merchants via the Internet still express trepidation about completing a similar purchase transaction on their mobile phone. Corporate network managers who have come to terms with providing desktop Internet connectivity and Web-enabled remote e-mail services draw the line at enabling wireless devices to access internal enterprise data.

Some of these concerns are based simply on resistance to unfamiliar technology and reluctance to risk being an early adopter. Others stem from the memory of successful attacks on first-generation wireless networks and devices, when traffic traveled in the clear and a lack of familiarity with the encryption of all 2G transmissions. Notwithstanding these somewhat misguided perceptions, there are a number of real security issues that mobile commerce programs need to address.

Mobile devices are by definition small, frequently in motion, and a tempting target for thieves because of their intrinsic value as well as the sensitive information they may contain. Such devices are not designed to resist professional attacks, and it takes a professional only a few minutes to extract all the data that has been stored in them. Theft of corporate mobile equipment that has been used to store passwords and access authorization codes can be a security nightmare for information technology managers. If consumer-initiated mobile commerce transactions move off the carrier network to a less secure messaging center or server, there may be vulnerability for eavesdropping, credit card fraud, or identity theft. Third-party applications that are downloaded onto the mobile device could open up a potential path for malicious viruses and rogue applications that could spread from user to user.

Fortunately, the basic principles of online security apply equally well to the wireless environment. These include the following:

- End-to-end encryption to protect against eavesdropping;
- Secure digital identity and authentication of users;
- Storage of private keys, financial, and other confidential information in the most secure region of the mobile device; and
- Testing and certification of all applications before downloading via an authorized application management infrastructure.

The tamper resistant properties of the SIM and the encryption of all network traffic provide a first line of security defense. The SIM is also the ideal location for storing private keys and other sensitive data. Mobile commerce users and content and service providers using mobile PKI can obtain digital certificates that authenticate their identity and create a nonrepudiation trail for wireless transactions. The GlobalPlatform infrastructure for secure application loading and unloading, originally designed for managing applications on financial services smart cards, has developed specifications for mobile application downloading and management.

These solutions have been tested and proven to protect effectively against the most common mobile commerce security risks, but a mobile PKI solution is expensive to implement and complicated to maintain. Biometrics provides another solution to authenticating the identity of mobile users and protecting the integrity of transactions. Mobile devices with integrated fingerprint sensors can securely identify the owner of the device, prevent any unauthorized users from accessing data or imitating the owner, and also generate a valid digital signature for binding mobile transactions.

Planning for any mobile commerce program should include evaluating, electing, and applying the most cost-effective solution that is appropriate for the risk involved in a particular type of transaction and then educating the target market for that program about the protection that comes with the every mobile interaction.

SUCCESS FACTORS FOR THE FUTURE OF MOBILE COMMERCE

Nothing stays still for long in the wireless world. Technology advances in mobile equipment already make the "futuristic" phones featured in 2000 seem quaintly old-fashioned. Today's phones support full-color multimedia, incorporate digital cameras and biometric sensors, and provide high-quality audio for downloaded music. Tomorrow's phones will no doubt offer characteristics that will redefine the limits of wireless communications. Prototypes of miniature phones that can be implanted in a tooth, flexible phones that we can wear, and phones with the computing power of fairly recent desktop PC generations are already rolling out of research laboratories. Some of them will become mere curiosities, and others will be embraced by millions of users.

The transition to 3G networks that is just underway will be followed by even faster fourth-generation (4G) network implementations within the decade. 4G cellular phones will deliver high-resolution movies and television programs to mobile subscribers, allow them to play high-resolution games with multiple players, and support applications that have yet to be invented. The next generation of networks will aim maximum data transmission speeds of more than 20 megabits per second. That is almost 2,000 times faster than the typical 2G network and 10 times faster than the much-anticipated 3G. Wireless innovation in general is inexorable, but the success of any specific mobile commerce application is dependent on a number of factors. Although it is difficult to ensure that a specific mobile commerce application will be a winner, the early adopters of mobile commerce have provided a roadmap to the core requirements for mobile commerce success in a region or vertical market. The key success factors are the following:

- A critical mass of wireless subscribers;
- Availability of a variety of mobile applications that are easy to use and cost-effective and that provide a visible value for the customer;
- Mobile devices that provide an enjoyable user interface for the applications;
- Interoperability of applications, devices, and networks around the world;
- Seamless roaming between wireless telecommunications networks, the Internet, and wireless local area networks such as WiFi;
- A business model that provides appropriate incentives for all participants needed to create a mobile value chain from application design to distribution, support, and billing; and
- Assurance that the mobile transaction and the identity and confidentiality of the customer are secure.

There are signs that the stakeholders of the wireless industry recognize that this foundation for mobile commerce success can only be built on truly open standards and cooperation among competitors to increase the market opportunity for all. In June 2002 more than 200 of the world's largest wireless carriers, mobile device and network infrastructure suppliers, and information technology vendors and content providers established a cooperative, global organization called the Open Mobile Alliance (OMA). According to its charter statement,

> The Open Mobile Alliance will collect market requirements and define specifications designed to remove barriers to interoperability and accelerate the development and adoption of a variety of new, enhanced mobile information, communication and entertainment services and applications, while ensuring the interoperability of mobile services across markets, terminals and operators, throughout the entire value chain.

This type of industrywide cooperation, if actually implemented, will go a long way toward addressing the remaining barriers to interoperability and adoption of mobile commerce applications.

CONCLUSION

The flurry of media coverage and optimistic revenue projections that accompanied the debut of mobile commerce several years ago echoed the overheated enthusiasm of early e-commerce development, and like e-commerce, the implementation of truly value-added applications lagged behind the initial growth projections. It is important to remember that many of the key ingredients for mass mobile commerce adoption are already in place and others are evolving quickly to stimulate the interest of a new generation of wireless customers. As long as developers, handset makers, infrastructure providers, and wireless operators work together to create mobile applications that deliver visible value to consumer and enterprise users, customers will try and ultimately buy the most relevant of these applications and services. It is still difficult to predict the exact timing of mass adoption of mobile commerce in different parts of the world, but it is certain that mobile commerce will continue to grow and expand in scope over the next decade.

GLOSSARY

2G Second generation digital cellular phone networks.

3G Third generation digital cellular phone networks.

3GPP The 3rd Generation Partnership Project, responsible for managing the standards that underpin the transition from GSM 2G to 3G networks worldwide.

Code Division Multiple Access (CMDA) A spread spectrum network technology for 2G digital cellular networks.

CDMA2000 The third generation upgrade of CDMA, providing faster data transmission speeds and better performance.

Cryptography The practice and study of encryption and decryption—encoding data so that the result can only be decoded by specific individuals.

Global system for mobile communications (GMS) A digital cellular phone technology based on TDMA that is the predominant system in Europe, but is also used around the world.

Mobile equipment (ME) A cellular handset, personal digital assistant, or other mobile device used by the wireless subscriber

Multimedia messaging service (MME) Enhanced messaging for digital cellular networks that provides graphics, animation and sound in addition to plain text.

Personal digital cellular network (PDC) The wireless network standards adopted in Japan for 2G wireless. PDC runs on the TDMA interface.

Subscriber identity module (SIM) A module that is inserted into a mobile device for subscriber identification and other security related information. GSM phones use a SIM smart card that contains user account information. SIM cards can be programmed to display custom menus for personalized services.

Short message service (SMS) A message service offered by digital cellular networks.

Time division multiple access (TMDA) A type of multiplexing in which two or more channels of information are transmitted over the same link by allocating a different time interval ("slot" or "slice") for the transmission of each channel.

Wireless application protocol (WAP) An open, international standard for applications that use wireless communication. WAP provides a complete environment for wireless applications that includes a wireless counterpart of transmission control protocol/Internet protocol (TCP/IP) and a framework for telephony integration such as call control and phone book access. WAP features the wireless markup language (WML), a streamlined version of hypertext markup language for small screen displays. It also uses WMLScript, a compact JavaScript-like language that runs in limited memory.

CROSS REFERENCES

See *Bluetooth^{TM}—A Wireless Personal-Area Network; Mobile Devices and Protocols; Mobile Operating Systems and Applications; Propagation Characteristics of Wireless Channels; Radio Frequency and Wireless Communications; Wireless Application Protocol (WAP); Wireless Communications Applications; Wireless Internet.*

REFERENCES

Abbott, L. (n.d.). What is m-commerce? *MobileInfo.com*. Retrieved June 15, 2002, from http://www.mobileinfo.com/Mcommerce/

CDG, CDMA Development Group (n.d.). Retrieved April 1, 2003, from http://www.cdg.org/worldwide/cdma_world_subscriber.asp

Ernst & Young (n.d.). Beyond e-commerce: Wireless and mobile business and the search for value. *The Ernst & Young Thought Center.* Retrieved June 15, 2002, from http://www.ey.com/global/Content.nsf/US/Issues_Perspectives_-_Index_-_Alphabetical

International Telecommunications Union (ITU) (2002, March). World Telecommunication Development report, executive summary. Retrieved June 15, 2002, from http://www.itu.int/ITU-D/ict/publications/wtdr_02/material/WTDR02-Sum_E.pdf

Taylor Nelson Sofres (May 2002). Wireless and Internet technology adoption by consumers around the world. Retrieved April 1, 2003, from www.tnsofres.com/IndustryExpertise/IT/WirelessandInternetAdoptionbyConsumersAroundtheWorldA4.pdf

FURTHER READING

Dornan, A. (2000). *The essential guide to wireless communications applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall PTR.

Guthery, S., & Cronin, M. J. (2002). *Mobile application development for SMS and the SIM toolkit.* New York: McGraw-Hill.

Kalikota, R., & Robinson, M. (2001). *M-business: The race to mobility*. New York: McGraw-Hill.

May, P. (2001). *Mobile commerce: Opportunities, applications, and technologies of wireless business.* New York: Cambridge University Press.

Paavilainen, J. (2002). *Mobile business strategies: Understanding the technologies and opportunities*. Boston: Addison Wesley Professional.

Raina, K., & Harsh, A. (2002). *mCommerce security: A beginner's guide*. Berkeley, CA: McGraw-Hill Osborne Media.

Sadeh, N. M. (2002). *M commerce: Technologies, services, and business models.* New York: Wiley.

Strader, T., & Mennecke, B. (Eds.). (2003). *Mobile commerce: Technology, theory and applications.* Hershey, PA: Idea Group Publishing.

Mobile Devices and Protocols

Julie R. Mariga, *Purdue University*
Benjamin R. Pobanz, *Purdue University*

INTRODUCTION

Mobile Device History

In 1993, Apple Computer introduced the world's first personal digitial assistant (PDA), the Newton Message Pad. Concurrently, Sharp Electronics introduced their version, the ExpertPad. Other than the screen cover and internal pen cradle, the ExpertPad was identical to Apple's Newton. The Sharp ExpertPad combined Newton's intelligence technology with some sophisticated communication capabilities, however. For example, Sharp was able to synchronize the information from the ExpertPad with a personal computer. The synchronizing of information was exchanged through built-in infrared technology. Newton's hardware architecture had an ARM 610 processor and ran at 20 MHz with a resolution of 336 × 240 pixels, all operating on Apple's 1.3 operating system (OS). Soon after the ExpertPad, Sharp developed the Zaurus to compete head to head with the Newton. The Zaurus predominantly relied on handwriting recognition instead of a keyboard, which was added in 1995. In 1996, Palm Inc. introduced the Pilot 1000 and Pilot 5000 organizer models, adding another player in the mobile device arena. After its initial product release, Palm captured a significant market share, which reached 73.6% in 1999. However, by 2001 its market share declined to 35% ("Palm stays atop," 2002), because of the rise in popularity of Compaq's iPAQ and Microsoft's PocketPC (Allnetdevices, 2000).

MOBILE COMPUTING

Mobile Computing

Mobile computing merges cell phones and computer technology. Resulting from this merge are mobile devices that unite the mobile communications of the cell phone and the expanded functionality of a computer. One company specializing in this field is Symbol Technology, which designs products that use radio frequency identification (RFID), infrared, and wireless data transmissions. This chapter previews several devices and highlights the standard wireless communication protocols, so-called smart phones, which are beginning resemble PDAs and PocketPCs. Some of the IEEE wireless communication protocols that will be highlighted are the Institute of Electrical and Electronics Engineers (IEEE) Standards 802.11, 802.11b (Wi-Fi), 802.11a, and 802.11g. A brief background on Bluetooth (a short range wireless connectivity standard) and wireless application protocol (WAP), as well as RFID and global positioning system (GPS), are also addressed in limited form.

The Differences Between Mobile and Nonmobile

It is important to distinguish the differences between mobile and nonmobile devices. The distinction is dictated by the use of a physical connection in the form of a cable, in which the device is connected to a network. Devices using a network cable are categorized as nonmobile. Conversely, devices that use wireless connectivity via a wireless modem are mobile.

Different Styles of Mobile Devices

The IEEE essentially defines a mobile device as a small portable electronic organizer, PDA, smart phone, clamshell, cell phone, laptop, or a PDA and phone combination that allows a user to enter and manage data without a physical connection to a network. In addition, levels of information processing and data functionality are also used to classify mobile devices. Some high-end PDAs, like the Palm i705, have a color display. (The majority of PDAs have a monochrome or gray-scale display.) The current color range of the PDA and PocketPC models are anywhere from 256 to 64,000 colors. PDA screen displays are either active matrix or passive matrix. Active matrix (AM) liquid crystal display (LCD) uses transistors to control each pixel. AMLCDs are more responsive, faster, and can be viewed at larger angles than passive matrix displays, making them easier to read. Passive matrix displays use fine lines that intersect. The point of intersection is called an LCD element, which presumably either blocks or passes light through for viewing (IEEE, 2001). PDAs perform an information exchange or synchronization with a desktop computer by placing the PDA in a cradle, or docking station. The docking station connects directly to the desktop computer with a cable, and, using synchronization software, the data exchange is performed. Some models have infrared technology that can synchronize without the use of wires or a docking station. For infrared synchronization to occur, both devices must have the capability from a hardware and software standpoint. In general, the higher the central processing unit (CPU) speed, the faster the unit will execute tasks and the more expensive the unit. Nevertheless, higher speed processors require more battery power and may deplete batteries quickly. Some larger models use AAA alkaline batteries, and others use lithium-ion rechargeable batteries. Models typically have 2–32 MB of random access memory (RAM), and some models may have an expansion slot built into the device, allowing for a memory upgrade. In general, for PDAs, the degree of Web accessibility is not yet equal to what you gain from a full-fledged computer or PocketPC.

The Differences Between a PDA and a PocketPC

Although both PDAs and PocketPCs are mobile, the pinnacle distinction can be found in the device's ability to engage software applications and manipulate data. Mobile computing devices allow you to store, organize, send, and receive messages as well as access information without a physical hard-wire connection. Mobile computing devices have a CPU, which allows them to receive data, perform a specific process with the data, and produce output. A PDA lacks general-purpose application software needed to perform data manipulation functions like that of a mobile computer. Although the early PDAs paved the way for mobile computing to reach its current level of utility, today's PDAs are less functionally dynamic compared with PocketPCs.

Components and Data Input

Entering data is done either through a scaled-down keyboard or through a penlike stylus. The stylus is primarily used in the writing area called the graffiti pad. Either rechargeable or disposable batteries power all devices, as dictated by the manufacture. PDAs primarily use a variation of the Palm OS, and a Motorola 33-MHz DragonBall CPU, whereas the PocketPCs use a StrongARM 206-MHz CPU and run a Windows OS. The amount of memory will vary with the make and model of the device. Optional are springboards and memory sticks. The average PDA screen size is 3.6 inches, with resolution ranging from 160 × 160 to 240 × 320 pixels. On PocketPC devices, the viewable screen averages 3.5 inches with resolution of 240 × 320–pixels. PDAs typically have between 2–32 MB of built-in memory. Often 2-MB memory is generally sufficient for maintaining address books, an active calendar, taking notes, and loading useful programs. More memory may be necessary for storing larger files, such as digital photos or audio recordings. Increasing the amount of memory can be accomplished with small flash-memory storage cards. These cards are inserted into slots in the PDA. The memory in most PocketPC models typically ranges from 32–64 MB. Adding smart memory cards in either PDAs or PocketPCs is advantageous when applications or files call for increased memory.

There are essentially three data-entry options available to both PDA and PocketPC users: the mini portable fold-out keyboard, requiring a flat surface and thus limiting where the keypad can be used; the numeric keypad, which is identical to those on cell phones; and the stylus, which has similar functions to that of a mouse and is used for navigation.

Functions and Features

Most PDAs operate on a Palm OS. Data input can be screen based, keyboard based, or both. More sophisticated devices, such as the PocketPC, use the Windows-based OS (PocketPC). What distinguishes the PDA from the handheld or PocketPC is the lack of window's application software that is available and compatible for these devices. PocketPC is capable of running word-processing, spreadsheet, and money management applications and provides Internet and e-mail access. The Palm OS has some office management software, but it is not well-suited to Microsoft office applications. PDA devices are typically less expensive than mobile computing models, because most lack advanced software applications.

To increase the functionality of some devices, manufactures add certain features. For example, Handspring's eyemodule2 transforms the PDA into a digital camera by inserting the cartridge into the back. A GPS module can be inserted when needed. Some of the more advanced PDA models allow e-mail and Internet access that is done through a conventional or wireless modem hook up. Some Internet features common on most desktop computer may not be available to users on a mobile device, however. For instance, viewing information in certain file formats such as pictures, Adobe files, or various multimedia programs such as Shockwave or Real Player may be limited.

Mobile Devices in E-commerce

In the early 1990s, industry and organizations began to dispatch hardware and software applications to streamline workflow and improve communications. These early and primitive wireless applications where leveraged as a means of increasing worker and business productivity, particularly among field service personnel, who were able to reduce the use of paper printouts listing daily service orders by simply calling a dispatcher for information on the next service call. (Today service calls are sent to the field professionals' mobile device as a text or voice message, thus eliminating a call to dispatchers.) The early applications improved the communication process yet lacked any transaction process that generated revenue. For example, before mobile communication, service workers contacted a dispatcher for work orders. Now, using wireless devices, technicians are tracked via GPS, and work orders are automatically sent to the closest technician.

In today's e-commerce arena, business has evolved into mobile commerce (m-commerce) by leveraging the popularity of mobile devices. Revenue is no longer dictated by set business hours. In other words, just because the front door to the business is locked doesn't mean that the business is not generating revenue. A new brand of shopper has emerged from advances in technology and the Internet. According to McGuire (2001), mobile commerce is defined as the transaction of goods and services made by a buyer or seller using a data-enabled wireless device over a wireless data connection. Soon cell phones will be used to purchase snacks and drinks from vending machines. The use of Web-enabled smart phones, PDAs, and PocketPCs are transforming the way people do business, both in the business-to-business and business-to-consumer arenas.

MOBILE DEVICES

Palm Models

Palm i705

The main strength of the Palm i705 device is its built-in wireless modem that allows users to be connected to the Internet at all times. Its weaknesses lie in usability. For

Table 1 Mobile Device

Manufacturer	Model	OS	CPU	Screen	Weight
Casio	Cassiopeia *BE-300*	Windows CE	166 MHz /	3.5 in	5.9 ounces
	Cassiopeia E-200	Pocket PC 2002		240 × 320	5.9 ounces
Compaq	IPAQ pocket PC	Pocket PC 2002	206 MHz	3.7 in	6.7 ounces
Computer	H3835		StrongARM	240 × 320	
	H3850				
Handspring	Visor neo	PALM OS 3.5.2	33 MHz Motorola	3.0 in	5.4 ounces
	Visor pro		DragonBall	160 × 160	5.7 ounces
	Visor edge				4.8 ounces
	Visor prism				5.4 ounces
Hewlett	Jornada 565	Pocket PC 2002	206 MHz	3.5 in	5.9 ounces
Packard	Jornada 568		StrongARM	240 × 320	6.1 ounces
NEC	MobilePro P-300	Pocket PC 2002	206 MHz	3.8 in	7.9 ounces
	MobilePro 790		StrongARM	240 × 320	2 lb
Palm	Pilot 1000	Palm OS 3.5	33 MHz Motorola	3.6–3.8 in	5.7 ounces
	Pilot 5000	Uses faster-	DragonBall	160 × 160	5.8 ounces
	Palm IIIxe	seeming			5.6 ounces
	Palm IIIc	DragonBall EZ			6.8 ounces
	Palm V	processor and			5.4 ounces
	Palm VII	Palm OS 3.1+			4.4 ounces
	Palm M 100	Palm OS 4.0			6.4 ounces
	Palm M500	Palm OS 4.1			4.0 ounces
	Palm M505				4.9 ounces
	Palm i705				5.9 ounces
PC-EPhone	EPhone	Pocket PC	206 MHz	6.0 in	10.5 ounces
			StrongARM	240 × 320	
Blackberry	857	Proprietary	32bit 386	3.0 in	4.7 ounces
	957			160× 160	4.8 ounces
	5810				4.7 ounces
Sharp	Zaurus	Linux 2.4 QT	206 MHz	3.5 in	7.3 ounces
	SL-5500	Personal Java	StrongARM	240 × 320	
Sony	CLIE′ PEG	PALM OS 4.1	33 MHz	320 × 480	4.9 ounces
	T-615C		Motorola		
	CLIE′ PEG	PALM OS 4.1	DragonBall		5.7 ounces
	T-760C				
Symbol	PPT 2800	Pocket PC 2002	206 MHz	320 × 240	11.8 ounces
			StrongARM		
Toshiba	570	Pocket PC 2002	400 MHz	240 × 320	6.7 ounces

Sources: Blackberry (2003), Casio (2002), Compaq (2003), Handspring (2002), Hewlett Packard (2002), NEC (n.d.), Nokia (2003), Palm (2002), PC-EPhone (2003), Sharp (2003), Sony (2003), Symbol (2003), and Toshiba (2003).

example, it does not have a built-in keyboard and only has 4 MB of read-only memory (ROM) and 8 MB of RAM. (See Table 1.)

Palm m515, m505, and m500

The Palm m515 offers a color display and an adjustable backlighting feature, as well as 16 MB of internal memory. The internal memory can be used for working with application files such as Microsoft Excel spreadsheets and Word documents or viewing presentations and video clips. The m505 and m500 models extend PDA functionality with a secure digital and multimedia card expansion slot and a universal connector for memory. Additional features can be added that help with information backup, wireless modems, and camera options. The visible differences are found in the screen background; the Palm m500 handheld uses a high-contrast monochrome display on a white background and is limited to 8 MB of memory. The Palm m505 handheld has a 16-bit color display. The PalmVII is the first wireless Web-enabled device produced by Palm. This unit requires the user to uplink to receive e-mail, rather than synchronizing periodically throughout day.

Palm m100

In addition to the standard software applications, the m100 model has a the date book, address book, to-do list, memo pad, note pad, and calculator. It has a monochrome display, requiring two AAA batteries and has 2 MB of memory.

Handspring Models

Handspring Treo

The Treo is the latest mobile device from Handspring. The Treo combines the dynamics of data organization in a PDA with the communication features of a cell phone. This mobile computing device can send and receive e-mail and text messaging, has wireless Internet access, and can place phone calls. The Treo is a dual-band phone, thus allowing three-way calling. It uses the 3.5 Palm OS and 33-MHz Motorola DragonBall processor. The device has 16 MB of memory and a 160 × 160–pixel monochrome touchscreen. One advantage of this Handspring device is that it comes with a switch that enables easy toggling between the standard ringing tones or vibrate mode. It also has a dedicated message button allowing quick access to manage e-mail and short text messages.

Handspring Visor Pro

This PDA has 16 MB of ROM and 0 RAM. The Visor Pro runs on the Palm OS and weighs 5.4 oz. A weakness of this model is that it as a side-mounted infrared port, which can be a challenge when lining up infrared technology. The Visor Pro uses a 33-MHz DragonBall processor and has a 160 × 160-pixel touchscreen. The expansion slots can integrate a digital camera, GPS, and Visor-phone, which use Handspring's Internet browser, Blazer Browser.

Handspring Visor Neo

The Visor Neo uses the Palm OS and software applications. Using Motorola DragonBall VZ 33 MHz processor and 16 MB of memory, this unit can send and receive information with any infrared equipped Palm OS. The main issues challenging Handspring are primarily related to the business's production.

Sony Models

Sony Clié PEG-T 615C

This device uses a 33-MHz Motorola DragonBall processor and runs the Palm desktop and Palm OS applications. This Clié has 16 MB of RAM and a 320 × 320–pixel high-resolution color screen. It weighs 4.9 oz and has a memory slot expansion slot.

Sony Clié 760C

This model has 16 MB of memory and an additional slot for a memory stick, which is needed to view videos and listen to MP3s. The Sony Clié has a 320 × 320–pixel, high-resolution color screen and a center toggle switch, which is used to navigate through applications. A weakness with the 760C is that additional software is needed work with Microsoft Word and Excel files. Also, the screen is reduced in size to accommodate the graffiti pad.

Casio Models

Cassiopeia E-200

This device is a PocketPC running windows CE OS. With its 64 MB of RAM and a color-screen resolution of 240 × 320 pixels, the E-200 offers desirable functionality to the user. The only limitation is the lack of software applications.

Cassiopeia BE-300

This device is a PocketPC that runs 166-MHz processor and has 16 MB of RAM, a 320 × 240–pixel color screen, and weighs 5.9 oz. The E-200 has a Type II expansion slot available to a CF card. Data entry is done with a touchscreen and stylus. The lack of available software is the main drawback of this device.

HandERA Model

HandERA Model 300

The HandERA 330 weighs 5.9 oz and has 8 MB of RAM. It runs on the Palm OS. The gray-scale screen resolution is 240 × 320 pixels. The advantage of this device is that the entire screen can be used to work with data, and the graffiti window can be hidden or brought back on command. Minimal memory and the use of a slower and more cumbersome serial cable (compared with a USB or infrared port) to synchronize data are its limitations.

Hewlett Packard Jornada and Compaq iPAQ Models

iPAQ Models

The iPAQ H3870 Pocket PC includes Microsoft's Pocket PC 2002 OS, 32 MB of ROM, 64 MB of RAM, Intel's 206-MHz StrongARM SA-110, and a USB cradle. There are three models to choose from. The latest version is the H3870 PocketPC. It has embedded Bluetooth technology. This 3870 weighs 6.6 ounces and has a 3.7-inch color screen with a resolution of 248 × 320 pixels. It also has an Secure Digital (SD) card option. The iPAQ H3760 has 64 MB of RAM and is identical to H3870 with one exception: it does not have Bluetooth. The iPAQ H3850 also has 64 MB of RAM and uses a dual USB or serial cradle. It also incorporates a card slot and the use of voice-command and voice-control software. It also does not have Bluetooth.

The Jornada Model

The Jornada 564 PocketPC features a robust 206-MHz, 32-bit StrongARM processor and 32-MB synchronous dynamic RAM (SDRAM). Its 240 × 320–pixel display supports up to 65,536 colors and features a 3.5-inch color LCD touchscreen. Jornada's 565 Pocket PC also has a Type I memory slot. The Jornada 564 is capable of wireless and Internet functionality. A weakness of the Jornada is its small keypad, which makes data entry difficult. The Jornada 568 weighs 6.1 oz has a 206-MHz, 32-bit StrongARM processor; it has 64 MB of SDRAM and 32 MB of Flash ROM. The screen has a resolution of 240 × 320 pixels.

NEC Models

NEC MobilePro P300

This device weighs 7.9 oz, making it heavier than the 5.7-oz average. It has a screen resolution of 240 × 320 pixels. The P300 also supports wireless connections and secure remote access capabilities. The P300 has two expansion slots supporting a 32-MB Compact Flash (CF) card and a 32-Mb SD card. A weakness of this device is the magnetic cover used to hold the device shut; magnetic fields can damage data. The 2002 model will phase out the magnetic cover and reduce the weight to 6.7 oz.

MobilePro 790

This is a clamshell-style device that weighs 2 pounds and uses a 168-MIPS (million instructions per second) processor. It operates using Microsoft Windows for Handheld PC 2000 and features 24 MB ROM, 32 MB RAM, and internal 16-MB Flash Memory that allocates 14 MB available for storage. The 790, like Hewlett Packard's Jornada, is weak in keypad functionality for data entry.

PC-EPhone Model

PC-EPhone

The PC-EPhone claims to be a Palm-sized PC that integrates cell phone functionality. It uses the 206 MHz StrongARM processor and has a 4.0-inch display with 640 × 480–pixel resolution. It has 32 MB of RAM and 32 MB of ROM. Little is known about its durability and compatibility.

Research In Motion BlackBerry Models

The RIM 857, RIM 957 and RIM 5810

The RIM models of the BlackBerry line of mobile devices are at the cutting edge of wireless technology and business connectivity. Once turned on, BlackBerry mobile devices remain connected to a wireless network. It does not require synchronizing of applications; e-mail messages find the device. The models differ in their 8 or 20 line graphical display. Currently, only the 5810 provides service that integrates customers' Internet service provider e-mail account to the device. Depending on the user, the device can be a distraction because of its high level of connectability.

Sharp

Sharp Zaurus SL-5500

This device incorporates a 206-MHz StrongARM processor and a 3.5-inch thin film transistor (TFT) front-lit screen with 65,536 colors and a resolution of 240 × 320 pixels. The SL-5500 weighs 7.3 oz and has 16 MB of ROM and 64 MB of RAM. The two expansion slots house the Compact Flash and SD cards and can be handled simultaneously. The Zaurus has a built-in cursor key for simple navigation and runs Linux OS. One weakness is that this unit caters to more technical users with its customizable command line prompt option. Also, there is no technical support for Macintosh users, and because it uses the Linux OS, application functionality may be an issue.

Symbol Technologies

Symbol PPT 2800

The PPT 2800 is the "industrial strength" PocketPC, with 64 MB of RAM and 32 MB of ROM all enclosed in an outer casing that can withstand a 4-foot drop onto concrete. Its fundamental application environment is in large retail chains and in manufacturing industries. It weighs 10.3 oz and has 320 × 240–pixel resolution and an optional 16-level gray scale or 64-k color background. The PPT 2800 has the functional expandability to meet most applications. A definite benefit is the removable backup battery pack, eliminating the need to cradle the device to charge the battery. Also, the wireless Internal (WLAN), External (WWAN) antenna is built in, securing the modem from unexpected impact. The cost of the PPT 2280 is the main drawback. For the average user in the market for a PocketPC, the PPT 2880 may be cost prohibitive.

Toshiba

Toshiba 570 PocketPC

The 570 weighs 6.7 oz and is equipped with a 400-MHz Intel PXA250 processor. It also has a 3.5-inch TFT color display with 240 × 320–pixel resolution and 65,536 colors. The 570 has 64 MB of SDRAM and a slot for both the SD and CF card. Along with integrated Wi-Fi (IEEE Standard 802.11b), the 400-MHz processor and dual expansion slots offer several noteworthy benefits.

SMART PHONES

Nokia

Nokia 7650

The Nokia 7650 weighs 5 oz and uses multimedia messaging, accomplished by merging image, audio, and traditional voice calls. The 7650 also uses the Symbian OS, has an integrated WAP browser, and is outfitted with Bluetooth. A limitation is the sliding keypad, which may make it challenging to use with one hand.

The Nokia 9000i Communicator

The Nokia 9000× Communicator series has been around since 1996 and has evolved into a sophisticated palmtop computer, built in the form of a global systems for mobile communication (GSM) cellular phone. The Nokia 9000i has a clamshell design; when closed the 9000i operates as a normal cellular phone but with text messaging. When opened, the phone is transformed into wireless PDA, the top half delivering a 640 × 200–pixel LCD screen and the bottom half offering a keyboard. The Communicator uses the GEOS 3.0 OS and has a 24-MHz Intel 386 processor with 8 MB of memory. The memory is divided into three individual partitions, 4 MB allocated to the OS and application software, 2 MB for program execution, and 2 MB for data storage. The pitfalls are the small keyboard and the fact that users must navigate through screens using buttons.

Nokia 9290

The Nokia 9290 addressed several of the challenging issues with the 9000i and increased the amount of memory to 56 MB and talk time to 10 from 3 hours. The memory is divided with 16 MB committed to user needs, 8 MB to execution, 16 MB to application, and 16 MB to the memory card. The 9290 uses a 32-bit ARM RISC CPU on the Symbian OS.

Ericsson

Ericsson T86 tri-mode GPRS

The Ericsson T86 weighs less than 3 oz and runs on a Symbian OS. This phone has a color screen and built-in Bluetooth capability. It also has optional modules for wireless earplugs, a video camera, and a four-way navigational joystick. A weakness is that the screen size is tiny and the buttons are small.

Ericsson T61d

This device features voice-activated dialing and supports Bluetooth technology. With a calendar, calculator, and stopwatch, the T61d is blurring the line between cell phones and PDAs. As with most smart phones, the main challenge is screen size. The seven-line graphic display limits the amount of information that the user can view.

WIRELESS PROTOCOLS

Institute of Electrical and Electronics Engineers (IEEE)

The IEEE 802.11 standard is the foundation on which PDAs and PocketPCs operate. Products that employ this technology support a broad range of enterprises and individual users. With such explosive growth in wireless infrastructure and applications, understanding some basics about mobile computing technology and its limitations and variations can be helpful. This section touches on the protocols used for mobile communications and highlights some of the differences. This section of the chapter is a general overview, not a comprehensive analysis. (See Table 2.)

Table 2 Smart Phones

Manufacturer	Model	Screen	Weight
Nokia	9000i	Gray Scale	13.9 ounces
	9290	Color	8.6 ounces
Ericcson	T86 tri-mode	Color	3.0 ounces
	T61d	Color	3.1 ounces

Sources: Nokia (2002) and Mobileinfo (2003).

Over the past 15 years, the use of mobile computing devices has increased rapidly as cellular and digital services have become more affordable. Businesses are incorporating cell phones and PDAs as a way for employees to manage their work. This equipment also provides more efficient customer service. Unfortunately, as the popularity of mobile units increased, compatibility issues have become problematic. To remedy this, the IEEE partitioned the 2.4-GHz microwave band into two separate sections, frequency hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSS). The standards that emerged are 802.11 (FHSS), 802.11b (DSSS), 802.11a (DSSS), 802.11g (DSSS), and 802.11e (DSSS). Their characteristics are as follows: 802.11 uses the 79 distinct FHSS channels for transmission of data; 802.11b uses a more exotic encoding technique and 14 fixed frequency channels up to 11 Mbps transmission in the license-free 2.4 GHz band. Because of current market share, 802.11b is the current wireless standard supporting all Ethernet network protocols. Data transmissions via 802.11a can reach up to 54 Mbps in the 5-GHz band using an orthogonal frequency division multiplexing (OFDM) encoding scheme. Standard 802.11g addresses QoS side of wireless networking and is currently in the development stages. Currently IEEE is investigating how best to standardize 802.11e as a wireless security protocol (IEEE, 2002). The frequencies of 802.11b and 802.11g share wavelength space with microwave ovens and cordless telephones, causing signal collision and interference.

Wireless Application Protocol (WAP)

In 1997, Ericsson, Motorola, and Nokia formed the wireless application protocol forum (WAP Forum, n.d.). This forum was established to provide a worldwide open standard, enabling the delivery of Internet service to cell phones. WAP technology is aiming to be the global standard for mobile phones and wireless devices. The wireless application protocol supports standard data formatting and transmission for wireless devices. An important issue for WAP is that applications can be developed on any OS, thus providing interoperability among device families (WAP Forum, 2001). This could cause an unnecessary increase in the number of interfaces that Web developers will need to design and maintain. We are in the earliest stages of wireless application development, and many changes will occur in the future. It will be important to reassess strategic goals before taking action on cutting-edge and nonstandardized wireless protocols. See Figure 1 for WAP infrastructure overview.

Radio Frequency Identification

A basic RFID system consists of three components: the antenna, the transponder (tag), and the transceiver (tag reader). The antenna can be incorporated with the transceiver and decoder, becoming a reader. Readers can be handheld devices, such as the Symbol PPT 2280. RFID

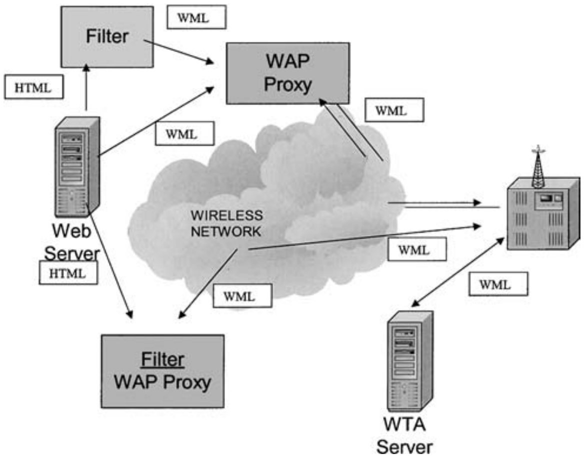


Figure 1: WAP infrastructure overview.

works as follows: the antenna emits radio signals to activate the tag; once the tag is activated, the reader either reads the tag or writes information to it (Table 2). The type of tag dictates the functionality. Tags are either active (read and write) or passive (read only). The range of data transmission can reach more than 90 feet, subject to ideal working environments.

Global Positioning System (GPS)

GPS is a method using satellites and radio frequencies to determine a specific position without making observations of one's surroundings. The GPS satellites orbit at about 12,000 miles above the Earth (Table 3). GPS devices can process signals received from four or more satellites' transmissions simultaneously to calculate the receiver's location within 6 feet (Punaridge, 2002).

Bluetooth: The Bluetooth Protocol Stack

Originally created by the Ericsson Company, Bluetooth is a short-range wireless technology similar to the 802.11 family. The use of Bluetooth in mobile computing is designed to link devices within a range of 30–35 feet. "Location independent" or ad hoc networks will benefit most from Bluetooth. Another feature is that Bluetooth performs authentication, payload encryption, and key handling. Because Bluetooth continuously hops between channels and switches between synchronous connection-oriented and asynchronous connection-less links, packet interference is primarily eliminated. With the multitude of compatibility options associated with Bluetooth, and with the increase of devices being manufactured with Bluetooth capabilities, it will likely surpass 802.11b in the arena of wireless communication within the next 2 years (Egan, 2001). (See Table 3.)

Table 3 Wireless Application Protocol Infrastructure Overview

Applications Layer
Transport Protocol Layer
TCP/IP HID RF COMM
Logical Link Control Layer
Connections linking to devices are established and released
Link Manager Layer
Link Manager Protocol maintains connections
Baseband Layer
Coding/Decoding, packet handling and frequency hopping
Radio Frequency Layer
Frequency combination, convert bits to symbols

CONCLUSION

The pace of emerging technology requires both developers and end users to continually revisit trends and business practices. With new technologies in development, today's methods and processes will be replaced at a moment's notice, and mobile computing and mobile networking are no exception. Over the next decade, mobile computing will evolve rapidly, surpassing the original desktop explosion. The IEEE standards briefly mentioned in this chapter are the catalysts through which mobile computing will continue to grow. These standards, too, will change over the next 2 to 5 years, driven by a market demanding faster transmission speeds and more dynamic functionality from the numerous mobile devices.

GLOSSARY

Ad hoc A networking mode used for establishing a network where wireless infrastructure does not exist or where services are not required.

Direct sequence spread spectrum (DSSS) A redundant radio frequency that operates between the 5.5 and 11 Mbps data transmission ranges.

Frequency hopping spread spectrum (FHSS) The transmission technology used in a wireless local area network in which the data hops in a random but predictable sequence from frequencies; data rate is between 1 and 2 Mbps.

Flash memory A special type of memory that is electrically erasable and read-only programmable; Flash Memory can be erased and reprogrammed in blocks instead of a single byte at a time.

Global positioning system (GPS) A system that incorporates the use of satellites for navigational and location identification purposes.

General packet radio service (GPRS) A standard for wireless communications that runs at speeds up to 115 kbps; GPRS standard supports a wide range of bandwidths and is suited for sending and receiving small amounts of data.

IEEE Standard 802.11 A wireless standard that applies to wireless local area networks and provides 1 or 2 Mbps transmission in the 2.4-GHz band using either the frequency hopping or direct sequence spread spectrum.

IEEE Standard 802.11a An extension of 802.11 that uses the direct sequence spread spectrum scheme to provide up to 54 Mbps in the 5-GHz band with less overlapping channels.

IEEE Standard 802.11b A wireless standard that operates in 11 channels and has the ability to move 11 Mbps in the 2.4-GHz band.

IEEE Standard 802.11g A faster and more secure wireless standard that operates on only three channels compared with 13 channels for 802.11a.

Radio frequency identification (RFID) A wireless technology that operates between 850 and 950 MHz and 2.4 and 2.5 GHz.

Secure Digital (SD) A nonvolatile memory device that identifies objects and act as a permission gatekeeper.

Synchronous dynamic random access memory (SDRAM) is tied to the system clock and is designed to be able to read or write from memory at the speed of the processor.

Wi-Fi Wireless fidelity, or IEEE Standard 802.11b.

CROSS REFERENCES

See *Bluetooth^{TM}—A Wireless Personal-Area Network; Mobile Commerce; Mobile Operating Systems and Applications; Wireless Application Protocol (WAP); Wireless Communications Applications; Wireless Internet.*

REFERENCES

Blackberry (2003). Retrieved April 23, 2003, http://www.blackberry.net/products/handhelds/

Casio (2002). Retrieved April 23, 2003, from http://www.casio.com/index.cfm

Compaq (2003). Retrieved April 23, 2003, from http://h18000.www1.hp.com/

Egan, B. (2001, October). *Mobile and wireless computing: The next user revolution*. Paper presented at the Lake Beuna Vista, FL October 8-12, 2001 Gartner Group Symposium.

Forrester Research (2001). Retrieved June 6, 2002, from http://www.forrester.com/home/0,6092,1–2,FF.html

Handspring (2002). Retrieved April 23, 2003, from http://www.handspring.com/

Hewlett Packard (2002). Retrieved April 23, 2003, from http://www.hp.com/

IEEE (n.d.). Retrieved October 3, 2002, from http://www.ieee.org/portal/index.jsp?pageID = home

Jones, N. (2001, October). *Mobile commerce business scenario*. Paper presented at the Gartner Group Symposium.

Jüptner, O. (2001). 36 million emails per day. (n.d.). Retrieved March 24, 2002, http://www.e-gateway.net/infoarea/news/news.cfm?nid=1876

McGuire, M. (2001, October). *Mobile business markets: What can't users live without*. Paper presented at the Gartner Group Symposium, Lake Beuna Vista, FL, October 8–12, 2001.

Mobileinfo (2003). Retrieved September 2, 2002, from http://www.mobileinfo.com/

NEC (n.d.). Retrieved April 23, 2003, http://www.necsolutions-am.com/products/psgateway.cfm

Nokia (2002). Retrieved August 3, 2002, from http://www.nokiausa.com

Nokia (2003). Retrieved August 3, 2002, from http://www.nokiausa.com/ (Date of access: 8/3/2002).

Palm (2002). Retrieved June 10, 2002, from http://www.palm.com/

Palm stays atop handheld market. Retrieved January 26, 2002, from http://news.zdnet.co.uk/story/0,,t269-s2103166,00.html

PC-EPhone (2003). Retrieved April 23, 2003, http://www.pc-ephone.com/tech.html

Punaridge (n.d.). Retrieved from August 20, 2002, from http://www.punaridge.org/doc/factoids/GPS/Default.htm

Sharp (2003). Retrieved April 23, 2003, http://www.sharp.com

Sony (2003). Retrieved April 23, 2003, http://www.sony.com

Symbol (2003). Retrieved April 23, 2003, http://www.symbol.com/products/

Toshiba (2003). Retrieved April 23, 2003, http://www.toshiba.com

WAP Forum (n.d.). Retrieved November 14, 2002, from http://www.wapforum.org/

xAllnetdevices. Retrieved March 16, 2002, from http://www.allnetdevices.com/

FURTHER READING

Anatomy of IEEE 802.11b wireless. Retrieved October 17, 2001, from http://www.networkcomputing.com/1115/1115ws2.html

Comdex wireless technology. Retrieved November 20, 2001, from http://www.key3media.com/comdex

Compaq product page. Retrieved November 2002 from http://www.compaq.com

Goldman, J., & Rawles, P. (2000). Local area networks: A business-oriented approach. New York: Wiley.

Pulling the Internet's plug. Retrieved September 10, 2001, from http://www.networkmagazine.com/article/NMG20000710S0007

The Wireless age: Theories of connectivity. Retrieved April 22, 2002, from http://nytimes.com/search/abstract

Understanding 802.11b and Bluetooth. Retrieved November 4, 2001, from http://www.mightywords.com/browse/details_bc05

Webopedia. Retrieved December 6, 2001, from http://www.webopedia.com/

Wireless local area network IEEE 802.11. Retrieved October 12, 2001, from http://grouper.ieee.org/groups/802/11/main

Wireless revolutionizing campuses: Retrieved September 19, 2001, from http://thesource.micronpc.com/articles/050201.html

Mobile Operating Systems and Applications

Julie R. Mariga, *Purdue University*

INTRODUCTION

Mobile computing, defined as a generalization of all mobile computing devices including personal digital assistants (PDAs, e.g., Palm Pilots, Pocket PCs), smart phones, and other wireless communication devices, will to change dramatically in the next 2 to 5 years. There are a number of reasons for these changes, but two primary factors are the convergence of next-generation handhelds and high-speed wireless technology. The operating systems (OSs) found in today's handhelds will provide the foundation for future devices and applications. The world of technology is changing in many ways. Important considerations for companies are the challenges, trends, and opportunities of deploying mobile computing technologies. Companies need to ensure they have a mobile strategy in place to stay competitive and capitalize on new developments in the mobile computing area.

Many factors drive companies to grow their mobile computing infrastructure, including decentralized workforces, telecommuting, travel, device capabilities and proliferation, mobile infrastructure focus, networking choices, cost of ownership issues, mobile business to employee transactions, and companies' control of handheld devices. Which OS should companies or individuals implement? It depends on a number of factors. One important issue to consider is which application(s) need to be used on the device. Answering this question may help to eliminate certain operating systems and devices. Another important factor to consider is portability of applications, which is important because devices change rapidly. Portable applications can be used on new devices without having to be rewritten or upgraded. If applications are developed in a language that allows for portability, such as Java, then these can be deployed to a wide range of devices, including handhelds that support various operating systems, embedded Linux devices, and pure Java devices. Another important issue to consider in selecting an OS is what type of development tools are available as well as the number and strength of the programmers available to create and maintain applications. Currently, the Palm OS supports the largest number of packaged applications. Many of these, however, are better suited for individual rather than business use.

According to Jones (2001a), there are four main factors driving the mobile business phenomenon: (a) economics, (b) business needs, (c) social trends, and (d) technology. Economics includes the falling prices of mobile airtime and the inexpensive cost of devices. Jones stated that over the next 5 years, costs will continue to decrease, which will allow for new mobile applications to be developed and Bluetooth chip sets to cost under $5. This will enable electronic devices to be networked. Business needs include organizations requiring new types of mobile applications to increase customer service and allow for better supply chain management. With regard to social trends, in many countries, mobile devices have become a lifestyle accessory, mainly among younger adults. As young adults continue to want more functionality from their devices and applications, there will be a mix between the mobile technology and entertainment and fashion. Finally, new core technologies such as WAP, i-mode, Bluetooth, and 3G networks are enabling a new generation of mobile applications. As these four factors evolve, they will continue to push the growth of the mobile business arena.

COMPUTING ARCHITECTURES

It is important to understand the overall computing architecture when using mobile devices and developing mobile applications. There are two general architectures, the occasionally connected application model and the tiered computing architecture. This occasionally connected model would be used when the client is not permanently connected to the rest of the system. An important consideration with this model is that the developer should not make assumptions about data accessibility. The majority of business logic, processing, and will be done on the backend or server. The tiered computing architecture generally includes three types of architecture that fall into the tiered architecture model. The first is the one-tiered approach, in which the mobile device is used only as a display device, and the software, data validation, and business rule enforcement take place on the backend or server. In a two-tiered architecture, the business rules are divided, with some performed on the client-end application running on a dedicated device and the remainder on the server. An advantage to this type of architecture is that the network traffic can be reduced and overall performance improved. The disadvantage is that maintenance can be difficult. When a business rule changes, modifications need to be made to both the client device and the server. The three-tiered architecture model places a middle layer between the client device and the server. The middle layer implements a majority of the business processing, and the client device handles the display functions; the server provides data storage. The biggest advantage to this architecture is that scalability is extremely high. Any changes to business rules can be made on the middle layer. This type of architecture is also called n-tier architecture.

MOBILE COMPUTING OPERATING SYSTEMS

There are a number of choices available for both individuals and companies when selecting handheld devices. Before purchasing a device, a few important decisions need to be made about what OS will run the device and what type of applications will run on the selected OS. There are three primary operating systems to choose from for handheld devices: the Palm, Pocket PC, and Symbian operating systems. The following sections look at each of these operating systems. For the next 5 years, information system managers will be forced to confront the various mobile platform and OS dependencies. A key for building success in the mobile computing arena will be for organizations to have a strategic plan in place for how mobile computing fits with the overall business plans.

Palm Operating System

The Palm OS has been the most popular platform for handheld devices since 1997 when the first Palm Pilot was launched. Some advantages include its ease of use, market share, synchronization with many calendaring and scheduling applications, and its contact management systems. The Palm OS uses a simple handwriting recognition system called graffiti instead of a keyboard. A user can plug in a keyboard if desired. Since first introduced, the Palm OS has lost market share to devices running the Pocket PC OS and the Windows CE OS. Palm has created PalmSource as a Palm subsidiary, which will allow PalmSource the ability to concentrate on developing and enhancing the operating system. The Palm OS runs on the following devices: Palm, Acer, Handspring, HandEra, Kyocera, Samsung, Sony, and Symbol. As of this writing, the latest version is Palm OS 5, which has created new opportunities for end users and developers. It will continue to support the ARM-compliant processors from Intel, Motorola, and Texas Instruments, but it enhances the multimedia capabilities and incorporates more security options. It also provides for more wireless connections.

The Palm OS currently maintains a market share in the handheld device market even though it has lost market share in 2000 and 2001. With the Palm OS, software developers can build data applications to use with Palm devices, which can be implemented via wireless or synchronized data access to corporate data. According to Dulaney (2001), Palm OS devices will have the broadest appeal and application support for most organizations, setting the standard for handheld devices. The Palm OS has the following strengths:

- The number of partners working with Palm is extremely large.
- A large number of applications are available to run on the Palm OS.
- Palm has a healthy percentage of the market share.

Weaknesses with Palm and Palm OS are the following:

- The core OS functionality is limited (compared with Windows CE/Pocket PC).
- Palm as a company is undergoing major changes, and some question its leadership and future business directions

There is an ongoing debate in the industry about the future of Palm. Palm supporters are becoming concerned that the Palm market share is stagnating while Microsoft's Pocket PC is increasing. "The conventional wisdom is that Microsoft is gaining a lot of momentum. But Palm is still the market leader," said Alexander Hinds, the president and CEO of Blue Nomad, makers of the Wordsmith word-processing application for the Palm OS (Costello, 2001). One of the reasons for the sense of impending doom surrounding Palm is Microsoft's push into the enterprise market, an area it has traditionally dominated (Costello, 2001). Many industry experts believe that the Pocket PC will post large gains in the market in 2002 and that Palm will need to work more closely with third-party developers who are already strong in the enterprise to counter that move. Nonetheless, many of the vendors supporting the Palm OS are still confident in the future and believe that the release of Palm 5.0 will be a key in sustaining its market share. Many observers tout the functionality of the Windows CE platform, noting its native support for full-motion video, digital music, and high-resolution color screens as selling points for corporate customers. This situation hasn't been helped by the length of time Palm has let pass between major upgrades to its operating system. Palm OS 5.0, a major upgrade that will mark the platform's transition to the more powerful StrongARM processor, was released in July, 2002. Some of the key benefits to the new version of the Palm OS include the following:

- **Enhanced security—**To keep sensitive data private, Palm OS 5 offers 128-bit data-encryption services based on the de facto standard RC4 encryption algorithm from RSA Security. It also includes end-to-end security that is provided through Secure Socket Layer (SSL) services for e-mail, Web browsing, and online transactions.
- **Multimedia—**The new OS will have the ability to record sound and play CD-quality digital audio. Support for high-density screens (up to 320×320 pixels) doubles the Palm OS screen resolution, and new selectable color themes will let users customize devices.
- **Wireless—**In addition to current support for wide area network (WAN) and Bluetooth, Palm OS 5 supports 802.11b for connections to wireless local area networks.

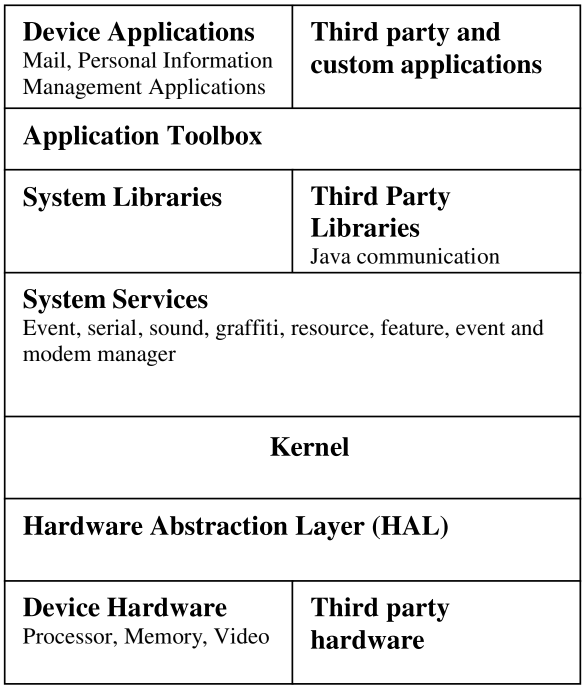

Device Applications
Mail, Personal Information Management Applications
Third party and custom applications
Application Toolbox
System Libraries
Third Party Libraries
Java communication
System Services
Event, serial, sound, graffiti, resource, feature, event and modem manager
Kernel
Hardware Abstraction Layer (HAL)
Device Hardware
Processor, Memory, Video
Third party hardware

Figure 1: Palm operating system architecture.

Primary Components

The Palm OS consists of five main components:

1. Palm OS software
2. Reference hardware design
3. Data synchronization technology for one-button synchronization
4. Platform component tools including an application programming interface (API) that enables developers to write applications
5. Software interface capabilities to support hardware add-ons

Figure 1 shows the overall architecture of the Palm OS.

Windows CE Operating System

Windows CE is an operating system developed by Microsoft for different types of mobile devices made by several leading consumer electronics manufacturers including Casio, Compaq, Hewlett Packard, and Symbol. Windows CE is a modular operating system that provides many building blocks for developers to choose from. It can be large or small, depending on the size and strength needed for a device. It provides users with the familiar look and feel of Microsoft Windows. Windows CE is a 32-bit OS designed to meet the needs of a broad range of devices, from enterprise tools (such as industrial controllers, communications hubs, and point-of-sale terminals) to consumer products (such as cameras, telephones, and handheld and home entertainment devices). A typical Windows CE-based embedded system is targeted for a specific use, and many analysts think that Windows CE should be used for any industry or business applications. Improved kernel services allow the OS to respond more quickly to various processing needs. The improved services make the OS ideally suited for industrial applications such as robotics, test and measurement devices, and programmable logic controllers. The OS interoperates easily with desktop environments that are based on Microsoft Windows NT operating system and Microsoft Windows 2000. This makes it the easier for overall enterprise system integration that combines small mobile systems with high-performance desktops servers and workstations. There are several applications that come with the Pocket PC including Pocket Word, Pocket Excel, and Pocket Internet Explorer, Microsoft money, active sync, note taker, and file explorer. Some of the features found in the OS include the following:

- **Increased Device Security—**enhancements include encrypted passwords and programming interfaces that allow third-party developers to extend anti-virus software.
- **Windows Media Player—**This enhancement delivers the ultimate digital media playback for mobile device users. With Media Player, a user can listen to digital music or watch a short video. The media player supports Windows media content, MP3 audio files, Windows media audio, and Windows media video.
- **Pocket Internet Explorer—**This feature allows users to surf the Web online or download Web pages to read while offline. It supports HTML and WAP sites.
- **eReader—**This feature allows users to read their favorite books on their mobile devices.
- **Pocket Outlook—**This is the mobile companion to Microsoft Outlook. Users can schedule appointments, manage contacts, read, write, and send e-mail.
- **Industry Standard Expansion Slot—**This feature allows users to expand the use of their device as their needs grow. Some users might use an expansion slot to add more memory or to add a modem, digital camera, or scanner.

Symbian Operating System

The Symbian OS is an advanced, open, standard system licensed by the world's leading mobile phone manufacturers. It is designed for the specific requirements of open and data-enabled 2G, 2.5G, and 3G mobile phones. The Symbian OS includes a multitasking kernel, integrated telephony support, communications protocols, advanced graphics support, data management, a low-level graphical user interface infrastructure, and a variety of application engines. Some of the key features of the operating system include the following:

- **Full suite of application engines—**Applications include a contact database, an alarm, an agenda, charts, a word processor, a database, and a help application.
- **Browsing**—This operating system includes the Opera Web browser, which has the ability to browse over global system for mobile communications, general packet radio service, code-division multiple access,

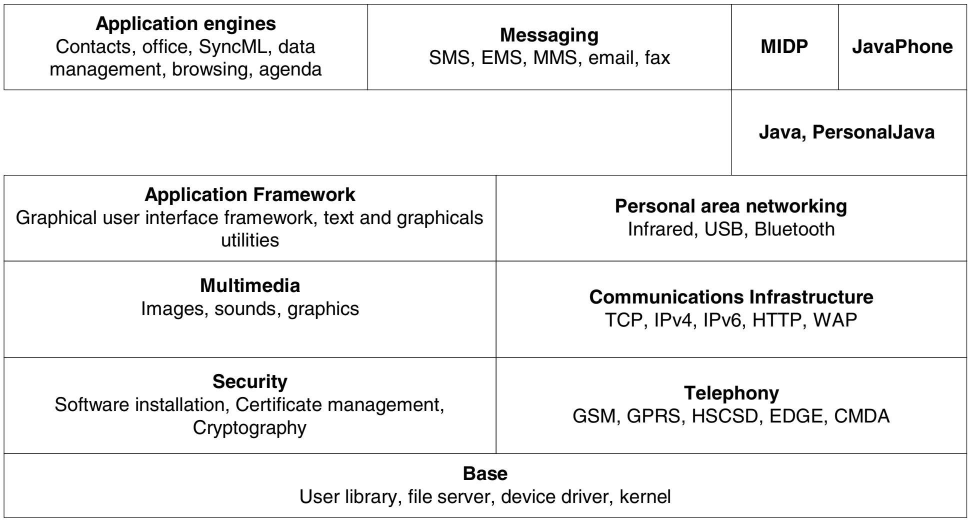

Application engines
Contacts, office, SyncML, data management, browsing, agenda
Messaging
SMS, EMS, MMS, email, fax
MIDP
JavaPhone
Java, PersonalJava
Application Framework
Graphical user interface framework, text and graphicals utilities
Personal area networking
Infrared, USB, Bluetooth
Multimedia
Images, sounds, graphics
Communications Infrastructure
TCP, IPv4, IPv6, HTTP, WAP
Security
Software installation, Certificate management, Cryptography
Telephony
GSM, GPRS, HSCSD, EDGE, CMDA
Base
User library, file server, device driver, kernel

Figure 2: Overview of the Symbian operating system.

and transmission control protocol/Internet protocol connections and to browse local files.

- **Messaging—**This allows users to create, send, and receive text, enhanced, and multimedia messages, as well as e-mail and faxes.
- **Multimedia—**The Symbian OS includes a graphics subsystem that allows shared access to the screen, keyboard, and pointing devices. It also includes an interface that allows for drawing to any graphics device. In addition, it provides audio recording and playback.
- **Communication protocols—**The Symbian OS supports numerous protocols used for communication such as transmission control protocol, user datagram protocol, Internet protocol v4, Internet protocl v6, point-to-point protocol, domain name system, secure sockets layer, transport layer security, file transfer protocol, hypertext transport protocol, wireless application protocol, and Bluetooth.
- **Mobile telephony**—The telephony subsystem provides multiple APIs to its clients,which offers integration with the rest of the OS to accomplish advanced data services. Some of the functionality included is phone and network information, access to the network names detected by the phone, information about the current network, and retrieval of signal and battery strengths and network registration changes.
- **Data synchronization—**The SyncML client is used for data synchronization. It includes a contact and an agenda database adapter so that the two databases can be synchronized. It also includes a database adapter that allows the Symbian client to synchronize its data with the backend database.
- **Security—**The Symbian OS security subsystem enables data integrity and authentication by supporting underlying secure communications protocols such as secure sockets layer and IP security. The security subsystem includes a cryptography module and certificate management module.
- **Software development—**The Symbian OS is delivered to its licensees and development partners in two products, the Symbian OS customization kit and the Symbian OS development kit. This allows other companies to develop software to be compatible with the Symbian OS.
- **Support for multiple user interfaces.**
  The features are shown in Figure 2.

J2ME Operating System

Java was introduced in 1995 with the idea of developing a language that would allow developers to write their code once and then run it on any platform supporting a Java Virtual Machine (JVM). In 1997, a new edition was released, Java 2 Enterprise Edition, to provide support for large-scale enterprisewide applications. The most recent addition to the family is the Micro Edition, which targets "information appliances" ranging from Internet ready television to cellular phones. Following is a summary of the available Java platforms:

- **Standard Edition (J2SE)—**This edition is designed to run on desktop and workstation computers.
- **Enterprise Edition (J2EE)**—This edition provides built in support for servlets, Java server pages, and XML. J2EE is aimed at server-based applications.
- **Micro Edition (J2ME)**—This edition is designed for devices with limited memory, display, and processing power.

J2ME is aimed at consumer devices with limited hardware capabilities. By introducing J2ME, devices no longer need to be static in nature. It allows users the option to browse, download, and install Java applications and content; devices running J2ME can access features inherent to the Java language and platform. J2ME capabilities vary greatly. Products may include cellular phones, PDAs, pagers, entertainment devices, and automotive navigation as well as others. An understanding of configurations and profiles is necessary to understand how J2ME can accommodate a broad range of electronics and devices.

Configurations

A configuration defines a Java platform for a broad range of devices and is closely linked to a JVM. It defines the Java language features and the core Java libraries of the JVM for that particular configuration. What a configuration applies to is primarily based on the memory, display, network connectivity, and processing power available on the device. On the Sun Microsystems Web site, the J2ME Frequently Asked Questions states, “The J2ME technology has two design centers, things you hold in your hand and things you plug into the wall.” Two currently defined configurations are the connected device configuration (CDC) and the connected, limited device configuration (CLDC) (http://java.sun.com/j2me/docs/). The CDC configuration has the following specifications:

- 512 kilobytes memory for running Java
- 256 kilobytes for runtime memory allocation
- Network connectivity, possibly persistent and high bandwidth

The CLDC configuration has the following specifications:

- 128 kilobytes memory for running Java
- 32 kilobytes memory for runtime memory allocation
- Restricted user interface
- Low power and typically battery powered
- Network connectivity, typically wireless, with low bandwidth

Profiles

A profile is an extension to a configuration. A profile provides the libraries for a developer to write applications for a particular type of device. The mobile information device profile (MIDP) defines APIs for user interface components, input and event handling, persistent storage, networking and timers, taking into account the screen and memory limitations of mobile devices.

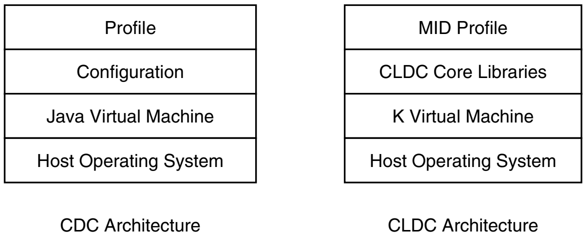


Figure 3: Overview of connected device configuration and the connected, limited device configuration architectures.

J2ME Architectures

The overall architecture starts with the host operating system followed by the virtual machine. The virtual machine can take one of two forms: For systems complying with the CDC configuration, it will be the traditional virtual machine. For systems complying with the CLDC configuration, it will be the K virtual machine (KVM) that meets the specifications as required by the configuration. See Figure 3 for an overview of the CDC and CLDC architecture.

FORECAST OF MOBILE COMPUTING

The future of mobile computing looks promising. Each of the operating systems described in this chapter has its strengths and weaknesses (Table 1).

Mobile Applications

According to Graff (2001), “By the year 2005, 65 percent of Fortune 2000 enterprises will be supporting mobile wireless applications, such as e-mail, Internet/intranet access or customer relationship management.” There are a variety of mobile applications, some aimed at the consumer and others at enterprises. Business needs new mobile applications to reduce costs; improve or increase customer service; satisfy the need for anytime, anywhere; and bridge business partners. For mobile applications to be successful, they first must meet some predefined business need. Once this is defined, the applications and solutions need to be location and device aware to be the most effective.

According to Egan (2001a), there are three applications architectures offline, on demand, and online. With offline applications, the client is able to perform functions. With

Table 1 Mobile Platforms Strengths and Weaknesses

Platform	Keys
Palm	*Strengths:* Market share, number of partners, and number of available applications *Weaknesses:* Somewhat hardware dependent and core operating system functionality is limited
Symbian	*Strengths:* Telephony area, number of partners, Eurocentric *Weaknesses:* Lack of marketing compared with Palm and Microsoft and many companies involved in development
Microsoft	*Strengths:* Integration with existing infrastructure and applications and performance *Weaknesses:* Anti-Microsoft sentiment among the public and no true Java support
Linux	*Strengths:* Open source code, inexpensive, and popular in Asian markets *Weaknesses:* No wide base of applications available, many variations of the core language, no owner

on demand applications, the client is intermittently connected, and some processing can take place on the client. With online applications, the mobile device is always connected, and much of the processing is done on the server. So which application architecture should be used? It depends on the defined business needs and goals, as well as the type of applications that will be implemented and used. Organizations will face various challenges when deciding which mobile technology to use. One challenge will be the human factor, which includes various social characteristics, fashion, and internal perceptions; in addition, many mobile applications will imply changes in human behavior. Another challenge is the business and economics of developing mobile applications. The challenge is to find the right business model because there are so many ideas about mobile applications. A successful application involves working with other business partners that may not have the same goals. As for economics, an operating system must be cost justified and needs the support of management. The final challenge will be with the technology itself. Mobile technology is evolving quickly, and few experts understand everything related to it. Two key areas are overall performance of mobile applications and security.

Generations of Mobile Applications

As mobile technology continues to evolve, so do mobile applications. The first generation of mobile applications was primarily focused on the organizer, which consists of contact and appointment features. These applications are still used heavily in handheld devices. The second generation of applications began to expand their application needs beyond the embedded applications. Many users want access to office and PC-based applications, such as Microsoft Office and e-mail. The third generation of applications will involve the requirement for continuous wireless connectivity to various applications. During this generation, data synchronization will become much more common. The fourth generation of mobile applications will be built based on the assumption that handheld devices are mainstream. This is estimated to occur in 2004–2005. The fifth generation will include applications developed for managing numerous devices and is estimated to become mainstream in 2006. Currently, messaging is the main application that end users require.

Concepts for Working With Mobile Applications

Some important concepts need to be considered when developing mobile applications. Among these is the obvious fact that the user is mobile, which has an impact on the application and content delivered. Mobile applications will more than likely run on a Web phone, a handheld device, a pager, a Web PC, or an information appliance. The four main applications pushing the wireless Internet include messaging, conversing, interacting, and browsing. The way these applications are accessed could be over a wide area network (WAN), a local area network (LAN), or a personal area network (PAN). These networks differ in power, bandwidth, and data rates and are important to define in the development of applications that will be accessed via mobile devices. Users will access or browse sites that keep their content current and easy to navigate.

Wireless Development Method

To develop software applications for wireless devices, a number of different methodologies are followed. Because wireless applications are different from traditional software applications, there are certain factors on which a developer should focus. Before writing the application, one should identify and study the intended audience. The three basic steps in studying the user are (a) developing a persona, (b) describing scenarios, and (c) creating storyboards. A persona is a single, concrete characterization of someone who uses an application. There can be more than one persona for each application, but developers should establish at least one. The creation of a persona involves the identification of the typical personality, writing a background for that personality, and describing characteristics and actions that define it and its typical activities. A persona should have a realistic personality and, at a minimum, should include a name, a title, a picture, biographical information, personality traits, and goals. The first paragraph of the persona's written information should introduce the user without reference to a specific technology. A scenario is a concise description of a persona using technology components to achieve a goal. The scenario considers how the user handles the hardware, how the application is operated, and how the content is used. A storyboard diagrams the entire story of application use, one screen at a time, and shows the display, navigation, and interaction for the wireless device. There are numerous programming languages available to develop applications, but the most popular wireless programming languages are Java, C, and C++.

A developer should also create a mobile database (with content) and design a logical application. The developer lays out a working model, builds wireless screens, and codes business logic to attach to real content. A mobile application should include back office administrative functions, keeping in mind that someone will have to manage the mobile user content on the server. Developing useful tools with good interfaces can help to simplify administrative tasks. Developers should also design a Web start page, which allows a user to set up an account and personalize the mobile application. A key to developing successful mobile applications is to allow users to personalize them to fit particular needs.

CONCLUSION

Wireless technology is an exciting area of study that will evolve considerably in the years to come. Companies seeking to develop such technology should have a business model in place before settling on a given platform; they should also have a strong understanding of the kind of customers they will serve. What type of device will these customers use? What type of content will they want to access with those devices? Developers should focus on simple applications, and, working in small teams, use rapid application development to create prototypes of their products. Because the technology is still evolving, companies in the wireless arena should expect both business and technical challenges. Selecting mature technologies while building flexible, adaptable infrastructures so that new protocols, operating systems, and applications can be added to the existing infrastructure will help developers of mobile technology meet the challenges of the market.

GLOSSARY

Configuration A Java platform for a broad range of devices that defines the Java language features and core Java libraries of the Java Virtual Machine of the particular configuration.

Java 2 Micro Edition (J2ME) An operating system designed for devices with limited memory, display, and processing power.

Mobile information device profile (MIDP) A profile that defines application programming interfaces for user-interface components, input and event handling, persistent storage, networking, and timers, taking into account the screen and memory limitations of mobile devices.

Mobile computing A generalization of all mobile computing devices including personal digital assistants, smart phones, and other wireless communication devices.

Palm OS The most popular platform for handheld mobile devices since 1997 when the first Palm Pilot was launched.

Profile An extension to a configuration that provides the libraries for a developer to write applications for a particular type of device.

Symbian OS An advanced, open, standard operating system that is licensed by the world's leading mobile phone manufacturers.

Windows CE An operating system developed by Microsoft that runs on various Pocket PC devices.

CROSS REFERENCES

See *Mobile Commerce; Mobile Devices and Protocols; Wireless Application Protocol (WAP); Wireless Communications Applications; Wireless Internet.*

REFERENCES

Dulaney, K. (2001, October). *Outfitting the frontline: Phones, PDAs, and strategies for use.* Paper presented at the Gartner Group Symposium, Orlando, Florida.

Egan, B. (2001a, October). *Mobile and wireless computing: The next user revolution.* Paper presented at the Gartner Group Symposium, Orlando, Florida.

Graff, J. (2001). *Wireless e-mail and messaging: You can take it with you.* Paper presented at the Gartner Group Symposium, Orlando, Florida.

Jones, N. (2001a, October). *Mobile commerce business scenario.* Paper presented at the Gartner Group Symposium, Orlando, Florida.

FURTHER READING

Air2Web. (2001). *A roadmap to wireless: The state of the technology.* Retrieved March 15, 2002, from http://www.air2web.com

Beaulieu, M. (2002). *Concepts for working with wireless applications.* Retrieved July 14, 2002, from http://www.informit.com

Broadbeam Corporation. (1999). *Introduction to wireless data.* Retrieved March 28, 2002, from http://www.broadbeam.com

Costello, S. (2001). *Comdex—despite Microsoft, developers stick with Palm—for now.* Retrieved November 16, 2001, from http://www.nwfusion.com/news/2001/1116palmms.html

Deitel, H. M., Deitel, P. J., Nieto, T. R., & Steinbuhler, K. (2002). *Wireless internet & mobile business: How to program.* Upper Saddle River, NJ: Prentice-Hall.

Donston, D. (2002). *Making a place for IM at work.* Retrieved July 14, 2002, from http://www.eweek.com

Egan, B. (2001b, October). *Wireless WANs and LANs: Even superman may be caught by this bullet.* Paper presented at the Gartner Group Symposium, Orlando, Florida.

Evans, N. (2002). *The m-business evolution.* Retrieved July 14, 2002, from http://www.informit.com

iConverse. (2001). *The impact of mobile application technology on today's workforce.* Retrieved March 14, 2002, from http://www.iconverse.com

Jones, N. (2001b). *Mobile e-business case studies.* Paper presented at the Gartner Group Symposium, Orlando, Florida.

Microsoft. (n.d.). *Microsoft Windows powered Pocket PC.* Retrieved March 28, 2002, from http://www.microsoft.com/catalog

Microsoft. (n.d.). *Product overview.* Retrieved March 28, 2002, from http://www.microsoft.com/windows/embedded/ce.net

Muchow, J. (2002). *The basics of J2ME.* Retrieved July 14, 2002, from http://www.informit.com

Nobel, C. (2001). *MIS 2002 only accesses Microsoft apps.* Retrieved July 14, 2002, from http://www.informit.com

Nobel, C. (2002). *Can new Palm OS take on the enterprise?* Retrieved July 14, 2002, from http://www.informit.com

Palm. (2002). *Palm OS 5 preview.* Retrieved March 28, 2002, from http://www.palmos.com

Simons, S. (2002). *SMS: Is it happenin' yet?* Retrieved July 14, 2002, from http://www.informit.com

Symbian. (n.d.). *Symbian OS version 7 functional description.* Retrieved April 2, 2002, from http://www.symbian.com

Synchrologic. (2000). *The handheld applications guide.* Retrieved April 2, 2002, from http://www.synchrologic.com

Synchrologic. (2001). *The CIO wireless resource book.* Retrieved April 2, 2002, from http://www.synchrologic.com

Synchrologic. (2001). *The future of enterprise mobile computing.* Retrieved April 2, 2002, from http://www.synchrologic.com

Multimedia

Joey Bargsten, *University of Oregon*

THE CONVERGENCE OF AESTHETICS AND TECHNOLOGIES

> *Interactive multimedia:* any computer-delivered electronic system that allows the user to control, combine, and manipulate different types of media, such as text, sound, video, computer graphics, and animation. (*Encyclopedia Britannica*, 2002)

Interactive Multimedia

In line with the above definition, this chapter discusses multimedia in its more recent digital, interactive connotations—limiting the discussion to work that is

- mediated by microprocessors or any technology that allows the display and interaction of various combinations of image, text, and sound;
- represented in whole or in part in digital form;
- often nonlinear and user-determined (interactive) in its form; and
- capable of transmission via networks, or published to disk.

In the larger context, however, multimedia refers to an integration of multiple art forms, as in "intermedia," a term first proposed by writer and experimental artist Dick Higgins of the Fluxus movement (fl. 1960–1973). Throughout the 20th century, intermedia artists created an aesthetic and conceptual framework, which integrated image, sound, text, theatrical performance, space, movement, traditional media such as painting and sculpture, and projected/transmitted media (still projection or slides, motion projection or film, video, television, radio). Intermedia included various combinations of media, performance, and concept, but was not necessarily built on a digital foundation. However, it is within this setting that interactive multimedia, and indeed digital culture at large, emerged.

Although many communicative forms since the printing press proposed new combinations of art and technology, it is distinctly digital, interactive multimedia that blurs the line between author and user. Reinvigorated by the technologies of computation, telecommunications, and audiovisual representation during the latter part of the 20th century, multimedia is now commonly regarded as a communicative form that realizes the possibilities of user-determined, computer-mediated, nonlinear, digital content, and its potential delivery across networks. In the minds of many, multimedia evolved to a digital art/design form as well—with even the once avant-garde notions of chance operations, nonlinear narrative design, nonsequitur of words and phrases, image collage, and all manner of digital synthesis (from sounds and voices to 3D virtual worlds and computer-generated actors) finding comfortable—if not always mainstream—acceptance.

Multimedia now suggests a "digital commons" or vast community of multimedia authors and users—in business, the arts, communications, education, government—all interconnected via the Internet or other electronic means. Users are accustomed to increasingly complex interfaces with fusional, user-friendly designs, whether the multimedia is a digital art installation, a video game, or a Web site delivered via computer or a wireless device. The continual development of multimedia authoring tools such as Illustrator, Flash, Director, and ProTools, as well as digital media technologies like Quicktime and Real Media, is an integral part of meeting both user and author expectations.

THE AUTHORING PROCESS

Multimedia authoring places new demands on everyone involved in this integrative process. Multimedia project managers would do well to possess a broad knowledge, not only of design principles and trends, but also the tools and techniques to implement them. Moreover, in projects utilizing such accessible applications as Macromedia Director, the team of designers/artists is often expected to double as programmers. Even in environments insistent upon dividing up the functions of the multimedia staff, the designers should nevertheless grasp the subtle transformation of media elements through simple, and not so simple, use of code, whereas the programmers should learn to discriminate between the various design styles that result from their code manipulation.

Role of Design

In terms of design principles, the multimedia author(s) should be highly sensitive to the users' experience of the project. In particular, the author focuses on the project's intended visual style, the balance of its various media elements, and its information design.

Visual Style

The "look" of a project contributes greatly to its overall tone and effectiveness. Visual design can be examined along the polarities of

- approach to imagery, from photorealism to abstraction;
- density of visual elements, from sparse to cluttered; and
- overall organization of visual elements, from orderly and balanced to spontaneous, even chaotic. (See Figure 1.)

Balance of Media Elements

In cohesive multimedia projects, the visual design "echoes" throughout the project's writing, organization, programming, and sound. The individual media elements—although each is given equal attention—should ultimately create an entirety. Any afterthought will be evident to the discriminating viewer.

Many artists achieve this cohesion with widely disparate elements, such as Dziga Vertov's film *Man with a Movie Camera* (1928), an early example of montage in cinema—or the editing together of contrasting shots. This approach translates to the world of multimedia in such animations as *Mumbleboy* (http://www.mumbleboy.com), with its beautifully oblique, non-sequitur imagery.

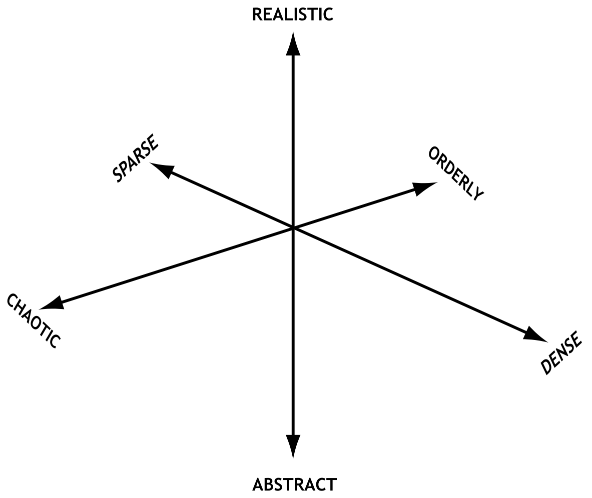


Figure 1: Design polarities.

Information Design

The audience's experience of the project is greatly enhanced by

- the consistency of actions and interactivity, such as buttons that always do what users expect;
- the logical sequence of events, such as the user logs in and arrives at the main interface page or the homepage—the deeper the user goes into the project, the more clicks it takes to get back out;
- the visually consistent navigation, always in the same location on the page—the visual look of the navigation reinforces the hierarchy of the information;
- the unambiguous directions, including dialogue boxes that avoid confusing language; and
- the project's adaptability to individual users—prompting them, for instance, to enter their own demographics or adjust control panels.

Information design is "the art and science of preparing information so that it can be used by human beings with efficiency and effectiveness" (Horn, 1999). Like many aspects of multimedia, information design draws upon multiple disciplines: psychology, library science, cartography, marketing, information science, visual design, etc. However, as a good starting point in the creation of a multimedia project, authors often ask themselves the following questions:

- Who will most likely benefit from, enjoy, or respond to my information/project?
- What am I trying to convey with this information?
- What led me to create this project?
- What do I hope to accomplish?
- Given a likely audience, what is the best delivery of the information (printed publication, Web, DVD, etc.)?

By answering these questions, the multimedia author clarifies the audience, tone, background, purpose, scope, and depth of the project.

Role of Testing

In testing or auditioning the multimedia project, the designer uncovers shortcomings in the interface's design and functionality, as well as determines the portability of the project across various networks, hardware configurations, and operating systems. In presenting the project to representative audiences, as well as complete novices to the subject at hand, the designer answers such questions as

- Does everything function as the designer expects?
- Does any aspect of the design get in the way of the user's expectations?
- Were assumptions made in the design of the project that are not understood by the intended user?

Understanding Digital Data

File Size and Compression

Digital data is stored in files characterized by file type (text, image, sound, etc.) and file size. The file size is affected by the amount of data, the data's quality (resolution and bit-depth, for example) and the data's compression (if any). The technology that compresses and decompresses data is generally referred to by the acronym "codec." After data are compressed then transmitted, it is decompressed for its intended use. Usually, a higher level of compression yields data of a smaller file size, but lower quality.

Far-reaching changes in digital culture followed the development of better codecs, capable of delivering a high-quality audio or visual experience while maintaining a small or reasonable file size. For example, the .mpeg-1, layer 3 audio codec—popularly known as MP3—involves compressing CD-quality music data to file sizes small enough for easy transmission via the Internet, making music-sharing sites like Napster.com an enormous success. After the dismantling of the Napster site following litigation brought against it by the Recording Industry Association of America (RIAA), other file-sharing sites emerged such as Gnutella.com and Kazaa.com.

QuickTime, Windows Media, .mpeg, and Real Systems are key technologies that develop or incorporate multiple codecs for different types of media. Each of these platforms offers various specialized codecs—for example, photo .jpeg for still images; MP3 or Qdesign for audio; Sorenson, QuickTime OfflineRT, or Indeo for video; and .mpeg-4 for video streaming. While most codecs are software-based, some—like certain versions of the Indeo video codec—require specific hardware. A complete listing of currently available codecs and their application to multimedia data can be found at http://www.discreet.com/support/codec/ (Discreet, n.d.).

Color Models

The proper use of color models, or color spaces, is crucial in maintaining color consistency in a multimedia project. In most digital media, either the additive (RGB) or the subtractive (CMYK) model is used as the basis for the project's color. In the additive model—used for screen projects such as Web or video—its basic color components of red, green, and blue light create pure white light when the three colors are combined in equal amounts. The subtractive color model (cyan, magenta, yellow, black [K]) is best suited for print technology and is based on inks or pigments rather than light. Thus, in creating a Web site, an author must adjust high-quality photo scans, .gif and .jpeg graphics, and Flash elements so that they occupy the same RGB color space. If images made in Flash for Web animation are later included in a print publication, the author must convert the Flash material from RGB to CMYK. See Figure 2.

Process Overview

Input

To take advantage of digital technologies such as reproduction, editing, transmission, and networking, the multimedia author must convert the project data to digital form. Various technologies exist for the conversion of analog data into digital form: scanning and digital photography for images; optical character recognition (OCR) for

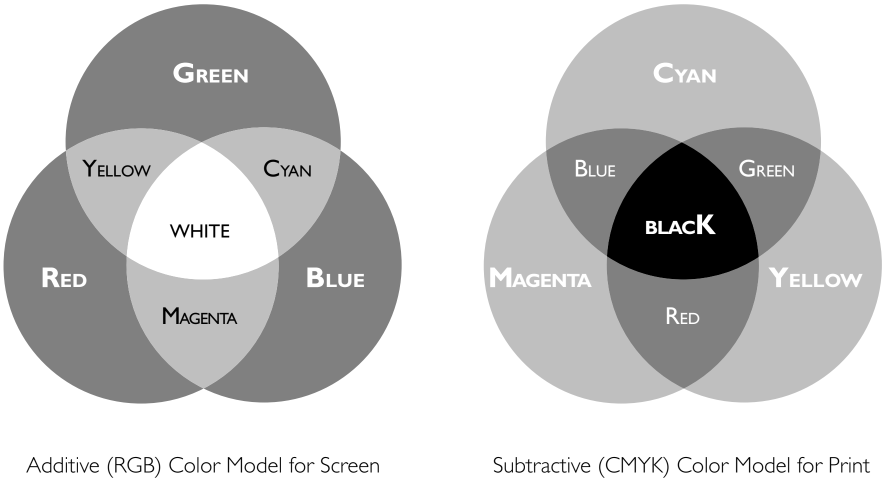

GREEN
YELLOW
CYAN
WHITE
RED
BLUE
MAGENTA
Additive (RGB) Color Model for Screen
CYAN
BLUE
GREEN
BLACK
MAGENTA
YELLOW
RED
Subtractive (CMYK) Color Model for Print

Figure 2: Color models.

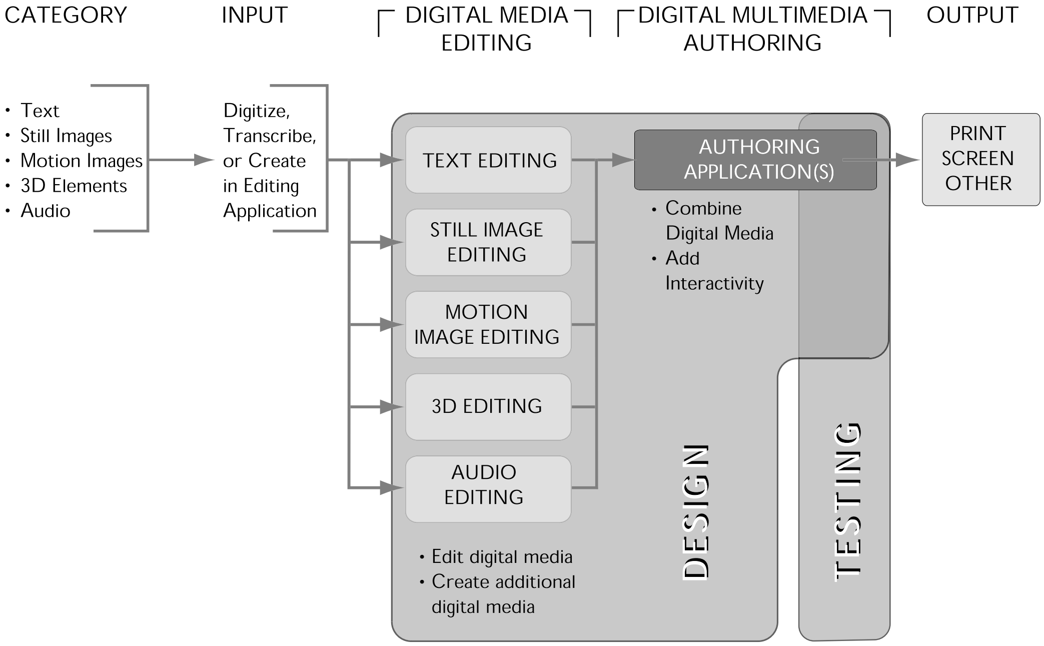


Figure 3: Digital multimedia—process overview.

text; recording, sampling, a-to-d (analog to digital) conversion for sound; and telecine and video capture for motion media (see Figure 3).

Once "digitized," the data can be manipulated within a multimedia application (for example, scanning a photo of the *Mona Lisa* into Adobe Photoshop and using Photoshop's pencil tool to add a mustache). Although dedicated applications emerged during the early 1990s for the processes of digitizing and editing media, many recent applications include both processes.

Editing and Authoring Applications

In multimedia, various media elements such as text, images, sounds, animations, and video are first created in an editing application suited to a particular media type; then the elements are combined and arranged in an authoring application that incorporates the various media elements with interactivity, usually in the form of a scripting language. (Note that while this chapter approaches editing and authoring as distinctly different environments, some multimedia applications that comprehensively integrate both functions have evolved.)

Linking/Embedding/Nesting

The media elements—text, images, sounds, animations, video—relate to the project file through linking, embedding, and/or nesting:

Linking. Linked elements are not incorporated within the project file, but links to these elements are established within the project file. The project file and all of its linked elements must be present for the project to play, display, or print properly.

Embedding/Nesting. An embedded media element is one copied in its entirety into the project file, usually making the project file size much larger. Once embedded, multiple references to the same media element in the project file can be made, usually at little or no increase in file size. Elements embedded within other elements, which are embedded in yet other elements, results in nesting. Nesting could involve, for example, animations of a circle graphic—getting larger and smaller—referenced respectively within the "over" and "up" functions (or states) of a button; multiple copies of this button are then referenced on a navigation strip, and multiple copies of the navigation strip are referenced on an interface (see Figure 4). In the prior example, in which all of the animations are vector-based—created within a single authoring tool that is vector-oriented—the file size and download time is minimized, contributing to greater economy of visual design.

If the multimedia author uses an application with scripting capabilities, the resultant programming has a nested structure as well. In the project file, the author starts with simple independent commands, which are called (referenced or gathered up in a larger action) by higher-level commands. In Figure 5 below, "handlers," or a series of commands in Macromedia Director's scripting language, are called within other handlers.

Output

The authoring application creates a project file that incorporates, and in some cases further edits, the media elements. The project file also provides the technology, or "engine," necessary to publish or play the project as a single, self-running, or standalone application from a disk or hard drive. Otherwise, the author publishes the project file without the engine, making the project smaller in file size and easier to transmit over the Web or other networks.

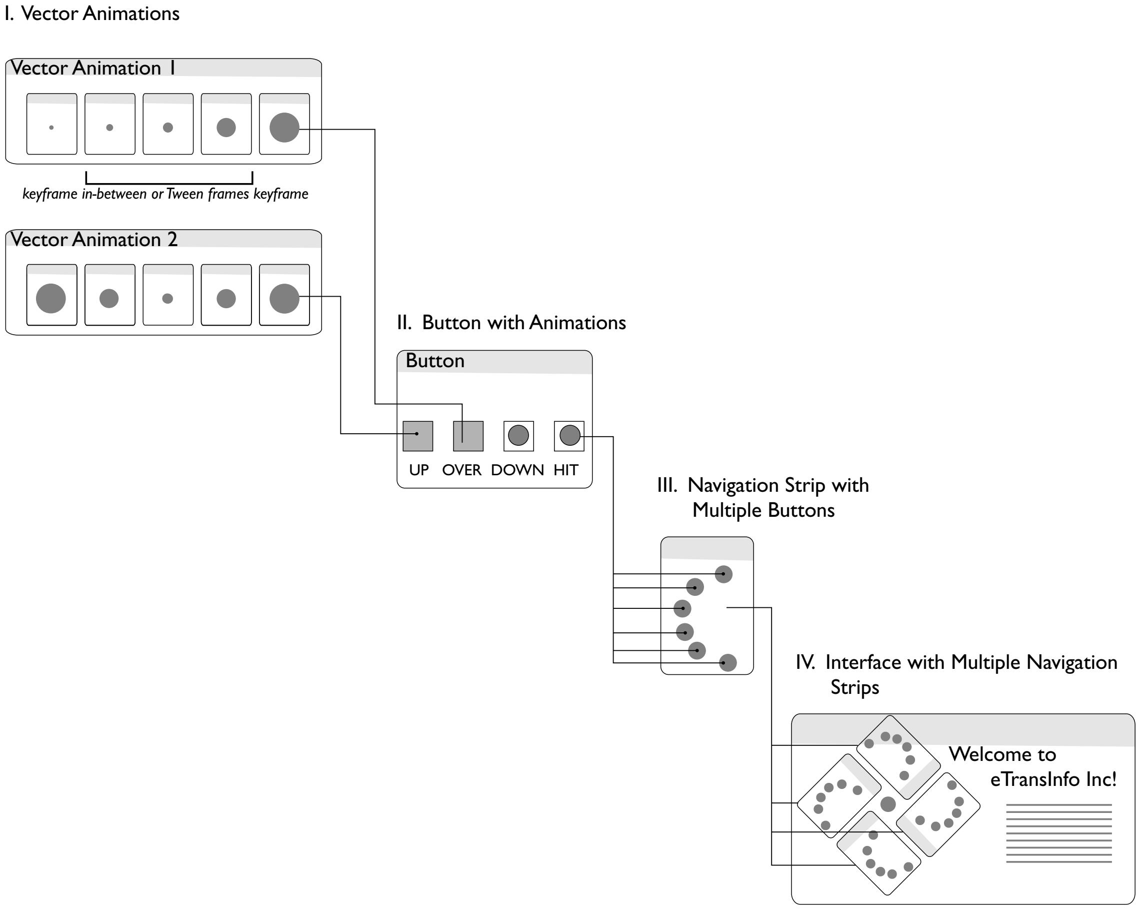

I. Vector Animations
Vector Animation 1
keyframe in-between or Tween frames keyframe
Vector Animation 2
II. Button with Animations
Button
UP
OVER
DOWN
HIT
III. Navigation Strip with Multiple Buttons
IV. Interface with Multiple Navigation Strips
Welcome to eTransInfo Inc!

Figure 4: Nesting.

In this case, the project is played on a Web browser that has the engine technology built into it in the form of a browser plug-in. The project file can also be published or exported as a file format that can be further imported into other authoring applications. (See Figure 6.)

THE SPECIFICS OF MULTIMEDIA DATA, ITS REPRESENTATION, AND ITS KEY EDITING APPLICATIONS

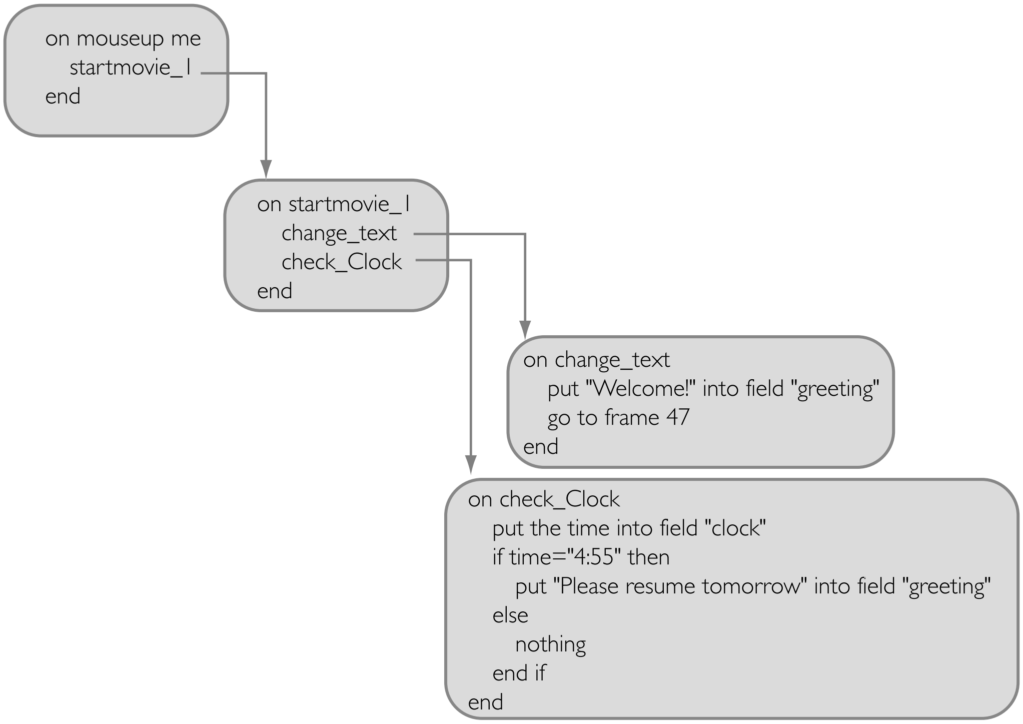
Text

Indicative of the power and pervasiveness of text in multimedia production is the slogan of Voyager, the hypertext and CD-ROM publishing company founded by Bob Stein: “Text: the final frontier.” Text is perhaps the most supported of all media types; it can be edited in virtually all the major editing and authoring tools and is commonly cut and pasted between applications (see Figure 7). A central component in most multimedia projects, text is unique in its ability to engage the senses of readers, creating characters, events, and entire sensorial worlds in their imaginations.

In multimedia production, text is generally presented as

- body text—a means of conveying information or instructions, in a continuous or linear form, or arranged discontinuously;
- display text—a visual design element—as is usually the case with titles; display text can also be the visual foundation of a work, as in the Juxt Interactive Web site (http://www.juxtinteractive.com) or in the Shredder Web site (http://www.potatoland.org), an example of digital art in the tradition of Dada and Lettrisme text experimentation;
- dynamic text—an element of user input or interactivity; for example, users enter departure time and destination on a travel Web site to immediately elicit/display information on airlines and hotel accommodations; and
- hypertext—a link to any other page or part of a page on the Internet or in an interactive project.

Although text-based works tend to be read linearly or from start to end, there were early literary experiments in nonlinear text, anticipating the user-determined presentations of the digital era. William Burroughs engaged in such an experiment, in the late 1940s, when he developed the process of the *cut-up:* he cut pages of his stories into four equal sections, then reassembled the sections



```
on mouseup me
   startmovie_1
end
```

```
on startmovie_1
   change_text
   check_Clock
end
```

```
on change_text
   put "Welcome!" into field "greeting"
   go to frame 47
end
```

```
on check_Clock
   put the time into field "clock"
   if time="4:55" then
      put "Please resume tomorrow" into field "greeting"
   else
      nothing
   end if
end
```

Figure 5: Nested scripting.

into new pages, subverting his control over the order and outcome of events (ArtMuseum.net, n.d.b). Likewise, digital multimedia users are accustomed to creating their own path through text, clicking on links and moving from one interface to another (somewhat unpredictably from the designer's standpoint). This nonlinearity is just one of text's adaptations to a digital setting. E-mail, for example, revived written correspondence, making it ubiquitous, and making its transmittal ever faster with newer forms such as instant messaging. E-mail, Web chatrooms, Web diaries or logs ("blogs"), and other text-based digital communicative forms are changing how we communicate—shaping language into more compressed, fragmented, and nonlinear form.

Still Images

In visual communication, the image provides a symbol or reference for more abstract words or concepts. Thus, the image is perhaps the most volatile and potentially ambiguous of media forms—owing to the tremendous influence of visual information from fine arts, advertising, communication design, and other cultural sources. In the hands of the digital artist, the image is yet more endlessly transfigurable.

In multimedia, bitmaps and vectors are the two major types of still images (see Figures 8a and 8b). Whereas bitmap files represent the photos manipulated, or material drawn, in bitmap applications, vector files represent imagery drawn in vector applications. In bitmap editing, Adobe Photoshop is the standard editor, as it has dominated the market since its release in 1990. "Photoshop" is often used as a verb, meaning to alter a photographic image through digital means—although Photoshop performs other functions besides bitmap editing (see Figure 9). In vector image editing, Adobe Illustrator and Macromedia Freehand are currently the dominant tools (see Figure 10).

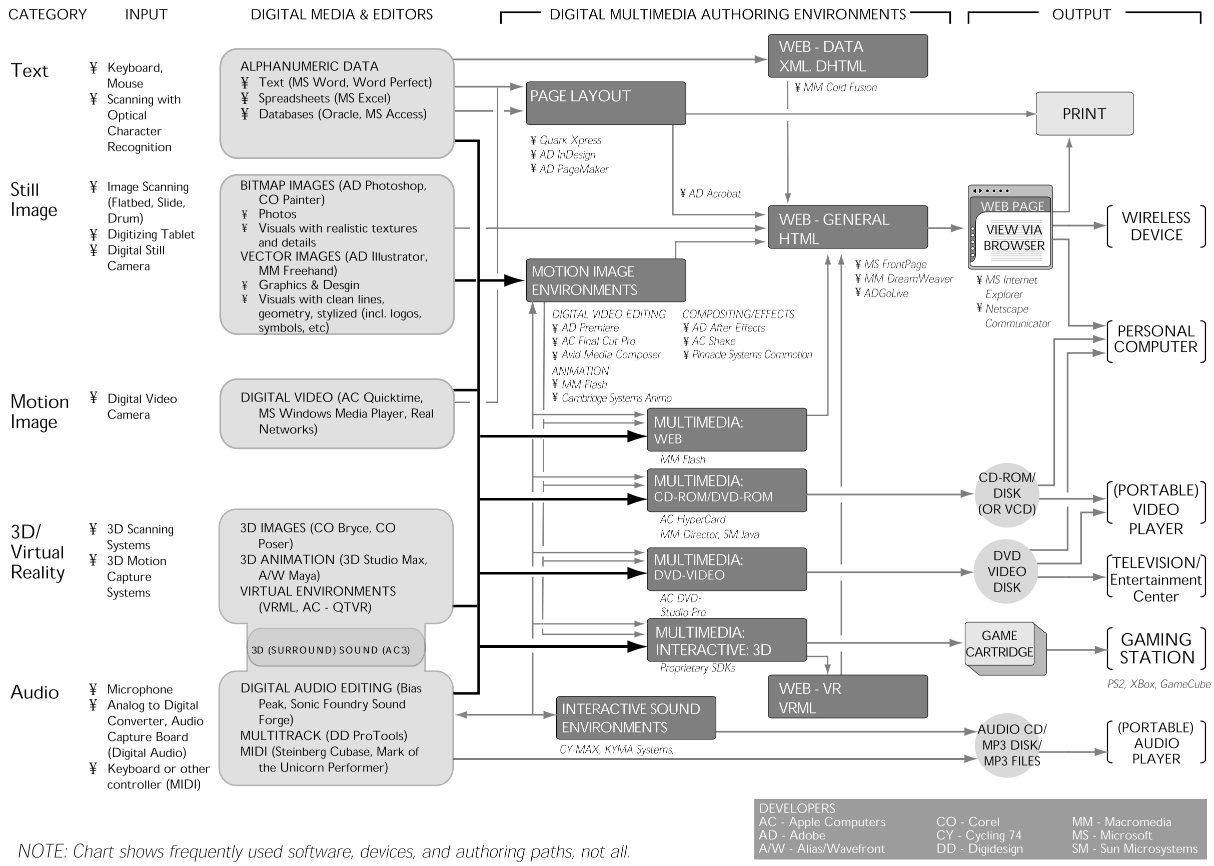
Bitmap Editing

The bitmap editing tools of Photoshop and similar applications include the following.

Painting and Drawing. Tools to add pixels to a digitalized image or a blank digital canvas typically employ a drawing or painting metaphor. A pressure-sensitive pad and handheld stylus (pen) often replace the mouse pad/mouse as the preferred method of input, attempting to make the creation of visual images more accessible. For example, Corel Painter was used in an interactive art installation in San Francisco's Zeum Children's Museum. The exhibit encouraged children to draw on electronic easels and view their artwork projected on large screens.

Image Selection Methods. Using a digitized image as a starting point, certain tools select the area of the image to be manipulated. Common among these tools are Photoshop's select-all command, the lasso tool (drawing around a selection), the marquee tool (dragging a rectangle or oval around a selection), and the pen tool (to draw a Bezier curve).

The alpha channel is a more advanced approach to selection. This tool creates a grayscale version of an image, in which the black pixels constitute the selected area, the



NOTE: Chart shows frequently used software, devices, and authoring paths, not all.

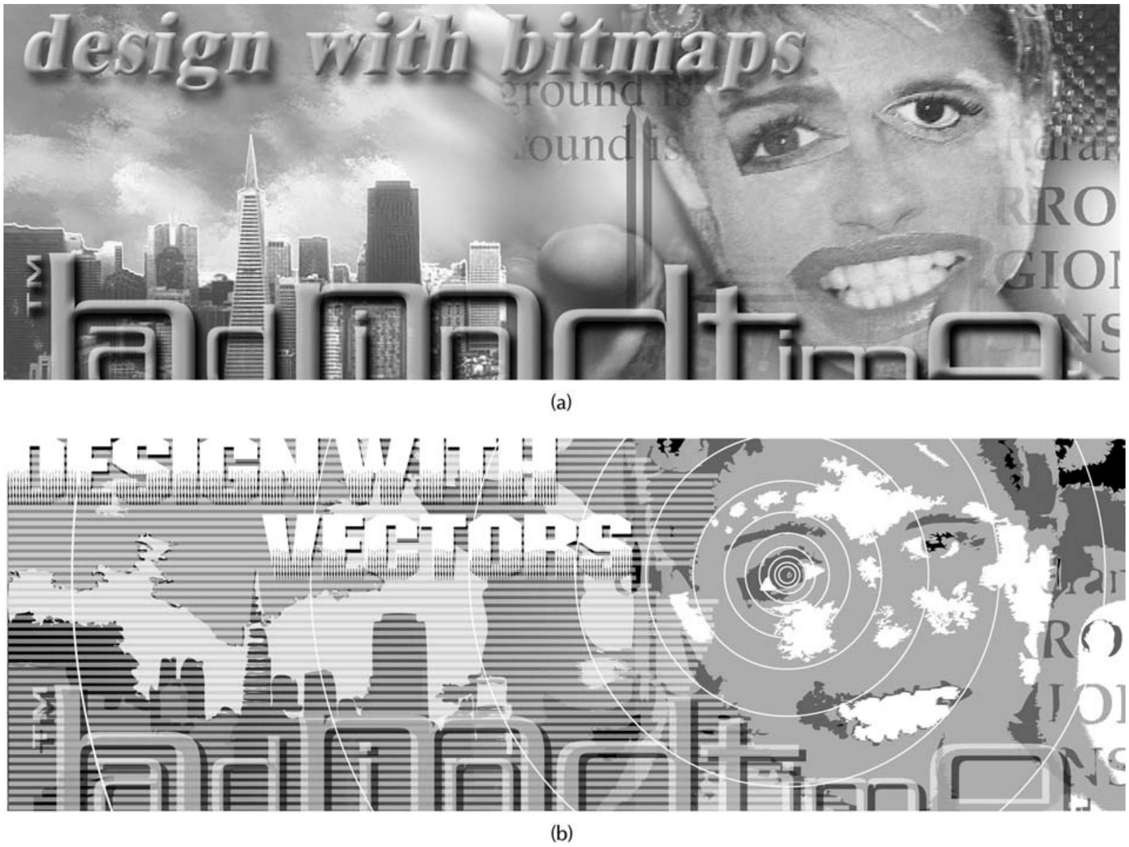
Figure 6: Digital multimedia—detail.

white pixels are unselected, and all shades of gray are partially selected. Selections can be imported from other programs, as well. This tool allows for complicated image selection and manipulation, such as accurately masking the outline of a person's hair against a visually "busy" background.

Cut, Copy, and Paste Operations. Once selected, the selection can be cut, copied, and pasted onto other parts of the image. Variations on these elementary operations are included in most bitmap editors—such as the rubber stamp or custom brush tools that are a hybrid of painting and copy/paste tools. Any selected area can be copied and

Media Type	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
Text	Microsoft Word 98	**.doc**	text only (**.txt**);	As is	Information Document	Use for "Read Me" or "Read First" informational documents when delivering applications via disk; general purpose cross-platform documents.
			Rich Text Format (**.rtf**);	As is	Formatted Information Document	For informational documents with more formatting preserved (tabs, indents, bold, italic, type size)
			.doc	As is	Text Document in MS Word	Standard editable text document with formatting; requires compatible versions of software.
			.doc	MS Power Point	Text for Presentation	Copy/paste in some instances more productive than import.
			.doc	Import to Adobe Acrobat	Export from Acrobat as Portable Document Format (.PDF) file	Preserves all formatting including font (typeface). Standard self-contained document for web delivery when consistent formatting is essential. Editable when opened within Acrobat full version.
			.doc	Link or import to Page Layout	Desktop Publication (DTP); layout integrating text, image and typography	Import or link as story in Quark, PageMaker, or InDesign. Most complete professional control over design elements, typography. Many output options, including .pdf files to deliver to service bureau or printer.
			HTML	Open in HTML editor	HTML page	For bringing large text documents into HTML for web delivery; suggest opening in true HTML editor for greater control and visual consistency.

Figure 7: Text/data.



(a)

(b)

Figure 8: Design with (a) bitmaps and (b) vectors.

made into a custom brush or rubber stamp in order to paint parts of the image onto other parts of the image or canvas.

Image Adjustment. Many aspects of an image can be altered or adjusted, such as brightness, contrast, color saturation, color balance, or color palette. Colors can be matched to a standard system of colors, such as the Pantone system, used throughout print and electronic industries.

Layers. Layering is another basic technique/feature common to most digital image programs. It involves placing one image on top of another, while controlling the amount of transparency or opacity of each image—providing for the building of multiple layers, each with its own imagery, graphics, or text. A subgroup of layers known as adjustment layers incorporates image adjustments like brightness/contrast, color saturation, balance, and visual definitions of shadow and highlight. The user can produce multiple versions of an image, with subtle differences in color balance, for example, so that each version suits a specific printer or screen.

Filters and Effects. The filters and effects tools offer a vast range of visual possibility. Bitmap images are altered, sharpened, blurred, pointillized, shattered, smoothed—taking on the characteristics of physical material (glass, plastic, wood, canvas, water)—or transformed into stylized, partly abstracted visuals, using the extreme sharpen, posterize, motion blurs, and displacement map features of Photoshop and similar programs.

Vector Editing

In contrast to the textured photorealism of bitmap graphics, vector art is the basis for the clean, sharp lines of graphic design, logos, packaging, signs, or vectorized images. Vector editing is dominated by such applications as Adobe Illustrator and Macromedia Freehand, unless the intent is animation; in which case, Macromedia Flash is the tool of choice. (See Vector Animation below.) Vector editing toolsets are similar to those in bitmap applications. However, because vector editors describe space, line, point, and color in mathematical terms—rather than in individual picture elements—vector images are displayed with equal crispness and acuity at any magnification or size. Vector shapes retain their clarity and scalability, while being pulled, twisted, and altered, because a vector image is made up of endpoints and the curves connecting them (Bezier curves).

Bitmap to Vector Converters

These tools—referred to as trace bitmap functions in such programs as Flash and Corel Draw—involve the conversion of bitmap images to vector-based graphics. A stylized, less realistic image results from this changeover, a process reminiscent of “posterization”—breaking an image

Media Type	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
Bitmap Image	Adobe Photoshop 7.0	**.psd**	Photoshop (**.psd**).	DVD Studio Pro	DVD Menu	Layers are preserved for roll-over and down state of DVD menu buttons. Create at 160-300 dpi at 720X540 pixels; then adjust image (proportions not constrained) to 720X480 pixels at 72 dpi for import into DVD Studio Pro.
			PICT	DVD Studio Pro	DVD Slideshow images; still menus.	Create at 160-300 dpi at 720X540 pixels; then adjust image (proportions not constrained) to 720X480 pixels at 72 dpi for import into DVD Studio Pro.
			BMP, PICT	Import into MS PowerPoint	Presentation	Create images at 160-300 dpi at size you wish them to be in your presentation, then adjust resolution down to 72-96 dpi (proportions and size constrained)
			PICT (16-bit)	Import into Macromedia Director	16-bit photorealistic graphic for multimedia authoring	Delivers 16-bit images in multimedia authoring program that are almost 1/4 the file size of comparable 24-bit BMP images.
			Photoshop EPS or TIFF	Page Layout Programs	DTP Publication	Preserves clipping paths and alpha channels in current versions of page layout software. Suggest 300 dpi or greater for high quality output.
			Save for Web:**GIF**	Link to HTML page	Web Graphic, 8-bit color	Create final image at size you want displayed on web page at 72 dpi. Recommended for graphics with solid colors and low bit-depth, not recommended for photorealistic images.
			Save for Web: **.JPEG**	Link to HTML page	Web Graphic, 24-bit color	Create final image at size you want displayed on web page at 72 dpi. Recommended for photorealistic images. Variable quality & file size possible; use quality above 30% only when image quality critically outweighs file size in importance .
			Save for Web: **.JPEG**	Import into Macromedia Flash	Photorealistic web graphic for Flash interactivity or animation	Use judiciously, as file size will balloon enormously with each .JPEG image. Useful as a single background graphic, for instance.
			Save for Web: **.PNG-8, PNG-24**	Link to HTML page	Web Graphic, 8-or 24-bit color	Saves as web graphic in .PNG format, which has image quality arguably superior to both .GIF and .JPEG, but is not universally supported in all browsers.

Figure 9: Bitmap still images.

Media Type	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
Vector Image	Adobe Illustrator 10.0	**.ai**	**.ai**	As is.	General purpose vector art and design.	For graphics and illustrations that demand a high degree of visual clarity and detail, and which tend to be more stylistically conceived than photorealistic in detail.
			Adobe PDF ;	Import to Adobe Acrobat	Export from Acrobat as Portable Document Format (.PDF) file	Useful for design-intensive small projects that incorporate text, graphics, typography (posters, announcements, CD covers, etc.). For export to web to preserve all aspects of design and formatting, or to service bureaus or printers for output.
			Illustrator EPS	Page Layout Programs	DTP Publication	Preserves clipping paths in current versions of page layout software.
			Common Bitmap file formats: GIF, PNG-8, PNG-24. BMP (BMP),), JPEG (JPG), Macintosh PICT (PCT), PCX (PCX), Photoshop 5 (PSD), Targa (TGA, TIFF (TIF),	- -	- -	Although possible to export vector art in these formats, the vector art becomes rasterized or converted into bitmap form, which elminates the advan tages of vector formats: small file size and sharp image quality at any magnification.
			Flash (SWF)	Import to Flash	Scaleable still-image graphic for Flash animation, interactivity or design.	For preliminary design for interactive or animation project in Flash. Import may not be successful on complex images that include multiple gradients or gradient meshes; copy and paste between Illustrator and Flash are also possible.
			Flash (SWF)	Upload to website, link directly to .swf file	Scaleable still-image graphic	Useful in displaying single-image vector art on the web.
			Windows Metafile (WMF)	Import into MS PowerPoint	Presentation	To import vector design elements with transparent backgrounds into presentations
			AutoCAD Drawing (DXF), AutoCAD Interc hange File (DXF),	Import into 3-D or CAD based application	Still vector model for further animation or modeling	For importing vector design elements into 3-D or CAD environment

Figure 10: Vector still images.

into irregular shapes according to the image's color content. This technique is only effective when using a limited color palette; otherwise, it produces a large number of complex vector curves and, thus, high-bandwidth file size—not recommended for slower computer processors.

Vector to Bitmap Converters

To manipulate vector art in a bitmap editing program, it is converted or rasterized into bitmap form. It can be cut, pasted, edited, and filtered like a bitmap, but it cannot be reconverted to the original vector art. Corel Draw and Flash are often used in this conversion.

Motion Images

Certain ambiguities surround the terms motion images, motion graphics, and animation. For this chapter, motion images refer to any image that moves, and motion images include motion graphics, animation, and video. Common to both motion graphics and animation is the process of setting a drawn figure into motion. The difference between the two is the figures: in motion graphics, they are typically text or geometric shapes, forms, or designs, and, in animation, they are generally drawings of characters (animals, humans, etc).

Motion images in multimedia are difficult to discuss outside the context of their creation in animation environments (see below). However, still image bitmaps and vector art—starting points for motion images—are often first manipulated outside the animation environment. For example, bitmap images saved as individual layers in Photoshop are imported into motion image environments like Adobe After Effects in order to activate them. Similarly, vector images created in Illustrator or Freehand are easily imported into Flash as the foundation for Web animations.

Video—although primarily edited in digital video environments—is often included under the umbrella of motion images, because individual clips of video are sometimes considered motion images to be painted on, filtered, and layered in some stylistic, painterly manner. After Effects or Pinnacle System's Commotion are well suited for this.

It is equally challenging to discuss motion images without speaking of the influences of the film and television industries, which have so informed the look of contemporary motion images. A key example are the titles for the film *Se7en*, designed by Kyle Cooper, immediately adopted as the "look" of the mid-1990s—disrupted, blurry, and shaking frenetically—appropriated by a generation of multimedia students. From the earliest animated cartoons, such as Windsor McKay's *Gertie the Dinosaur*, animated figures entered our cultural consciousness, only to emerge en masse when animation applications became available for the desktop.

3D in Multimedia

If the fine arts are an archetype for the abstracted look of an interactive Web site or motion graphics piece, then the entertainment industry provides the model for the specialized area of 3D design, inciting multimedia authors to create realistic representations of object, character, and place. Some companies that ultimately informed multimedia production are George Lucas' Industrial Light and Magic (the *Star Wars* series, *Terminator*, etc.), Pixar (*Toy Story, Antz, A Bug's Life*), and Sony Pictures Imageworks (*Stuart Little, Spider-Man 2*). The proliferation of feature-length 3D animations and motion picture special effects—as well as the explosion of commercial 3D video games—inspires independent multimedia authors to create this type of experience. This is not limited to the visual world: the 3D listening experience of 5.1 Surround Sound, originally introduced in full-length films, is now accessible in more generally used cost-effective 3D authoring environments.

Some specific examples of 3D in multimedia are the following:

- 3D elements incorporated into 2D graphics—such as realistically rendered buttons or design elements on a Web site (i.e., buttons that look like actual buttons or a doorknob with three dimensions);
- 3D data representation such as informational charts and graphics—often inserted into business presentations created in another authoring environment (i.e., Microsoft PowerPoint);
- walk-through interfaces or "virtual reality" tours of buildings, including interactive links on panoramic photos of locations;
- interactive in-the-round product demonstrations, including proof-of-concept virtual models (a premanufactured prototype of a product);
- animated figures—whether in a motion graphic, interactive game, or TV commercial;
- wireframe art—using only the outline, or wireframe, of a 3D figure or object;
- 3D animated figures in conjunction with more high-end applications like the motion-capture systems of Vicon (retrieved 20.iv.03 from http://www.vicon.com/); for example, a motion capture system scans people performing various actions, tracking the 3D motion of their joints with specialized sensors; the data for these actions are transferred to a 3D authoring environment, such as Alias/Wavefront's Maya, for figure animation; and
- high-end interactivity created with higher-level authoring platforms; for example, the user's or "active immersant's" experience—though primarily created in Maya or similar programs—is extended with 3D head sets, goggles, or joysticks linked to specialized controllers created in higher-level authoring platforms like C++, or in proprietary software developer kits (SDKs) for specific game platforms (PlayStation, Nintendo).

3D Production Process

In 3D programs such as Maya or 3D Studio Max, the process for creating 3D objects, characters, or settings is summarized as follows.

- Generally, 3D objects or settings are "sketched" as a series of basic geometric shapes, or "primitives," such as spheres, cubes, cylinders, cones, and polyhedrons. Since they are vector shapes, the author edits them in terms of their Bezier curves; however, with 3D, the curves extend into three dimensions. Simultaneous views of an object—front, side, top—are available in most 3D programs.

- To assemble the edited primitives, the author determines the motion of their various parts. If building a human figure, for instance, the author approximates the rotation of an elbow or knee.
- The author creates "texture maps," which are detailed 2D bitmap images of a particular physical texture (such as fur or corrugated aluminum) applied to figures, objects, and scenes to provide life-like effects.
- The author designs/programs the lighting and camera movements (panning, tracking, dolly, crane, zoom). In a 3D animation, the author employs Bezier curves to map the path of the camera through space (see Vector Editing).
- At the end of this process, the author can export 3D material to an interactive application like Macromedia Director or, for interactive Web delivery, Macromedia Flash. The 3D material can also be imported into proprietary or high-end programming environments to create interactive games, kiosks, or other products.

Software such as 3D Studio Max, Maya, and SoftImage—while creating rich and complex 3D worlds—place high demands on computer hardware. Processor power and speed are considerations, as are disk space, memory, and rendering time—the time needed to create, process, and output each frame of an animation or effect. Often, even modest 3D projects require powerful video or graphics cards for a high-quality render in a reasonable period of time.

Beyond hardware and software considerations, 3D authors face a somewhat steep learning curve. Using software like Softimage or Maya, the author is involved in every aspect of 3D design. Authors with less experience, or limited resources, have the option of using 3D tools such as Corel Bryce (landscapes, settings, weather, and modeling), Corel Poser (human figures), and Swift 3D (geometric elements for inclusion in Flash-based Web design). These programs, while not as full-featured as Maya or 3D Studio Max, are still useful in creating 3D elements for integration with 2D design in a motion graphic, interface, or short animation.

Audio

The use of recordings as a means of experimenting with sound began with the composer Edgard Varese, who created collages of music and sound in the 1920s and 1930s, recording on multiple disks and, later, tape recorders. The exploration of music and sound—through the cutting, splicing, and layering of tape—was called "musique concrete," with echoes now in contemporary techno and hip-hop. Still another turning point in the manipulation of electronic sound was the first commercially produced electronic synthesizers, created by Robert Moog and others in the 1960s. The Moog analog synthesizer ultimately led to the digital instruments of the 1980s—specifically the establishment of MIDI (musical instrument digital interface), the standard protocol for connecting digital music hardware and software. Various digital audio editing software soon followed, as well as the development of multitrack recording/editing environments, which allow the designer to layer multiple audio tracks, much as graphic designers layer images.

Functional Categories of Sound

Music, background sound, and effects, as well as spoken text and dialogue, all have an enormous impact on the overall success of a multimedia project. Regardless of the level of the multimedia production, its sound elements contribute greatly to the user's overall immersion.

The most common functions of sound in the multimedia environment are listed below, followed by the tools—both software and hardware—needed to produce and manipulate sound elements. See Appendix C for further examination of textures and sound design organization.

Cue Effects. These sounds correspond to a particular visual element or action in a Web interface or interactive application, cuing the user, for instance, to the location of buttons or rollovers. The user receives a signal that is clear and immediate, yet unobtrusive. More sophisticated interactive work builds subtle variation into each instance of a button sound, often providing the user with a more sonically organic experience.

Background Sounds and Effects. Background sound or "room tone"—like the clinking of silverware and plates in a diner—is perhaps most effective when strongly suggested at the beginning of a scene, then reduced in dynamic or absent as other functional elements move to the foreground (like the actors' voices or music).

Background effects are usually linked to specific events or actions. In a linear work (an animation, a motion graphic build, a digital movie clip), an effect accompanies and amplifies a particular action or event in the unfolding of the narrative or imagery.

Background sounds and effects are more cohesive if placed within the same "sonic space." In a scene that takes place in a cavernous cathedral, for example, all of the sound effects need varying adjustments of reverberation to give them all an equally spacious quality. Sonic space also refers to characteristics of ambience, echo, delay, and equalization, applied uniformly to each sound, achieved by applying digital signal processing (see DSP below).

Narration and Dialogue. Narration and dialogue include live-action, voice-over, over-dubbing, or ADR (automatic dialogue replacement). For best results in voice recording, an actor is recorded in a controlled sonic environment—ideally, a sound studio where various applications optimize the voice's sonic elements. Often a high-quality, voice-optimized microphone such as a cardioid is used. Compression–expansion hardware reduces the level of ambient noise while boosting the signal of the voice.

Musical Score. When incorporating music into a multimedia project, the major considerations are likely the relative density, complexity, or texture type of the musical elements (acoustic, instrumental, vocal, or electronic). A rock song, for example, is often composed of slowly changing electronic harmonies (ambient), driving acoustic drum and bass tracks (rhythmic), and a solo sung by the vocalist or played by the lead guitar (solo/foreground).

Media Type	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
digital audio	bias peak VST v. 3.0	.aif or .sdii	.aiff .wav	other sound editors, multitrack editors	digital master audio file for multiple applications	For creating master audio files. Unless storage space is an issue, save uncompressed stereo files sampled at 44.1khz in 16 bit sound, OR at 48kHz in 24 bit sound for professional audio environments.
			.aiff .wav	digital interactivity (Flash, Director)	digital audio file for multimedia authoring.	Unless storage space is an issue, save uncompressed stereo files sampled at 44.1khz or 48khz in 16 bit sound. Authoring programs tend to apply compression during the publishing process (usually MP3)

Figure 11: Digital audio.

Sound Production Tools

After sounds are recorded live, or created electronically (digital audio editors, MIDI sequencing, synthesis or sampling, etc.), they are further processed in digital audio editors and/or combined in a multitrack environment, optimizing them for a multimedia project.

MIDI (Musical Instrument Digital Interface). MIDI is the standard file format or language for digital electronic instruments. To use the MIDI protocol, the author must have MIDI equipment—previously marketed only as dedicated hardware, but now incorporated into other authoring applications like QuickTime.

The music performance data—played on a MIDI controller such as keyboards, drum pads, or MIDI wind instruments—is transcribed in terms of which notes are played, at what time, by which instrument, as well as many other nuances of performance. In contrast to sound files like MP3, MIDI data files do not contain actual recorded sounds, and, thus, are extremely small. These data are stored and edited by means of the MIDI software editor known as the sequencer, and then ultimately played back through the MIDI hardware or software modules that convert the MIDI data into actual sound.

Digital Audio Editors. These editors manipulate sound files, recorded by way of microphones, and digitized via sound-capture technology. These editors are well suited to a number of basic tasks, including cut/copy/paste functions in time scales down to the millisecond. Most applications offer a batching option, allowing the user to perform the same operation on multiple files—a great convenience when, for instance, the media artist is downsampling, normalizing, and applying digital signal processing to numerous files. Some examples of digital audio editors are Sonic Foundry Sound Forge and Bias Peak VST (see Figure 11).

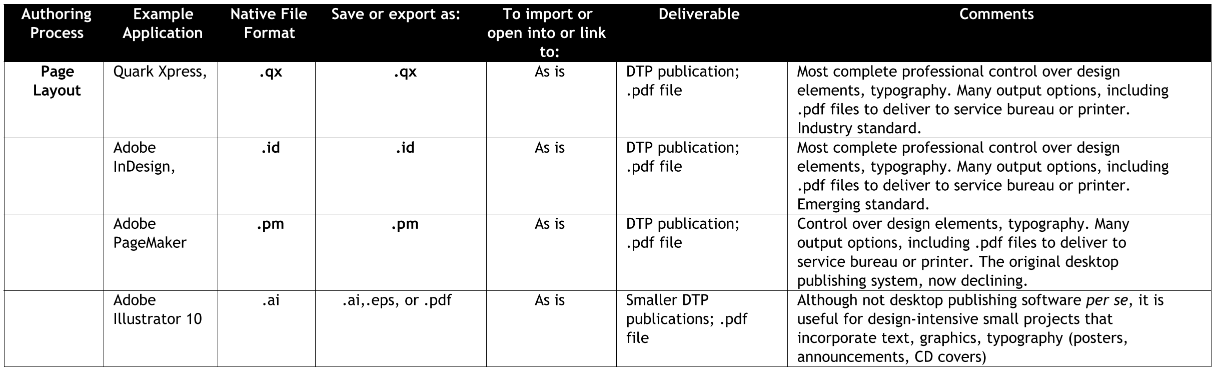
Multitracking. Multitracking is a basic approach to combining multiple sound files—digital audio files, as well as MIDI data synchronized to the multitrack software. Multitracking creates a musical score, or combines dialogue, background sounds, and cue or Foley effects into a soundtrack, all in a postproduction environment. The process assigns tracks to the sound input and adjusts such parameters as the equalization of the sound's output as well as its position (i.e., left and right for stereo; left, right, front, and center for Surround Sound). These tracks are layered, and/or digital signal processing is added to them individually or in groups (see DSP below). The ProTools systems by Digidesign are widely used multitrack tools (see Figure 12).

Digital Signal Processing (DSP). DSP is an additional functionality available to some digital editing, multitracking, and MIDI sequencing applications. Although the filtering of images (as discussed above in Bitmap Editing) involves digital signal processing, this process is generally only referred to as DSP when working with audio. Using formats such as Microsoft DirectX, VST (Virtual Studio Technology), AudioSuite, and Digidesign TDM (Time Division Multiplexing), the DSP plug-ins emulate sound processing—including reverb, chorus, delay, gate, compression, normalize, fade in/out, and others—previously only available through dedicated audio studio hardware.

Loop Generators/Editors. Loop generator software produces rhythmic and ambient textures, often employing a screen metaphor—visually and functionally—of a rack of sound gear (sampler, drum machine, mixer, synthesizer, sequencer, and bass). Popular with dance clubs DJs who operate the loop generators on laptops, this software also exports music mixes in common digital audio formats such as .aif, MP3, or .wav for use in other multitrackers

Media Type	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
multitrack sound	ProTools	project file + linked media files (usually .aif or .wav)	.aif , .wav, Quicktime	other sound editors, multimedia editors, digital video editors	digital audio file for multimedia authoring.	Most authoring applications accommodate stereo files and export mixed or compressed mono or stereo.
			6 mono .aif or .wav files, or stereo files (to be later split into mono files)	AC3 application	Dolby 5.1 compliant surround sound	Additional software will be required to mix multiple channels spatially.

Figure 12: Multitrack digital audio.



Authoring Process	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
Page Layout	Quark Xpress,	**.qx**	**.qx**	As is	DTP publication; .pdf file	Most complete professional control over design elements, typography. Many output options, including .pdf files to deliver to service bureau or printer. Industry standard.
	Adobe InDesign,	**.id**	**.id**	As is	DTP publication; .pdf file	Most complete professional control over design elements, typography. Many output options, including .pdf files to deliver to service bureau or printer. Emerging standard.
	Adobe PageMaker	**.pm**	**.pm**	As is	DTP publication; .pdf file	Control over design elements, typography. Many output options, including .pdf files to deliver to service bureau or printer. The original desktop publishing system, now declining.
	Adobe Illustrator 10	.ai	.ai,.eps, or .pdf	As is	Smaller DTP publications; .pdf file	Although not desktop publishing software *per se*, it is useful for design-intensive small projects that incorporate text, graphics, typography (posters, announcements, CD covers)

Figure 13: Page layout.

and multimedia authoring applications. Examples of loop generators/editors are Propellerheads' Reason and Sonic Foundry Acid.

Image-to-Sound Generators or Vice Versa. Applications such as U&I Software's MetaSynth take a PICT image file and treat it as a "score," playing the visual elements in sonic terms. The author can draw or paint lines, shapes, or other visual elements in the image window, and MetaSynth renders and plays the result in stereo. The author saves the score as a .pct file for further visual editing, or as an .aif audio file for playing, which as might be expected from a program that inhabits the dual worlds of image and sound.

Alternatively, sound can be visually represented by applications such iTunes on the MacOS and WinAmp Media Player. These graphic generators provide a kaleidoscopic wash of colors and visual textures in synch with the music.

Video Jam Software. Video jam software is equal parts music application and video controller. It manipulates and plays multiple clips of digital and live video in real time, accompanying the music played at dance clubs or concerts. Principal software includes Vjamm, developed for members of the media ensemble Cold Cut, as well as U&I Software's Videodelic. Computer keyboard commands, mouse motion, and MIDI data trigger such effects as video color cycling and rotation, kaleidoscope, and superimposition. The software reads MIDI data for the parameters of the video playback, as well as plays the various MIDI sound modules.

Hardware for Sound Production

See Appendix B for a list of hardware in setting up a small sound production studio.

INTEGRATION OF MEDIA TYPES: AUTHORING APPLICATIONS

Some well-known integrative applications are discussed below. They are listed in ascending order from authoring applications that combine the fewest media elements—desktop publishing applications, HTML editors, music interactivity software—to applications that combine virtually all media elements, ranging from HyperCard to Director to DVD Studio Pro.

Desktop Publishing Applications

These applications perform page layout and formatting to create a final printed product—typically integrating text with numerical charts and graphic images imported from other applications (see Figure 13). Examples of desktop publishing software are Adobe PageMaker, Adobe InDesign, and QuarkXPress. With their relative ease of use, these digital applications, indeed, brought document production to the desktop, while creating a heightened enthusiasm for typography and print design, ultimately fueling electronic design.

HTML Editors

Some basic HTML editors are Notepad (PC) or BBEdit (Mac). Users of these editors employ HTML code—the basic programming language in the creation of Web sites—and Web browsers such as Internet Explorer decode the HTML code and display the Web site. HTML editors, thus, facilitate the insertion of text or other media elements into Web pages. Using appropriate plug-ins, the code also links to other media elements such as Adobe Acrobat .pdf files. Interactivity is coded into the Web site with the use of compatible coding languages such as JavaScript, Java, and XML—provided the viewing browser has plug-ins for these. For less-experienced Web designers, Web content editors such as Macromedia's Dreamweaver and Microsoft's FrontPage allow the designer to create, edit, and incorporate Web content, working entirely from the screen (what you see is what you get), while manipulating minimal, if any, HTML code.

Motion Image Environments

Motion (or moving) images in multimedia began with repetitive bitmap animations, a succession of still images. Motion image authoring, now seamlessly built into interactive applications, exploits the possibilities of combining motion graphics, animation, and video.

Bitmap Animations/Motion Graphics

Animated .gifs (Animations or Motion Graphics)

In the early days of the World Wide Web, motion was generally limited to the animated .gif format. Typically, these motion images are first manipulated in a bitmap editing programs, then exported as .gif files to applications with animating capabilities like Director or Flash, and exported as an animated .gif. Usually no more than a series of bitmap images, animated .gifs tend to loop incessantly, which give many Web pages, ironically, a static quality. Primarily used in banner advertising and as "eye candy," they've become a source of irritation to Internet audiences, but were, however, put to ingenious use by Web artists like Joan Heemsker and Dirk Paesmans, the experimenters behind Jodi.org (ArtMuseum.net, n.d.a).

Animated .gifs are a declining media due to Macromedia Flash's advances in vector motion images and interactivity. However, bitmaps are central to such industrial-strength animation software as Animo by Cambridge Systems, which incorporates aspects of both bitmap and vector editing. The Animo system scans and arranges individual drawings, creating bitmaps that make up the frames of an animation. The animator then applies vector-based outlining and paint to each image.

Compositing and Effects

The highly effective layering and compositing tools in motion graphics and animation applications are often used to combine live-motion video with more stylized or special effects; thus video is included in the discussion of motion image applications. The following techniques and their applications are common in the editing and compositing of motion images.

General Purpose Compositing. With applications like Adobe After Effects, the user changes any aspect of the image (size, perspective, color, location in frame, filtering, etc.) using the application's timeline to determine the rates of change. Specifically, while using a system of keyframes, the user sets a beginning and end state, and the application fills in all states in-between. (See Figure 4.) Many layers of multiple "effects" can be incorporated in this manner. Whether starting from a still or motion image or video, After Effects produces a composite project file or "comp" that previews the project. The comp is then rendered into a series of bitmap images and exported to a motion format such as QuickTime, in order to play in real time, or to include in other projects.

High-End Compositing. Apple Computers' Shake software leads the next generation of compositing/special effects applications for bitmap motion images, as evidenced by numerous Academy Awards for special effects in major motion pictures that used Shake. Rather than a timeline/keyframe approach, Shake is an icon-driven, object-oriented effects generator with a more extensive tool set. In this process, the author determines effects for each icon. The author then scripts the variability of the effects within each icon, so every instance of the icon has a random range of effects. This results in a sophisticated yet more organic experience.

Go to http://www.apple.com/shake/ to view Shake-enhanced films winning Academy Awards for Special Effects.

Bitmap Rotoscoping. Rotoscoping entered the digital world with a number of applications such as Pinnacle Systems' Commotion Pro and Synthetik Software's Studio Artist. In rotoscoping, the designer draws or paints on each bitmap or live motion frame, giving the realistic images a more stylized, painterly, or cartoon-like look, but retaining the life-like motion. To be played as a sequence, these files are exported to common motion formats like QuickTime.

Vector Animations/Motion Graphics

Vector animations revived the cartoon and short animated film genre, realizing them anew in digital terms, while making the production of animation and motion graphics more accessible to all. While several vector animation and motion graphics programs now exist (Live Motion by Adobe and ToonBoom by Toon Boom Technologies), the market still belongs to Flash, with an installed base of over one million authors.

See http://www.wired.com/animation for examples of short Flash animations (retrieved 20.iv.03).

In Flash, the user explores the full range of vector-art in motion. The user can create frame-by-frame animations or motion graphics, or nest them within other animations/motion graphics. All are small in file size and easy to deliver over the Web, or easily exported to digital video or other motion media. Flash creates, for instance, streaming animations, so that part of the animation plays while the remainder of it loads into the user's random access memory. Interactivity is attached to any visual element—an unavailable feature, at least to this extent, with other motion image applications.

With Flash's vector rotoscoping capability, a series of bitmaps or a video clip is first imported, and then "hand-painted" with vector art. The bitmap series or video is then deleted, leaving the vector art layer behind, with extremely life-like motion. (See http://www.sfdt.com/xiao-xiao/.) Alternatively, the "trace bitmap feature" in Flash or other software automatically converts bitmap to vector.

Richard Linklater's animated film Waking Life (2001) is an exceptional example of using both rotoscoping and trace-bitmap techniques. He shot the film using the Sony TRV-900 camera (a low-end professional camera), with the film then digitally stylized into Flash-like animation through software developed by Bob Sabiston, and digitally rotoscoped by over 30 artists.

Digital Video Editing

Digital video is popular with students and small businesses as well as large companies and production houses for television and film. Many film and TV editors—as well as independent digital film directors such as Steven Soderberg (*Full Frontal*, 2002) and Mike Figgis (*Timecode*, 2000)—opted to shoot and edit on desktop video systems. For the full-length feature, *Cold Mountain*, the director Anthony Minghella used a suite of three desktop video editing systems at one third the cost of a comparable dedicated system (Crabtree, 2002).

Capabilities

As of this writing, many digital video editing applications are able to

- capture or digitize video footage into easy-to-handle units or clips for editing;

Media Type	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
digital video	Final Cut Pro Adobe Premiere	DV Stream	.pict or .bmp	import to image editing or page layout program	frame capture from video	Interlaced video can create artifacts, especially on objects in motion, which will require additional image editing.
			.mov (Quicktime)	Quicktime Pro	Progressive download or Streaming Video for web	Can be hinted for variable bit-rate streaming (i.e., performance quality scaled to internet connection speed). Recommend Sorenson III codec.
			.mov	Flash	Video for Web	For short -form video clips (2 minutes and under). Recommend Sorenson Spark or Squeeze codecs.
			.mov	Director	Video for CD-ROM or DVD-ROM	For long -form video clips to be played back from within the Director Executable or Projector. A variety of codecs appropriate (MPEG 1 or Sorenson 1 are arguably the best for cross-platform delivery)
			DV Stream MPEG-2 encoding	DVD Authoring (DVD Studio Pro)	DVD Video	For short - and long-form video to be played back through a standard commercial DVD player.

Figure 14: Digital video.

- log, document, and label footage for automated capturing;
- specify in and out points of edits (beginnings/ends of shots);
- generate transitions between scenes, such as fades;
- layer multiple clips, often incorporating compositing and other motion image tools;
- further composite motion graphics and animations and combine them with live motion;
- mix soundtracks and perform other sound editing; and
- export to a variety of formats with various compression schemes—from low-bit-rate streaming video for the Web, to medium-bit-rate video for interactive applications (CD-ROM), to high-bit-rate video for DVD (see Figure 14).

Technologies

Some notable digital video technologies are as follows:

QuickTime. In 1991, Apple Computer's QuickTime media platform was one of the first technologies to pave the way for video production on the desktop. Although the image produced was not much larger than a postage stamp, QuickTime was a breakthrough technology, the first to attempt playing many types of digital media. Its current version, QuickTime 6—built around the .mpeg-4 international standard—supports over 50 types of digital media.

Digital Video on the Web. In the mid 1990s much research centered on the development of codecs, reflecting the emergence of the World Wide Web and the desire to transmit video via the Internet. Better codecs—like Sorenson and Cinepak—facilitated the transmission of video data at more acceptable rates. Even while transmission speeds and codecs improve, the visual quality of compressed video—for the Web for instance—still does not surpass video on a standard DVD disk.

Media 100. In 1995, the Media 100 system presented an integrated hardware/software solution for desktop video editing, a viable alternative to dedicated digital editing systems. Media 100 produced full-frame, full-motion digital video with affordable video capture hardware, suitable for high production environments. The current industry standard for high-production, Avid Media Composer, is similar in design to the Media 100.

The DV Chain. A portable and cost-effective solution for digital video editing was introduced in 1999 with the miniDV tape format. The DV chain or process begins with video digitized via a camera, rather than through an additional capture board in a computer. Then a tool like Adobe's Premiere or Apple Computer's Final Cut Pro edits the footage, which is then printed back to tape via the miniDV camera. This approach was possible with the development of a high-speed transmission protocol known as IEEE 1394 (iLink [Sony] or Firewire [Apple]).

Digital Video Editing for Film and HDTV. In conjunction with software like Cinema Tools, current digital video editors like Final Cut Pro edit HDTV and 16 and 35 mm film, as well as digital video. After the HDTV or 16/35 mm film is digitized in a lab via telecine, Final Cut Pro edits it and generates a cut list or edit decision list (EDL). An EDL is a list of cuts, transitions, motion effects, and titles, which is then printed back to film or HDTV.

Audio Interactivity Applications

Max (Cycling '74 Software) may be the ultimate in integrative audio applications. Not only does it combine multiple sounds, music tracks, and visual elements, it incorporates interactivity, possibly to a larger degree than any other music application. Originally developed at IRCAM—the venerable electronic music research center in Paris—its inventors used light, heat, and motion sensors to trigger sound production from Max-built sound generators and MIDI devices. The composer has seemingly endless choices in interactive sound sources (light and heat sensors, toasters, live animals). Max uses traditional analog synthesis modules, emulating them digitally, and allowing the user to engineer sound via a flow-chart. Although not for the novice sound designer, Max promises a rich and immersive sound experience.

HyperCard

HyperCard is one of the first integrative multimedia authoring applications for authors with little or no programming experience. Created in 1986 by Bill

Authoring Process	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
Interactivity, high bitrate (Disk & Kiosk Delivery)	Macromedia Director 8.5	**.dir**	**.pict or .bmp**	page layout or bitmap image editors	still image from animation or interface	For documenting an animation or interactive work at a specific magnification (screen shot).
			publish as .dcr	Link to HTML page	interactivity or animation for web delivery	Secondary deliverable of Director. Although generally larger file sizes result compared to Flash, Director for the web retains its robust processing of bitmap, interactivity, and sound elements.
			create projector (.exe)	- -	stand-alone application or player (projector)	Primary deliverable of Director. Enables robust processing of long-form video, vector animation and interactivity, and interactive imagery and sound.
			publish as Java class	linked to HTML page	Java-based interactive project	Suggest extensive testing, as not all Director interactivity will function when translated to Java.
			export as .mov	digital video editor (Final Cut Pro)	Animation or Linear media to be incorporated into digital video	For converting animations to video; for integration into digital video editing environment, compositing with live motion, output to DVD, etc.

Figure 15: Director.

Atkinson, HyperCard effectively integrates image, motion, sound, text, and programming codes—creating standalone multimedia. Using the card-stack metaphor, the author assigns a card to each media element and its associated links to other media elements or scripts. The eventual stacks of cards are all contained within the application. Although still in use today as a rapid-development and prototyping tool for application design, HyperCard's feature set falls behind SuperCard and MetaCard, two other authoring programs sharing its card-stack approach.

Director

Marc Canter created Video Director in 1987—later renamed MacroMind Director, and now Macromedia Director. Its metaphor is a movie production, with casts (media elements), scripts (using the scripting language Lingo), actors (later called sprites), and a stage (the application's window)—upon which all action eventually takes place in a standalone project. With developments in desktop video (QuickTime technology), Director supported video by 1993. In early 1996, Macromedia released Shockwave, a browser plug-in allowing Director projects to be played over the Internet. Although still not ideal for video on the Web, Shockwave laid the foundation for highly interactive Web experiences, supporting bitmap animation, sound, text, and graphics (see Figure 15). Shockwave brought the complexity of Director-based projects to the Web, such as those with multiple streams of music, narrative, and interactivity (see http://www.badmindtime.com).

Macromedia Flash quickly appropriated Director's spotlight on the Web, but Director is still the preferred authoring environment for disk delivery (CD-ROM and DVD-ROM) and for kiosks. The current version of Director supports 3D interactive objects as well as multiuser interactive Web environments ranging from chats to multiplayer online games. With its solid support of video

Authoring Process	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
Interactivity, low bitrate (Web & wireless Delivery)	Macromedia Flash MX	**.fla**	**.pict or .bmp**	page layout or bitmap image editors	still image from animation or interface	For documenting an animation or interactive work at a specific magnification (screen shot).
			.ai	page layout or vector editors	still image from animation or interface	For using the imagery of an animation or interface in a broader number of professional design or print applications: packaging, promotional, publications.
			animated .gif	link to HTML page	animated .gif element	For exporting animation or motion graphic elements for inclusion in web projects for older, non-Flash enabled browsers (user base declining)
			.swf	link to HTML page	interactivity or animation for web delivery	Primary deliverable of Flash. For creating interactive web interfaces, advertising, and animation. For fixed size in a browser window.
			.swf	direct link within a browser	interactivity or animation for web delivery	For creating interactive web interfaces, advertising, and animation. Size of project scales as user scales browser window; some additional functionality (i.e., calling javascript) may be lost.
			.swf	play through Flash Player	interactivity or animation for web delivery	For delivering interactive web interfaces, advertising, and animation on disk or independent of a network connection.
			.swf	import into other authoring program (Director)	Flash-created graphics, interactivity components or animation	For bringing vector-based elements already created in Flash into Director. Very useful for repurposing media already created for web to disk or kiosk projects. Communication between .swf elements and Director project is seamless. Multiple .swf elements and multiple videos playing in the same window will be very processor-intensive.
			.mov	digital video editor (such as Final Cut Pro or Adobe Premiere)	Animation or Linear media to be incorporated into digital video	For converting vector animations to video; for integration into digital video editing environment, compositing with live motion, output to DVD, etc.

Figure 16: Flash.

Authoring Process	Example Application	Native File Format	Save or export as:	To import or open into or link to:	Deliverable	Comments
high bitrate video	Apple DVD Studio Pro	--	Multiplexed VIDEO_TS files	Opens in DVD playing application or commercial player	DVD-Video Disk	Incorporates MPEG2 encoded video, PCM or AC-3 encoded audio, subtitle tracks, along with .PICT or .PSD image files (used for interactive menus or slideshow images)

Figure 17: DVD authoring.

on disk, and its accommodation of a wealth of file types, Director maintains its user base.

Flash

In 1996, Macromedia acquired a small vector-based animation program called Future Splash as a lower-bandwidth complement to the often weightier and primarily bitmap based Director. Future Splash became Flash and—while initially limited in its scripting capabilities—it has supplanted Director as the dominant tool for authoring animation, motion graphics, and interactivity for the Web, with an estimated 96% of current Web browsers with the Flash plug-in (Cantrell et al., 2002) (see Figure 16). With Versions 4 and 5 of Flash, a more robust scripting language—ActionScript—was incorporated. ActionScript—a relative to the HTML-friendly language JavaScript—became a quick favorite with both Web designers and programmers, producing the fabled first generation of Flash artists such as Yugo Nakamura (see http://www.yugop.com) and Joshua Davis (see http://www.praystation.com).

With the current version of Flash (MX, 2002), video support was added, which had long been a Director advantage. Flash's partnership with the video codec developer, Sorenson, coinciding with the rapid growth of broadband users, set the stage for Flash's Web-based video streaming system, without the licensing fees or server enhancements associated with the major video streaming platforms. This feature, however, has yet to penetrate the market perhaps because Flash is most effective with short two to three minute Web videos. (See http://www.madonna.com.)

DVD Studio Pro

The DVD, which stands for Digital Video Disk, is perhaps the most integrative delivery option for multimedia users. Apple Computer's DVD Studio Pro's icon-driven approach to DVD authoring guides the author through importing and linking media elements, attaching scripts and interactive buttons for each, and/or alternate soundtracks, camera angles, subtitles, etc. Among other media elements, DVDs incorporate audio and video of the highest quality. Because of the versatility of DVD Studio Pro and other DVD authoring environments, DVDs play on both commercial DVD players (DVD-Video) and personal computers (DVD-ROM) (see Figure 17).

DVD-Video

DVD-Video is for play on commercial DVD players. Disks made through such applications as DVD Studio Pro conform to the DVD-1.0 universal standard, supporting high-quality .mpeg-2 encoded digital video, multiple AC3 Surround-Sound audio tracks, and up to 32 language subtitle tracks.

Although the user control of DVD-Video is limited to the buttons on a remote control, the DVD can support slide shows, motion menus, multiple camera angles, alternative versions, and Web links and other user-determined paths through the content. The slide show feature, poised to replace more conventional methods, projects high-quality images, while providing page forward/reverse options for still images or video clips.

DVD-ROM

As alternatives to the CD-ROM's computer-mediated interactivity, DVD-ROMs are intended for use on a desktop or laptop—holding at least seven times the data of a standard 650-Mbytes CD-ROM. With no restrictions on the type of data DVD-ROMs hold, they store offline Web sites, multiple interactive projects, or even DVD-Video material, all interconnected.

Future Integration and Delivery

As multimedia moves from personal computers to smaller handheld devices and wireless communications, authors, designers, and engineers face challenges of infrastructure, design, and usability. As opposed to the material on innovative large plasma screens, Internet content, scaled to fit hand-held screens, will undoubtedly be limited in scope. Likewise, voice-activated commands will present challenges to multimedia authors as they design its functionality into interfaces. New aesthetics and styles will invariably evolve, despite multimedia's preoccupation with designing access to information and communicating and displaying it effectively. Even as such wireless interconnectivity between small mobile devices and global networks has been accomplished, at least in part, by such software as BlueTooth, the visual design of the communication has yet to adapt.

CONCLUSION

The integration of a wireless future with the visual and sonic richness of multimedia is one of the major initiatives of the media, telecommunications, and design industries. Multimedia, however, is not merely limited to the concerns of commerce; it occupies a larger realm of digital culture and must cultivate experimentation and constant re-examining of what it means to communicate through images, text, sound, and motion. From time to time, it relentlessly hangs on to its roots in intermedia, incorporating live actors and installation, and continuing to push boundaries. Multimedia, thus, maintains its own negotiation of aesthetics and technology. As seen in its central role in creating cultural forums like the Internet, multimedia will no doubt maintain an impact on how we work and play, how we communicate, and how we experience the world.

APPENDIX A: AUDIO INPUTS FOR MULTIMEDIA PRODUCTION

Source	Channels	Sample rates	Bit depth	Compression	Comments
Audio CD	Stereo	44.1 kHz	16-bit	None	Standard CD-quality sound
Mini disk recorder	Stereo	44.1 kHz	16-bit	Contains transcoding; slightly better in quality than MP3 files	Suitable for field recording
MP3 audio from Web	Stereo	44.1 kHz	16-bit	MPEG Layer 3 audio compression (MP3)	Standard for audio files exchanged on the Internet; good perceptual compression throughout audio range
DAT (digital audio tape)	Stereo	48 kHz (standard); also 32-, 44.1-, and up to 96-kHz rates)	16-bit (standard); also 24 and up to 32 bit	None	Standard professional quality master recording
Audio from mini DV camera	Stereo	48 kHz (standard); also 32 and 44.1 kHz	16-bit	None	Also suitable for field recording. Be sure capture settings on your DV editor match your sample rates, or sound will not sych to image.

APPENDIX B: HARDWARE FOR SOUND PRODUCTION

- Microphones:
  - Stereo dynamic (for live recording via MiniDisc)
  - Lavalier
  - Large diaphragm vocal studio mic
  - Shotgun mic (optional)
- MiniDisc Recorder (can substitute mini DV camera, which gives you higher quality sound, but can also be more cumbersome)
- DAT (digital audio tape) Recorder (studio deck)
- Appropriate cables and connectors
- MIDI keyboard (or other input device)
- MIDI sound module
- MIDI interface
- Sound card capable of digitizing analog sound (can also use capture audio only from mini DV camera, but not as elegant)
- Mixer—at least 2 mic inputs and 4 stereo line inputs
- Reference monitors—powered monitors do not require an amplifier

APPENDIX C: SOUND DESIGN TERMINOLOGY

Levels of Organization of Sound

Organizational level	Definitions
Sound parameters	Pitch, dynamic, duration, timbre. All sonic events have these parameters.
Gestures	Change of sound parameter(s) over time.
Textures	The combinations of a single or multiple voices (sound sources) engaged in similar or contrasting activity.
Form	The mix of successive and/or simultaneous textures and the change of these textures over time.

Sound Parameters

Sound parameter	Definitions
Pitch	Frequency of a sound, expressed in cycles per second (Hz). Range: 20 to 20,000 Hz, usual limits of human hearing.
Dynamic	Amplitude or loudness of a sound, expressed in decibels (dB). Range: 0 (nearly inaudible) to 140 dB (threshold of pain).
Duration	Length of time a sound lasts, expressed in units of time (milliseconds, seconds, minutes, hours).
Timbre	Unique characteristic tone color or signature of a sound, most often expressed verbally (bright, dark, hollow, etc.).
Location	Spatial position of the source of sound, determined by placement on stage or within a given sonic environment (such as left-right placement in stereo audio or 3D placement in surround-sound audio).

Gestures

Changing any sound parameter results in a gesture. The change can range from *subtle* and *gradual* to *dramatic* and *immediate.*

Sound parameter	Gestures resulting from subtle, gradual change	Gestures resulting from dramatic, immediate change
Pitch	Pitch vibrato, slides (glissandi)	Melodic statements
Dynamic	Crescendo, diminuendo	Dynamic accents
Duration	(The relative change of density of activity over time)	
Timbre	Timbre modulation	Hocket, or note-to-note instrumentation or sound source change
Location	Gradual or variable shift in a sound's apparent source	Marked or dramatic vectors of sound across or through the sonic environment

Textures

Textures are the combinations of a single or multiple *voices* (sound sources) engaged in *similar* or *contrasting* activity. Although there are many possible textures representing these various combinations, many styles and genres of musical (and sonic) activity can be examined according to the three following texture types:

- **Ambient textures** are slowly evolving sound "landscapes" that have discernible pitch content (as in the Western notion of harmony), or they may not (as in a live recording of a rain forest). Their organization over time is characterized by a relatively slow rate of change. Ambient textures can seem so static, because change takes place so gradually that listeners are sometimes unaware of a specific change.
- **Rhythmic textures** are characterized by regular, predictable, and symmetric patterns of sounds that may or may not be of discernible pitch. With any sound repeated at regular intervals creating a rhythmic layer, multiple rhythmic layers then can be combined, with the result sometimes perceived as a single (although dense or complicated) rhythmic texture. Sometimes rhythmic textures that become highly predictable seem static to listeners, receding into the background of their attention.
- **Solo or foreground textures** are any sounds in a musical work that become so prominent that all other musical elements recede to the background (such as instrumental or vocal solo). Solo elements typically have a clearly discernible pitch content (melody for example).

Form

Form is the mix of successive and/or simultaneous textures, along with gestures and individual sound events often articulating change in the mix of textures (i.e., beginning, key sections, end).

Although your approach to form can be very spontaneous or very structured, you need to be able to communicate your formal desires if your work involves more than one person. You can do this best by creating a score, which is a timeline with verbal or graphic descriptions of the kind of activity you want to create.

ACKNOWLEDGMENT

I thank my wife Deborah Jones for her help, patience, guidance, and expertise in crafting this chapter into its finished form. No words can express my gratitude to her; she has a gift for language that has shaped this entire work.

GLOSSARY

See Appendix C for additional audio terms and information.

AC3 A digital audio file format that supports Surround Sound audio.

ADR (automatic dialogue replacement) The process of recording the voice of an actor after the film or video of the action has been shot, and synchronizing the recording with the film or video; also known as postsynch.

ActionScript Macromedia Flash's object-oriented scripting or programming language.

.aif (audio interchange format) A digital audio file format that supports mono or stereo (2-channel) audio.

Analog Any media element (sound, image, text, etc.) in its native state that has not been converted to digital form.

Bandwidth See bit-rate.

Bezier curve A curve generated between any two points in the creation of a computer graphic, while using vector-art applications. (See vector art/vector-based graphics.)

.bmp (bit-map photo) File format that supports bitmap graphics, used primarily on the PC platform.

Bit-depth The measurement of data or information contained in the smallest unit of a digital media type. A pixel, for instance, is the smallest unit of an image; its bit-depth measures the number of bits in a pixel. For sound, the smallest digital unit is a "sample"; its bit-depth measures the number of bits per sample.

Bit-rate Amount of data transmitted over a network or between electronic devices, expressed in kilobits or megabits per second. Also known as data rate, data transfer rate, and throughput.

Bitmap graphic A digital image made of a grid or matrix of individual pixels.

Cardioid A microphone type, also known as unidirectional, with greatest sensitivity to sound in front and to the sides of the microphone. This sensitivity to sound, or pickup pattern, resembles a heart shape, which gives this microphone its name.

Chorus Audio effect of generating multiple versions of a sound, usually at slightly different pitch levels than the sound, and playing them with the original sound, resulting in a "thicker" or more dense sound.

Codec (compression–decompression) The process of compressing and decompressing digital motion data—scaling the data to meet the bandwidth capacity of various networks.

Compression–expansion Contracting (compression) or augmenting (expansion) the total dynamic or amplitude range of an audio recording.

Data through-put See bandwidth.

Delay Audio effect of generating a copy of a sound, played with the original sound file, but at a slightly later time (usually measured in milliseconds) than the original.

Equalization Any process of increasing or decreasing the intensity of an audio signal (or sound) within a particular pitch range.

Foley The process of creating sound effects for pre-existing motion media (film, video, or motion graphics).

Gate Audio effect of adding reverb or other effects to a sound while that sound is at a specific dynamic level or threshold.

.gif (graphic interchange format) A file format that supports a color palette of 216 standard colors used on the Web.

HTML (hypertext markup language) The standard language or code used in Web design, using a system of tags to describe the layout of image and text on a page; e.g., <HEAD><TITLE>My Web Page</TITLE></HEAD>.

Hypertext Text supporting a system of links through which viewers can select their own path through the information.

Java The object-oriented multimedia programming language used in the creation of Web-based projects so they can run on any Java-enabled platform or browser. Java was developed by Sun Microsystems.

JavaScript A programming language for Web interactivity developed by Netscape. Although similar to Java in many features, it is not the same as Java.

.jpeg (joint professionals electronic graphic) The bitmap file format that supports 24-bit color, supporting variable image quality—thus, often used with photos.

Keyframes The frames in which the main actions change in an animation or motion graphic.

Lingo The scripting or programming language of Macromedia Director.

Live motion Action or images captured in real time with a video or motion picture camera.

Live synch sound Sound recorded during the process of filming or videotaping, usually refers to the synchronization of live sound to the corresponding action filmed or recorded.

Motion menus An interface that incorporates video or other motion imagery.

.mov A digital video file format of QuickTime.

MP3 Audio compression process that creates quality sound recordings with small file sizes. Refers to a coding scheme that handles audio compression (Layer 3) in the MPEG-1 video codec.

.mpeg (motion picture expert group) A series (.mpeg1, 2, 4, 7, 21) of file formats developed by Motion Picture Expert Group, primarily for the delivery of digital video or audio across various platforms.

Normalization Digital signal processing effect that locates the loudest point of a sound file, raises that point to a specific audio level, and proportionally scales the remainder of the sound file to that level.

Object oriented Programming procedure that uses higher-level rules (parent or subroutine) to create multiple, interactive lower-level environments (child or object). First developed in the programs Simula 1 and Simula 67.

.pct or .pic (i.e., PICT file) A file format developed by Apple Computers supporting 16- or 32-bit-depth bitmap graphics.

.pdf (portable document format) File format that displays text and layout design in a Web browser, retaining such features as specific typefaces, letter spacing, and other typographical features.

Pixels (picture elements) The smallest unit of a digital bitmap image; these squared-shaped units come together in a grid to form a bitmap image.

Plug-ins An application added to an existing program, providing additional/specialized functionality.

.psd (Photoshop document) The file format for files created in Photoshop.

QuickTime Technology by Apple Computers to support multiple media types in digital imagery, audio, and video, and its delivery across various platforms.

Rasterize Process of converting vector-based art or graphics into bitmaps.

Reverberation Audio effect of adding the ambience or sound reflection of a particular room (large room, small hall, concert hall, auditorium, etc) to a sound.

Subwoofer A loudspeaker that reproduces sounds at the lowest pitch range (bass) of the audio spectrum—generally at a frequency between 20 and 150 cycles per second.

.swf (Shockwave Flash) The standard file format for displaying vector-based animation and interactivity on the Web; these files are created in Macromedia Flash or in other applications.

Surround sound The use of multiple loudspeakers to simulate a 360° sound environment.

Telecine The process of digitizing motion picture film footage for use in digital video editing applications.

Vector-art/vector-based graphics Digital imagery created by applications that use mathematical formulas to describe points, lines, curves, and areas in an image.

Video streaming Digital video incrementally downloaded to a Web page. The video file plays while it is being downloaded.

.wav File format that supports digital audio, used primarily on the PC platform.

CROSS REFERENCES

See *Downloading from the Internet; File Types; Interactive Multimedia on the Web; Virtual Reality on the Internet: Collaborative Virtual Reality.*

REFERENCES

ArtMuseum.net (n.d.a). Pioneers: Jodi, beyond interface. Retrieved April 20, 2003, from http://www.artmuseum.net/w2vr/timeline/Jodi.html

ArtMuseum.net (n.d.b). Pioneers: William Burroughs, cut-ups. Retrieved April 20, 2003, from http://www.artmuseum.net/w2vr/timeline/Burroughs.html

Cantrell, C., et al. (2002). *Macromedia Flash enabled: Flash design and development for devices*. Indianapolis, IN: New Riders.

Crabtree, S. (2002, December 4). Cut to the chase. Retrieved April 20, 2003, from http://www.hollywoodreporter.com/hollywoodreporter/emmys/article_display.jsp?vnu_content_id=1772771

Davis, J. (2003). *Flash to the core: An interactive sketchbook*. Indianapolis, IN: New Riders.

Discreet (n.d.). Retrieved April 20, 2003, from http://www.discreet.com/support/codec/.

Encyclopedia Britannica (2002). Retrieved April 20, 2003, from http://www.britannica.com

Horn, R. (1999). In Jacobson, R. (Ed.), *Information design* (pp. 15–30). Cambridge, MA: MIT Press.

Nakamura, Y., et al. (2000). *New masters of Flash*. Birmingham, UK: Friends of ED.

FURTHER READING

General Multimedia

Cotton, B., & Oliver, R. (1997). *Understanding hypermedia 2000*. London: Phaidon Press.

Packer, R., & Jordan, K. (Eds.). *Multimedia from Wagner to virtual reality.* Retrieved April 20, 2003, from http://www.artmuseum.net/w2vr/

Purgason, T., Blake, B., Scott, P., & Drake, B. (2001). *Flash deConstruction: The process, design, and actionscript of juxt interactive.* Indianapolis, IN: New Riders.

Tan, M., et al. (2002). *Flash math creativity.* Birmingham, UK: Friends of ED.

Veen, J. (2001). *The art and science of Web design.* Indianapolis, IN: New Riders.

New Media and Digital Culture

Druckrey, T., with Ars Electronica (Eds.). (1999). *Ars Electronica—Facing the future*. Cambridge, MA: MIT Press.

Gelernter, D. (1998). *Machine beauty*. New York: Basic Books/Perseus Book Group.

Lunenfeld, P. (Ed.). (1999). *The digital dialectic: New essays on new media*. Cambridge, MA: MIT Press.

Manovich, L. (2001). *The language of new media*. Cambridge, MA: MIT Press.

Riccardo, F. (1998, October). *Field notes on the secret war between value and meaning in digital culture*. Retrieved April 20, 2003, from http://callmoo.uib.no/dac98/papers/ricardo.html

Design and Still Imagery

Adobe Creative Team (1995). *Adobe Print Publishing Guide*. Mountain View, CA: Adobe Systems.

Bergsland, D. (2002). *Introduction to digital publishing*. Clifton Park, NY: Thomson Learning.

Carter, R., Day, B., & Meggs, P. (2002). *Typographic design: Form and communication* (3rd ed.). New York: Wiley.

Dillon, G. (2000). Dada photomontage and net.art sitemaps. Retrieved April 20, 2003, from http://faculty.washington.edu/dillon/rhethtml/dadamaps/dadamaps2b.html

Holtzschue, L., & Noriega, E. (1997). *Design fundamentals for the digital age*. New York: Wiley.

Meggs, P. B. (1998). *A history of graphic design* (3rd ed.). New York: Wiley.

Digital Film and Motion Images

Katz, S. D. (1991). *Film directing shot by shot: Visualizing from concept to screen*. Studio City, CA: Michael Wiese Productions.

Long, B., & Schenk, S. (2000). *The digital filmmaking handbook*. Rockland, MA: Charles River Media.

Newton, D., & Gaspard, J. (2001). *Digital filmmaking 101: An essential guide to producing low-budget movies*. Studio City, CA: Michael Wiese Productions.

Rees, A. L., (1999). *A history of experimental film and video*. London: British Film Institute Publishing.

Rodriguez,. R. (1995). *Rebel without a crew, or how a 23-year-old filmmaker with $7,000 became a Hollywood player*. New York: Penguin.

Wright, S. (2002). *Digital compositing for film and video*. Woburn, MA: Butterworth-Heinemann.

Music and Sound

Chadabe, J. (1997). *Electric sound: The past and promise of electronic music*. Upper Saddle River, NJ: Prentice Hall.

Prendergast, M. (2000). *The ambient century: From Mahler to Trance—The evolution of sound in the electronic age*. New York: Bloomsbury.

Roads, C. (2001). *Microsound*. Cambridge, MA: MIT Press.

Rockwell, J. (1983). *All American music: Composition in the late twentieth century*. New York: Knopf.

Weis, E., & Belton, J. (Eds.) (1985). *Film sound—Theory and practice*. New York: Columbia University Press.

Multiplexing

Dave Whitmore, *Champlain College*

INTRODUCTION

In recent years, there has been phenomenal growth in both the number of Internet users and in the ways in which the Internet is used. Today's applications are much more than just Web page access. They include music-quality wide-band audio, video, voice, and multimedia applications. This growth in applications and in the bandwidth that is required to support them has put enormous pressure on the Internet infrastructure.

There are three types of key players on the Internet today that are concerned with bandwidth availability: businesses that offer services through the Internet, consumers who acquire these services, and service providers that provide the connections between businesses and consumers. Each of these constituencies has a stake in access bandwidth.

Many businesses now depend on the Internet for all or for a significant part of their revenue. If customers can't access a Web site, the company will lose money. Depending on the company, these losses could be significant and business threatened. Internet businesses not only require bandwidth for routine operations but also for backup applications. Failures due to network congestion or outages due to physical damage to circuits are no longer tolerable. Companies are seeking redundant network paths and backup bandwidth strategies that will guarantee a high degree of system reliability. Consumer expectations have changed, too. The old joke about WWW meaning "world wide wait" is no longer funny, and consumers are demanding faster response times. The growth of the cable modem and digital subscriber line (DSL) markets is clear evidence of this fact. Service providers include Internet service providers (ISPs) that furnish the consumer's portal to the Internet. Other service providers are telecommunications companies that furnish connections between consumers and ISPs and between ISPs and the Internet backbone.

Fast and reliable access to adequate bandwidth depends on making the best possible use of the existing infrastructure. Building and maintaining outside plant equipment such as fiber-optic networks and submarine cables are expensive, and it is to everyone's advantage to use this infrastructure as efficiently as possible. Multiplexing provides a way to make the best use of existing equipment to promote high bandwidth, high-reliability connections. Consumers and businesses alike benefit from fast and reliable access. All three groups benefit from reduced costs.

Definition

A simple working definition of multiplexing is the transmission and reception of two or more electronic signals over the same medium. In this definition a signal may be voice, data, video, facsimile, wide-band audio, or a combination of these. The medium may be copper wire, coaxial cable, fiber-optic cable, or a radio frequency. Figure 1 is a schematic diagram of the process. On the left side of the figure is the multiplexer. Its function is to collect the different input signals and organize them for transmission over the single shared medium. On the right side is the demultiplexer. Its function is to break up the composite signal and recover the individual components. In between is a shared medium over which all of the individual signals travel. All multiplexing schemes use this basic concept but differ in the ways in which the signals are combined for transmission over the medium. Most multiplexing systems operate at the physical layer (Layer 1) of the Open Systems Interconnect (OSI) model.

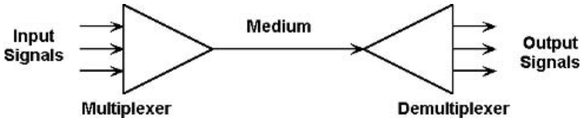


Figure 1: Schematic diagram of the basic multiplexing process.

Rationale

The cost of installing, operating, and maintaining transmission media is high. Multiplexing makes it possible to make the maximum use of expensive media through high utilization of an expensive resource. Here is an example. Think of the cost to provide transatlantic fiber-optic cable service. The initial capital outlay is significant. Thousands of miles of fiber-optic cable must be manufactured, the services of a special ship must be acquired and repeaters must be installed along the route. There is a significant labor cost as well. The subsequent maintenance cost to keep the cable functioning is also high.

The operator of such a cable facility must maximize its use if the rather significant initial and continuing costs are to be recouped and a profit made. Clearly it would not be reasonable for such a cable to carry only one telephone conversation or one stream of data. In this example, multiplexing can be used to put hundreds or thousands of voice, data, and video signals over this single physical medium. Clearly this is an efficient use of the resource.

Multiplexing methods today center on digital techniques. There is no distinction in today's world between voice, data, or video. The ability to handle any arbitrary bit stream makes it possible to support these different types of payloads; and, inside today's global digital networks, all traffic is digital.

Historical Antecedents

The problems associated with multiplexing today are very similar to those encountered by telecommunications pioneers in the 19th century. Then, as today, the problems of providing capacity at low cost are critical.

The cost to install telegraph poles and wires in the American west in the mid-19th century was high. The terrain is difficult and the distances long. As use of the telegraph increased so did the pressure to make the best use of these limited physical circuits. By the 1870s, there were several schemes for multiplexing up to four telegraph signals on the same wire. (It must be remembered that in that era electricity was a new science, and the discipline of "electronics" was several decades away.) Achieving even this limited degree of multiplexing required a considerable degree of inventiveness and ingenuity.

During the latter part of the 19th century, the explosive growth of the telephone system mirrored the growth of the telegraph system. The Bell System was faced with the same problem the telegraph operators had: how to make the best use of limited wire circuits. The problem for the Bell System was more complicated, however, because of the higher bandwidth requirements for voice. In that era it was thought that voice signals required a bandwidth of about 2 KHz. The available telegraphic systems that were capable of multiplexing simple "on" and "off" telegraphic signals were unable to handle the bandwidth required for voice.

It wasn't until about 1914 that a viable multiplexing system was invented for voice circuits. This system used the newly invented vacuum tube. The first of these systems was capable of carrying four simultaneous telephone conversations over one circuit using frequency division multiplexing. The Bell System continued to develop vacuum tube analog systems for voice multiplexing throughout the first half of the 20th century. The analog approach to multiplexing limited the number of voice channels that could be carried, however. In addition, analog technology was prone to failure and required frequent maintenance.

By 1960, Bell Labs had invented a digital multiplexing system called the T-1. The T-1 carrier was first used in 1962 to link telephone central offices, and it was later offered to individual customers. Later the Bell System deployed faster carriers capable of handling many more telephone calls. The next major step in digital multiplexing came when fiber-optic facilities were deployed beginning in 1977. This technology vastly expanded the number of signals that could be carried over a single medium.

It can be seen from this brief history that multiplexing was developed and refined in the context of the public switched telephone network (PSTN) for voice applications. On the surface it appears that voice applications have nothing to do with data applications or Internet access and that voice and data represent two different worlds. This is not the case. The work on multiplexing for the PSTN that was originally done for digital voice applications has been extended to data and Internet access as well. Therefore, I discuss multiplexing in both the voice and data contexts. The similarities will become readily apparent.

FREQUENCY DIVISION MULTIPLEXING

Frequency division multiplexing (FDM) was the first technology used to multiplex voice signals. It was adapted later to handle special data communications problems such as the support of multiple "dumb terminal" connections to a mainframe. FDM was used before the invention of time division multiplexing (TDM), and after their deployment FDM systems were largely abandoned. The basic concepts of FDM have been resurrected in wavelength division multiplexing, however, a technique used to enhance the capacity of fiber-optic systems.

FDM Concepts

In FDM, a separate carrier frequency is allocated to each of the signals to be multiplexed. The spacing of each carrier frequency exceeds the bandwidth of its payload, a practice that provides a guard band that keeps adjacent carrier frequencies from interfering with each other. (This interference is called crosstalk.) The individual carrier frequencies are modulated by the input signals. Modulation may be amplitude modulation (AM), frequency modulation (FM), phase modulation (PM), or some combination of these. All of the input signals appear at the same time on the shared medium. The signals are separated only by the different carrier frequencies. This is an important contrast to TDM.

The following example illustrates the basic concepts of FDM. Commercial FM broadcasting in the United States uses frequency division multiplexing. The Federal Communications Commission (FCC) has defined carrier frequencies from 88.1 MHz to 107.9 MHz for commercial FM broadcasting. Stations are assigned to one of the carrier frequencies in this range, and each carrier frequency is separated from its nearest neighbor by 200 KHz. For example, permitted carrier frequencies are 88.1 MHz, 88.3 MHz, 88.5 MHz, and so on. According to FCC rules, a station operating on one of these frequencies is allowed to use no more than 75 KHz of bandwidth. This bandwidth is centered on the carrier frequency. Because there is 200 KHz separating each carrier frequency, there is 125 KHz left for guard bands.

FDM Technology

FDM is based on analog technology. In the early days of telecommunications, digital technology was not available, so engineers designed multiplexing systems using oscillators and filters with analog devices such as vacuum tubes, inductors, capacitors, and resistors. These components must be tuned to specific carrier frequencies, and the successful operation of a multiplexing system depends on frequency stability. If the frequency drifts on either the transmitting or receiving ends, the performance of the multiplexing system is degraded. For example, think about a commercial FM receiver with a frequency that changes over time. The station will appear to "drift," and the listener must retune the receiver at regular intervals.

In an analog multiplexing system, component values will change because of age, exposure to heat, and other environmental factors. As a result, the frequency characteristics of the system will change. These changes lead to "drift," which degrades the quality of the multiplexed signals.

FDM and Telecommunications

Here is an example of an FDM multiplexing system used in voice communications. This configuration was used by the Bell System for many years and is defined in the G.232 standard now maintained by the International Telecommunication Union (ITU). Designed to multiplex 12 telephone signals onto a single medium, it supports a frequency range of 200 Hz to 3400 Hz for each input channel, a range that is adequate for voice grade signals. The bandwidth is therefore 3200 Hz. Each of the signals occupies a 4-KHz range, and therefore the guard bands are 800 Hz. Table 1 shows the carrier frequencies for each of the 12 signals defined in the ITU G.232 standard.

The shared medium must be capable of handling a range of frequencies from 60 KHz to 108 KHz. Total bandwidth in this modest system is 48 KHz. Wide bandwidth of the shared medium is the price to be paid for using FDM. Figure 2 is a simplified example illustrating multiplexing using three of the voice channels just described. Note that the three signals are transmitted at the same time on the shared medium and are separated from each other only by different carrier frequencies. An actual voice FDM system would use two of these configurations, one for transmitted voice and one for received voice. Figure 3 shows a 12-channel FDM equipment bank used in an analog microwave system.

Table 1 Standard Carrier Frequencies for G.232 Voice Multiplexing

Input Signal Channel	Carrier Frequency Band
1	60 KHz to 64 KHz
2	64 KHz to 68 KHz
3	68 KHz to 72 KHz
4	72 KHz to 76 KHz
5	76 KHz to 80 KHz
6	80 KHz to 84 KHz
7	84 KHz to 88 KHz
8	88 KHz to 92 KHz
9	92 KHz to 96 KHz
10	96 KHz to 100 KHz
11	100 KHz to 104 KHz
12	104 KHz to 108 KHz

FDM Today

The data and telephony applications of FDM have, for the most part, been phased out in favor of time division multiplexing, which, as with other digital techniques, is much more stable and reliable than FDM. There are, however, several important applications of FDM that are used today for Internet applications. These include conventional 56-Kbps dial-up modems, asymmetric digital subscriber lines (ADSL), and wavelength division multiplexing, which is used to enhance the capacity of fiber-optic carriers.

TIME DIVISION MULTIPLEXING

TDM is a digital technique used in wide-area networks such as the PSTN, T-carriers, and optical carriers. It is the dominant multiplexing technology today for voice and data applications. TDM-based technologies are widely used to connect users to Internet service providers and to connect Internet services providers to the Internet backbone. TDM is based on digital technology, and it is more reliable and stable than FDM. It is also capable of handling much higher bit rates.

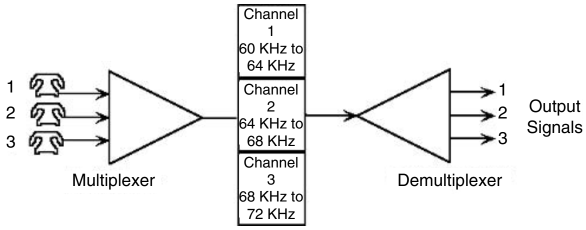

1
2
3
Multiplexer
Channel 1 60 KHz to 64 KHz
Channel 2 64 KHz to 68 KHz
Channel 3 68 KHz to 72 KHz
Demultiplexer
1
2
3
Output Signals

Figure 2: Example of simplifed voice frequency division multiplexing (FDM).

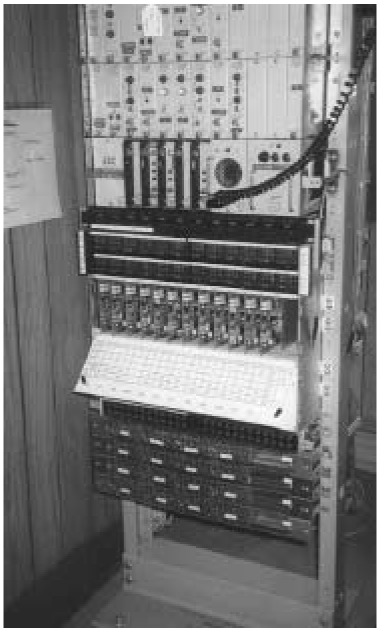


Figure 3: A 12-channel frequency division multiplexing (FDM) equipment bank used in an analog microwave system for telephony. Copyright © 2002 Dave Whitmore. Reprinted by permission.

TDM Concepts

In contrast to FDM, TDM allocates a slice of time known as a time slot on a single aggregate channel to each individual signal. Each signal, in turn, is transmitted on the shared carrier, and only one signal at a time is allowed to use the shared carrier. When all of the signals have been given their turn, the first signal to have been sampled is given another turn. This concept is in contrast to FDM in which each signal occupies the shared carrier at the same time. It may be helpful to think of it this way: With FDM, a user is assigned "some" portion of the aggregate channel "all" of the time. With TDM, a user is assigned "all" of the aggregate channel "some" of the time. Because there is no guarantee that each signal will arrive at precisely the time it needs to enter the shared medium, there is usually some buffering in the multiplexer. Data may have to be held for a short time until its time slot becomes available. A rough analogy can be found in a highway. Cars enter a highway from a ramp. One car at a time goes on to the highway, and once on the cars stay in line. (No passing or lane changing is allowed.) The individual cars are analogous to the individual input signals, and the highway is analogous to the shared medium.

There's one other important point about the highway analogy. In a two-lane highway, cars travel in one lane to go in one direction and in the other lane to go in the opposite direction. In other words, there are two separate lanes, one for each direction of travel. In a TDM system, there are two paths as well—one path for transmitted data and another for received data. A T- or a fiber-optic carrier will have two network connections, one for transmitted data and another for received data.

Figure 4 illustrates the TDM concept for a data application. In this illustration, there are three input signals A, B, and C. Each of the signals is sent over the shared medium one at a time: A is first, followed by B, followed by C, and then the cycle repeats. Note that at any one time, only one of the signals is present on the shared medium. Time separates the signals from each other. This is in contrast to FDM in which all signals are present at the same time on the medium, and separation is achieved through different carrier frequencies.

Synchronization is vital to this process. Both the multiplexer and the demultiplexer must be closely synchronized to ensure that each of the individual multiplexed signals is properly recovered at the far end of the circuit. Without synchronization, this process will quickly fall apart. Note that Figure 4 shows a system for carrying data in one direction only. A mirror image of this configuration would be required to transmit data in the opposite direction.

The T-1 Carrier

The T-1 carrier is a basic building block for voice and data systems. It was developed at Bell Laboratories and was first deployed in the Bell System in 1962. Initially, the T-1 carrier was used to connect telephone central offices together, and it provided two major advantages. First, like FDM, it reduced the number of physical wires required to connect central offices together. Second, and unlike FDM, the T-1 carrier was digital and was easy to install and maintain. For example, there were no analog components to cause filters and oscillators to change their electrical characteristics, a problem that plagued the earlier FDM systems. Eventually, T-1 circuits were made available to individual customers. Initial applications were for point-to-point connections between customer-owned data or voice system end points. Later, the telephone companies introduced T-1 offerings for connecting Private Branch Exchange (PBX) systems to the PSTN. Customers could purchase a T-1 for their own use starting in the early 1970s when the T-1 circuit was first introduced to the market.

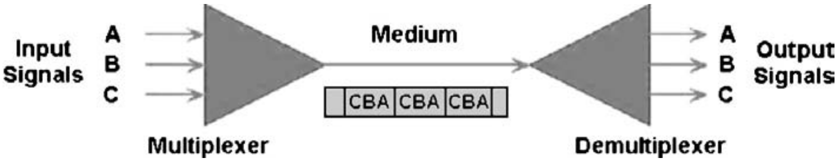


Figure 4: An example of time divison multiplexing.

Figure 5: T-1 terminating equipment. Copyright © 2002 Shepard Communications Group, LLC. Reprinted by permission.

As the Bell System operating companies became more familiar with T-1 technology, the costs to customers eventually went down. At the same time, the original voice applications of the T-1 were extended to include data. This was a relatively easy transition because of the digital nature of the T-1. Today T-1 circuits are routinely used to connect ISPs to the Internet backbone and to connect individual businesses and organizations to ISPs.

The T-1 is the simplest form of TDM multiplexing for voice and data circuits, and understanding a T-1 is essential to understanding more advanced forms of time division multiplexing. T-1 refers to a detailed and precise physical and electrical set of specifications. (See http://www.ansi.org/rooms/room_9/public/1998feb/atis_6.html for a copy of these standards.) A T-1 carrier is a two-pair (four-wire) circuit that supplies 24 voice-grade channels or 1.544-Mbps aggregate capacity. This data rate may seem arbitrary to data users, but the origin of this rate goes back to voice multiplexing requirements. Figures 5 and 6 show T-1 terminating equipment located at a large telecommunications service provider.

Figure 6: T-1 terminating equipment. Copyright © 2002 Shepard Communications Group, LLC. Reprinted by permission.

T-1 Configuration and Data Rate Derivation

A major part of the development of the T-1 centered on the representation of an analog voice signal in digital format. Bell Laboratories determined that an 8-bit sample of an analog voice signal provided enough resolution and dynamic range to represent voice conversations adequately.

A T-1 is designed to carry 24 voice conversations. Each voice conversation is sampled in a round-robin fashion. In other words, channel 1 is sampled, followed by channel 2, and so on until channel 24 is sampled. The next sample is channel 1 and so on. Once each 8-bit sample is made, it is placed on the shared carrier in this round robin fashion. The carrier must support 24 channels times 8 bits per channel for a total of 192 bits for one full cycle of all 24 channels. A framing bit is added to preserve synchronization, and this additional bit creates a frame of 193 bits.

In the digital world, we don't have a continuous analog signal but have instead a signal that is a series of discrete digital samples. How fast do we need to sample to preserve voice intelligibility and recognizability? Harry Nyquist, a Bell Labs scientist, determined that it is necessary to sample at least twice as fast as the highest frequency that you wish to reproduce. If we assume a very generous bandwidth of 4000 Hz for a voice signal we need to sample twice as fast or 8,000 times per second. This is called the Nyquist sampling frequency.

Therefore, for the T-1 carrier, we need to complete our "round-robin" of samples 8,000 times per second. This gives us an aggregate bit rate of 193 bits per frame times 8,000 frames per second for a total of 1.544 Mbps. This is the origin of the oft-quoted T-1 data rate. The bandwidth of a single channel is of interest as well. Each channel has 8 bits and is sampled 8,000 times per second. This means that each individual channel can support 64,000 bps, the bandwidth required to reproduce a digitized voice signal optimally.

Channel Usage and Configuration

A T-1 circuit can be configured as 24 discrete 64-Kbps channels or as a single 1.5-Mbps channel. This latter configuration is useful for point-to-point data exchange or Internet access to an ISP. There is a middle ground between these two configurations, however. It is possible to take the total aggregate bandwidth and divide a portion into voice channels and another portion into one or more data channels. This is the function of hardware called a channel bank. The channel bank provides terminations at both ends of the T-1 circuit for whatever configuration of voice and data is required. Channel banks provide high flexibility and low cost.

North American Digital Hierarchy

The T-1 carrier is the beginning of a family of T-carrier products. Increased experience with the T-1 and improvements in digital technology have allowed the development of even faster T-carrier multiplexing products. T-carriers were realized in both copper wire and in microwave

Table 2 Basic T-Carrier and DS-n Characteristics

SIGNAL	TRANSMISSION FACILITY	RATE (MBPS)	DS-0 CHANNELS	TRANSMISSION MEDIUM
DS-0	—	0.064	1	Copper wire
DS-1	T1	1.544	24	Copper wire
DS-1C	T1C	3.152	48	Copper wire
DS-2	T2	6.312	96	Copper wire, microwave
DS-3	T3	44.736	672	Microwave, fiber
DS-3C	T3C	90.524	1,344	Microwave, fiber
DS-4	T4	274.176	4,032	Microwave, fiber

circuits but the major increase in carrier speed was achieved when fiber-optic cable was introduced. These circuits were called optical carriers (OC).

It soon became apparent that a system was needed for classifying multiplexed data that was independent of the T-carrier designations. The North American Digital Hierarchy is such a classification system. It is based on the single 64-Kbps channel, and this bit rate is called a DS-0 signal. Table 2 contains a summary of the T-carrier hierarchy.

Overhead

Multiplying the number of DS-0 channels by 64 Kbps will always yield a number less than the aggregate bandwidth. For example, 24 × 64,000 bps = 1,536,000 bps, whereas the T-1 is rated at 1,544,000 bps. This gap of 8,000 bps is attributable to overhead for synchronization and control. The disparity between the bandwidth allocated to the user and the total aggregate bandwidth is present in other multiplexed carriers as well. This can be verified by choosing an aggregate data rate from Table 2, multiplying the number of DS-0 channels by 64 Kbps and subtracting the two rates. The difference is attributable to overhead.

International Standards

European carrier standards are different from North American standards. For example, the E-1 standard roughly corresponds to the T-1 but has a different aggregate data rate and DS-0 channel capacity. The E-1 has a rate of 2.048 Mbps and supports 32 DS-0 channels. The good news is that the DS-0 standard of 64 Kbps is shared between North American and Europe, and most telecommunications switches have the ability to translate between these formats. Table 3 is a comparison of the digital transmission technologies of North America, Japan and Europe. It is important to note that E-carrier is the global standard, and T-carrier is a variation on it. Because the United States (and a few other countries) chose to use T-carrier, it is incumbent on U.S. carriers to perform all T-1 to E-1 conversions.

SYNCHRONOUS OPTICAL NETWORK

Standards exist for multiplexed data rates up to DS-3. All manufacturers of equipment accept these standards, which ensure interoperability of equipment. There are some problems with these standards, however. First, the North American Digital Hierarchy is, well, North American. International standards differ from the North American rates. The incompatibility can be managed, but it is cumbersome for international operations. A second problem is that the data rates that go beyond DS-3 involve proprietary transmission protocols and naturally this creates more compatibility problems.

The synchronous optical network (SONET) was first deployed in the 1980s as a response to the need for multiplexing standardization and for increased speed requirements. SONET is the dominant protocol found in optical networks today and is based on international standards. Connections between systems in different countries using equipment from various manufacturers are much easier to thanks to these standards. SONET provides high multiplexed data rates that support bandwidth-hungry applications. The Internet uses SONET technology for the backbone, connections between ISPs and between ISPs

Table 3 T-Carrier and DS-n Characteristics for North America, Japan and Europe

	NORTH AMERICA		JAPAN		EUROPE	
Level	Circuits	Speed	Circuits	Speed	Circuits	Speed
DS-0	1	64 Kbps	1	64 Kbps	1	64 Kbps
DS-1	24	1.544 Mbps	24	1.544 Mbps	30	2.048 Mbps
DS-2	96	6.312 Mbps	96	6.312 Mbps	120	8.448 Mbps
DS-3	672	44.7 Mbps	480	32.06 Mbps	480	34.368 Mbps
DS-4	4,032	274.17 Mbps	5,760	400.4 Mbps	1,920	139.3 Mbps

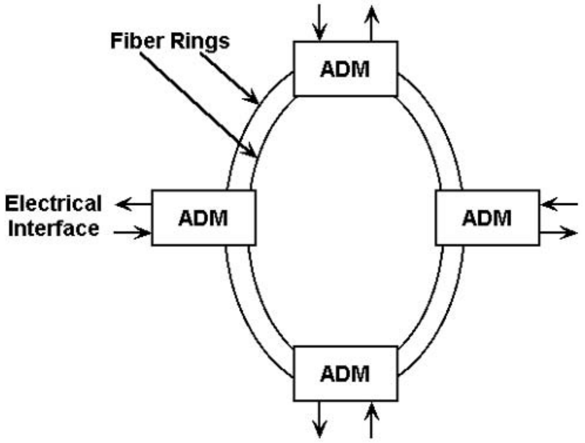


Figure 7: Example of a synchronous optical network (SONET) ring.

and their larger customers. TDM is the fundamental form of multiplexing used in SONET, but FDM-based dense wavelength division multiplexing can be used to combine multiple SONET channels on a single fiber for transport.

SONET Architecture

SONET systems usually consist of one or more rings of fiber-optic cable. Interspersed around the ring are add–drop multiplexers (ADMs), which are optical-electrical-optical conversion devices. ADMs serves three basic functions. First, they handle the transition between wired formats and the optical formats, so that payload components can be added and dropped as required. Most customer termination equipment today have electrical interfaces, and ADMs handle this conversion. Second, ADMs handle optical signal conditioning. Like all signals, the light pulses in a SONET ring will be attenuated or distorted over long distances, and ADMs function as a digital repeater to restore signal quality. Third, ADMs handle network failures. If one leg of the SONET ring fails, the network will heal itself automatically within about 50 milliseconds. This feature promotes high reliability in a SONET multiplexer system. Figure 7 shows a sample SONET ring.

Synchronous Digital Hierarchy

The synchronous digital hierarchy (SDH) provides a set of international standards for optical multiplexing. Unlike the North American Digital Hierarchy, SDH is an internationally accepted standard. Table 4 shows the characteristics of each type of optical circuit in the SDH. It's tempting to think of the DS-0 column as representing "telephone calls," but remember that optical carriers are digital and indifferent to the origin of the bits that they carry. Telecommunication service companies provide digital bandwidth that can be used for any application. The rather curious and seemingly arbitrary multiples of 64 Kbps are a reminder of the telephonic origins of the high-speed digital carriers of today. SONET and SDH permit service providers to allocate bandwidth to customers without the need to do elaborate hardware reconfigurations. Bandwidth flexibility is a major positive feature of SONET and SDH.

Cascading Multiplexers

TDM multiplexers are used to aggregate signals that are themselves aggregated. In other words, there can be a succession of multiplexers to collect individual data streams and consolidate them into larger and larger single data pipes. Figure 8 shows an example of cascading multiplexers, starting with a T-1 signal and ending with an OC-48 signal.

WAVELENGTH DIVISION MULTIPLEXING

Wavelength Division Multiplexing Concepts

A SONET system uses TDM to multiplex data by pulsing monochromatic light through glass fibers. It is possible to pass several different colors of light through the same fiber at the same time, however. The individual colors are independent of each other because they have different wavelengths or frequencies. This is a situation analogous to sunlight, which contains many frequencies (wavelengths) of light. A prism, for example, reveals that white light actually has various color components.

It is possible to send several wavelengths of light simultaneously into a fiber-optic cable using different wavelength lasers, and each of the wavelengths can be

Table 4 North American and International Optical Carrier Standards and Data Rates

Sonet Optical Level (OC)	Synchronous Digital Hierarchy Level (SDH)	Electrical Level (North America Only)	Gross Data Rate (MBPS)	Payload Data Rate (MBPS)	Overhead Rate (MBPS)	Equivalent DS-0 Channels
OC-1	—	STS-1	51.840	50.112	1.728	672
OC-3	STM-1	STS-3	155.520	150.336	5.184	2,016
OC-12	STM-4	STS-12	622.080	601.344	20.736	8,064
OC-48	STM-16	STS-48	2,488.320	2,405.376	82.944	32,256
OC-96	STM-32	STS-96	4,976.640	4,810.752	165.888	64,512
OC-192	STM-64	STS-192	9,953.280	9,621.504	331.776	129,024
OC-768	STM-256	STS-768	39,813.120	38,486.016	1,327.104	516,096

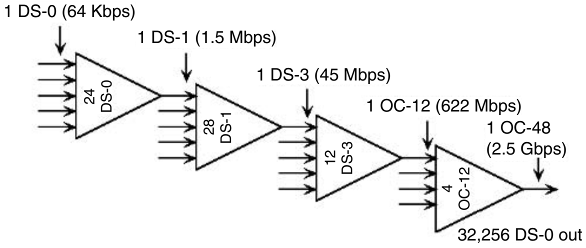


Figure 8: Cascaded multiplexers.

pulsed independently. A separate train of TDM pulses can be sent using each wavelength, and in this way the capacity of a single optical fiber can be dramatically increased.

In wavelength division multiplexing (WDM) there is both TDM and FDM occurring at the same time. First, the individual input data streams use TDM to encode multiple channels of data. Second, each stream of data is assigned a separate frequency (or wavelength) of light, and this is an example of FDM. Therefore, there are two separate types of multiplexing occurring at the same time. Figure 9 shows a simple conceptual example of WDM in which there are three input bit streams: ABC, DEF, and GHI. Each uses TDM to contain several channels of information. For example, in the first bit stream, TDM separates the A input channel from the B input channel and the C input channel. The same process takes place for the other two bit streams.

The multiplexer on the left takes the three TDM multiplexed bit streams and transforms them into a train of pulses on the shared fiber-optic cable medium. Each of the three input streams is given a unique and distinguishable light color before being combined with the other colors on the single fiber-optic cable. (The reference to color here may be somewhat misleading. The driving lasers in a real WDM system operate in the infrared range, not the visible light range.)

The three channels of data coexist on the fiber-optic cable at the same time. The different frequencies of light separate each of the channels from each other and make them independent. Within one color TDM is used to encode several channels of data. This can be thought of as a multilane highway in which several streams of vehicles can travel in parallel at the same time.

Light and FDM

The fundamental idea behind FDM is to separate different channels of data by using different carrier frequencies for each channel. The FM radio example that was cited earlier in this chapter used carrier frequencies ranging from 88.1 MHz to 107.9 MHz. Each carrier frequency is separated by 200 KHz. Different wavelengths (colors) of light also have different frequencies. For example, the ITU has defined the “C” category of WDM optical channels as wavelengths of light from 1,528 nm to 1,560 nm. These wavelengths correspond to 196.10 THz to 192.10 THz, respectively. (Note that wavelength and frequency have an inverse relationship. As wavelength gets shorter frequency gets higher.) The ITU standard defines 80 intervals in this range. Each interval is about 0.40 nm wide, resulting in a bandwidth of 50 GHz for each channel. This latter specification is important when considering the carrying capacity of each of these light carriers. Each interval must accommodate both the payload bandwidth as well as the guard bands.

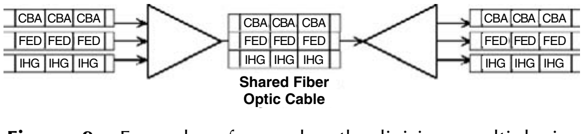


Figure 9: Example of wavelength division multiplexing (WDM).

Most WDM systems use two fiber channels, one for transmitted data and one for received data, but it is possible to use a single optical fiber for both transmission and reception. Two wavelengths of light can be used in the same medium, one encoded with transmitted data and the other with received data. This is similar to the FDM technique used on the telephone local loop to achieve full duplex data transmission with conventional modems.

WDM Technology in Practice

There are two keys to make WDM work in practice. The first key is to control precisely the wavelengths of light on both sides of the multiplexing process. On the transmit side, the lasers used to encode the TDM pulse stream must produce light of consistent and accurate wavelength. The lasers must be stable and must not drift in frequency over time. The detectors on the receiving side must be able to discriminate between the different wavelengths of light and must be sensitive to only a single wavelength. Naturally, both the transmitting laser and the corresponding receiving detector must be precisely matched in frequency as well.

The second key is to control the spacing of the light channels. Channels with very close spacing have the potential to carry large amounts of data because many channels can be accommodated in a single optical fiber system. There is a cost to this approach, however. The hardware must be able to discriminate reliably between individual light frequencies and extract the TDM data from only one of many channels. Closely spaced light frequencies make this task harder. This process is considerably easier for frequencies that are further apart but at a cost of fewer multiplexed channels of data and a lower aggregate bit rate.

Coarse Wavelength Division Multiplexing (CWDM)

Coarse wavelength division multiplexing (CWDM) systems require relatively less stringent hardware specifications than do dense wavelength division multiplexing (DWDM) systems. A typical CWDM system uses channels spaced at 20 nm, and an 8-channel system typically uses wavelengths of 1,470 nm, 1,490 nm, 1,510 nm, 1,530 nm, 1,550 nm, 1,570 nm, 1,590 nm, and 1,610 nm. In terms of frequency, a 20-nm spacing corresponds to 6.4 THz, a wide margin for optical hardware. If each of these eight channels carries an OC-48 (2.5-Gbps) signal, there is an aggregate capacity of 8 × 2.5 Gbps for a total of 20 Gbps.

Dense Wavelength Division Multiplexing (DWDM)

Dense wavelength division multiplexing (DWDM) achieves higher aggregate bandwidth by reducing the spacing of adjacent channels. DWDM is the prevailing optical multiplexing technique in use today. Manufacturers such as Lucent continue to refine laser and fiber technology, and any statements of capacity made here are likely to be surpassed quickly. Today, common DWDM spacings are 1.6 nm (200 GHz), 0.8 nm (100 GHz), and 0.40 nm (50 GHz). (Work is underway to achieve spacings of 25 GHz and lower). Most current DWDM systems are capable of carrying 160 distinct OC-48 signals for an aggregate bandwidth of 400 Gbps.

Using today's technology, TDM limits the upper speed of a single SONET channel to about 10 Gbps (OC-192), but most WDM systems today use 2.5 Gbps (OC-48) as a more conservative data rate. Recent laboratory work has shown that DWDM systems consisting of 320 40-Gbps channels are feasible for field deployment.

STATISTICAL TIME DIVISION MULTIPLEXING

By themselves, FDM and TDM do not make efficient use of the aggregate bandwidth. Each input signal is guaranteed the same bandwidth regardless of the actual need for bandwidth. In some cases, the allotted bandwidth for an input channel may not be adequate to support the input data stream. In other cases some channels may be underutilized, and the aggregate data stream has unused capacity. Statistical time division multiplexing (STDM) attempts to correct this imbalance by dynamically allocating bandwidth depending on past data speed requirements. STDM eliminates the guarantee of bandwidth for one particular channel in exchange for allowing a channel to exceed the defined bandwidth some of the time.

The key for successful operation rests on the definition of "some of the time." If all channels need to exceed their defined base bandwidth all of the time, then STDM won't work. In this case, the aggregate bandwidth will be oversubscribed. When this occurs, the multiplexer will attempt to implement flow control procedures with the subchannels to reduce demand temporarily. If some input channels don't require their full bandwidth, however, then some of this excess capacity can be diverted to those channels that do need bandwidth. A successful STDM system has two main measures of success: (a) It will have a high percentage use of the aggregate bandwidth, and (b) individual channels will experience minimum congestion and wait time.

STDM: Dumb Terminals

One of the earliest uses of STDM was for support of "dumb terminals." Dumb terminals provide only a limited range of functionality, such as a screen and a keyboard, because most of the computing takes place at a remote computer, such as a minicomputer or a mainframe. Dumb terminals are not personal computers, although a PC can provide dumb terminal services through a terminal emulation program. Dumb terminals tend to be bursty devices. In other words, data may come in short, intense bursts followed by longer intervals of relative inactivity. For example, a few commands typed at the keyboard may elicit a screen of data followed by a period of idleness until the next string of commands is typed.

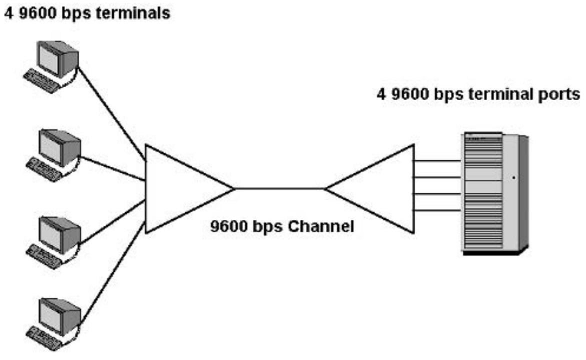


Figure 10: Statistical time division multiplexing (STDM) example.

Figure 10 shows a simple configuration in which four dumb terminals are linked by a single communications circuit to a remote minicomputer. In this example, there are four remote dumb terminals on the left, each of which operates at 9,600 bps. On the right side of the figure, we have four serial ports into a minicomputer, each of which operates at 9,600 bps. A single 9,600-bps channel connects the multiplexers. In theory, this is a terrible configuration because it seems like the common channel is overcommitted. This would be true if each of the devices were operating at full speed 100% of the time. Dumb terminals are bursty, however, and it is unlikely that each of them would be operating full speed at all times. Instead, it is assumed that there is a probability that at any one time several of the terminals will be idle, and the full bandwidth of the channel would be available to whichever single terminal needs it.

To simplify the calculations: If each of the four terminals has a one-in-four chance of being active at any one time, then that terminal will have access to the full shared channel bandwidth of 9,600 bps. Because there is a one in four chance that any terminal will be active, this means that the shared channel will be utilized 100% of the time. This is an efficient use of what is presumably an expensive common resource.

CONCLUSION

This chapter has covered the concepts, theory, and practice of multiplexing. The history of multiplexing provides a context for understanding current technologies and future possibilities. Frequency division multiplexing and time division multiplexing have been covered in depth, as have practical applications such as T-carriers, optical carriers, SONET, and wavelength division multiplexing. Anyone concerned with Internet bandwidth allocation will encounter these concepts and technologies, and this chapter provides a solid foundation for further work. Readers should consult the references provided (see Further Reading) for greater detail on any of the concepts encountered here.

GLOSSARY

Add–drop multiplexers Hardware devices on a synchronous optical network or synchronous digital hierarchy ring that provide signal regeneration, network failure management, and optical to electrical conversion, so that payload components can be added or dropped as required.

Bandwidth The capacity of a channel. In analog systems, bandwidth is measured in Hertz. In digital systems, bandwidth is measured in bits per second (bps).

Carrier frequency In frequency division multiplexing, carrier frequency is the signal on which the information to be carried is imposed.

Channel bank A hardware multiplexer used to combine several signals on a shared channel.

Crosstalk Interference between adjacent multiplexed signals.

Dumb terminal A user device with a keyboard and monitor but with very limited processing capacity.

Frequency division multiplexing (FDM) An analog technique for encoding many narrow band signals on a single wide band channel.

Guard band A portion of a carrier that contains no information but provides a buffer between adjacent channels.

Multiplexing The transmission and reception of two or more electronic signals over the same medium.

Nyquist sampling rate The minimum sampling frequency required to sample an analog signal digitally to preserve a specified upper frequency limit.

Public switched telephone network (PSTN) The worldwide voice network.

Synchronous digital hierarchy (SDH) The European equivalent of synchronous optical network.

Synchronous optical network (SONET) A fiber-optic based data transmission system.

Time division multiplexing (TDM) A digital technique of encoding many signals on a single shared channel by allocating a separate time slice to each signal.

Wavelength division multiplexing (WDM) A technique for increasing the capacity of a fiber-optic medium by encoding different input signals using light of different wavelengths (frequencies).

CROSS REFERENCES

See *Convergence of Data, Sound, and Video; Integrated Services Digital Network (ISDN): Narrowband andBroadband Services and Applications; Public Networks; Wide Area and Metropolitan Area Networks.*

FURTHER READING

The following references cover the early development of multiplexing techniques:

Ault, P. (1974). *Wires west.* New York: Dodd, Mead.

Fagen, M. D. (Ed.). (1975). *A history of engineering and science in the bell system, the early years (1875–1925).* Holmdel, NJ: Bell Telephone Laboratories.

General references to multiplexing can be found in a variety of standard telecommunications reference books and texts:

Bell Telephone Laboratories. (1977). *Engineering and operations in the bell system.* Holmdel, NJ.

Dodd, A. (2002). *The essential guide to telecommunications* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Freeman, R. (1996). *Telecommunication system engineering* (3rd ed.). New York: Wiley.

Hieroki, W. (2000). *Telecommunications* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Noll, A. M. (1991). *Introduction to telephones & telephone systems* (2nd ed.). Boston: Artech House.

Rosengrant, M. A. (2002). *Introduction to telecommunications.* Upper Saddle River, NJ: Prentice-Hall.

Stallings, W. (2001). *Business data communications.* Upper Saddle River, NJ: Prentice-Hall.

The next group of references cover fiber-optic technology, SONET, and WDM techniques:

Goff, D. R. (2002). *Fiber optic reference guide: A practical guide to communications technology* (3rd ed.). New York: Elsevier Press.

Hecht, J. (1999). *City of light.* New York: Oxford University Press.

Hecht, J. (2001). *Understanding fiber optics* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

Shepard, S. (2001). *Optical networking crash course.* New York: McGraw-Hill

Multiplexing standards can be found on the following Web sites:

http://www.itu.org (Date of access September 18, 2002)

http://www.ieee.org (Date of access September 18, 2002)

http://www.ansi.org (Date of access September 18, 2002)

http://www.eia.org (Date of access September 18, 2002)

http://www.telcordia.com (Date of access September 18, 2002)

General information on multiplexing can be found at the following Web sites:

http://williams.comp.ncat.edu/Networks/multiplexing.htm (Date of access September 18, 2002).

http://www.wmpenn.edu/PennWeb/Academic/ArtsTech/CompSci/Networks/Multiplex/Multiplex.html (Date of access September 18, 2002)

General reference for telecommunications:

http://campus.champlain.edu/faculty/whitmore (Date of access September 18, 2002)

Fiber-optic resources can be found at:

http://www.sonet.com/edu/edu.htm (Date of access September 18, 2002)

http://www.nortelnetworks.com/products/01/sonet/collateral/sonet_101.pdf (Date of access September 18, 2002)

http://www.lucent.com (Date of access September 18, 2002)

http://wdm.pennnet.com/home.cfm (Date of access September 18, 2002)

T-1 Reference Material:

http://www.dcbnet.com/notes/9611t1.html (Date of access September 18, 2002)

http://www.arcelect.com/RJ48C_and_RJ48S_8_position_jack_.htm (Date of access September 18, 2002)

http://www.ansi.org/rooms/room_9/public/1998feb/atis_6.html (Date of access September 18, 2002)

N

Nonprofit Organizations

Dale Nesbary, *Oakland University*

INTRODUCTION

Facing competition on a number of levels, nonprofits are turning to the Internet and e-commerce to solicit funds and seek volunteers. Accordingly, many nonprofit and nongovernmental organizations are working to determine the best methods of integrating Internet and e-commerce into their organizations. This chapter examines volunteerism and fundraising as they are affected by the Internet and e-commerce. Along with a literature review, case studies of nonprofit Internet and e-commerce use are presented. Recommendations are made to nonprofit managers with respect to implementing Internet and e-commerce technologies. Finally, key nonprofit Web directories and search engines are listed to give the reader additional examples of nonprofit use of the Internet for e-commerce purposes.

It must be noted that not all nonprofits are created equal. Some are large and complex. The National Multiple Sclerosis Society fits this example well. Others are local in scope with a specific focus. H.A.V.E.N., a domestic violence shelter serving the Pontiac Michigan area fits this example. This chapter paints nonprofits with a broad brush and does not attempt to identify them by size, funding, complexity, or purpose. It does intend to discuss, in some detail, their use of the Internet and its components, particularly the World Wide Web. Although it is a truism that these criteria obviously may be related to the extent to which a nonprofit may use the Internet, this chapter is a first step in determining how nonprofits use the Internet but not necessarily what factors influence their use of the Internet.

It must be also be noted that there exists a wide range of nonprofits, some of which rely extensively on gifts, other relying primarily on volunteerism, and still others relying on fee-for-service or grants to fund operations. This chapter focuses on the latter two categories of nonprofits.

LITERATURE REVIEW

There exists a growing body of academic and popular literature regarding how the Internet and e-commerce have impacted the world (“Internet Infatuation Filters,” 2000; Montagu-Pollock, 2001; Notovny, 2002). Much of it deals with the social sciences (e.g., *Social Science Computer Review*), education (e.g., *British Journal of Educational Technology*), and administrative issues (e.g., *Public Administration Review*). This literature supports the contention that technology is helping transform numerous institutions from being relatively slow and nonadaptive to more nimble and flexible (Western, 2000). Additionally, many nonprofit and nongovernmental organizations are working to determine the best ways to integrate the Internet and e-commerce into their organizations, some more successfully than others.

Moreover, the Internet has radically changed the way that business is conducted. Business, nonprofits, and government now have almost immediate access to clients and customers. This access has accelerated the economy, decentralized society, increased demands for accountability, and allowed nonprofits and funding sources to reach each other much more effectively (Strom, 2002). Nonprofits routinely use Internet tools and techniques to reach their audience. For example, online public service announcements are a natural extension of the Web's ability to reach a larger market than before (Reidman, 1998).

Driven by the commercialization of the Web, use of the Internet has grown dramatically. Estimates range from 500 million to 1 billion users worldwide and substantial increases in government revenue sourced from Internet commerce (Nesbary, 2000). Independent Sector (2002) argues that "The true impact the Internet will have on civil society is in 'building community,' bringing together groups of citizens who are united by shared values working for the public good, often spanning international lines." Communities will enhance the abilities of nonprofits to provide services in a way that has not been achieved to date. Alternatively, Pliskin (1998) and others maintain that the Internet may cause disruption from traditional social and community connections such as work and family.

Although the Internet is experiencing tremendous growth and providing new channels for purchasing goods and providing services, it is also having an effect—not always a positive one—on the way these goods and services are supplied. This is evidenced by numerous dot-com failures in 2000 and 2001 (http://www.IDG.net) and a perceived lack of ethical and legal boundaries, causing some to describe this openness as "ominous" (Donaldson, 2001). This could be an indicator that many consumers remain wary of using the Internet for anything but research, communicating, and information gathering.

At the same time, nonprofits are increasing their audience and identifying a specific, more technologically advanced user (Pearson, 2002). This may lead to a gap or "digital divide" between Internet users and nonusers. Given this gap, it is not enough for nonprofits to simply advertise their services on the Internet, nor is it enough to assume that users will find a particular agency's Web site simply by browsing the Web (Blau, 2002).

To realize fully the Internet's potential, nonprofit organizations must assist "donors" and volunteers in locating their Web site easily, enable them to access desired information quickly, and keep them interested in using the site on a regular basis. This is best done by targeting potential users through advertising and use of meta-tags (key words) embedded in Web pages. "Push technologies" are also a useful technique in getting information in the hands of intended recipients (Allen, 2002). Push technologies actively seek out specific users through e-mail lists, pop-up advertising, and streaming audio and video. Internet surveys are a rapidly growing method of collecting information and donations (Mertler, 2002; Witte, Amoroso, & Howard, 2000). Combining Internet technologies with traditional recruitment and fundraising strategies is the best strategy overall to accomplish the goal (Gardner, 1998).

NONPROFITS AND E-COMMERCE FUNDRAISING

Nonprofits are clearly taking advantage of Internet fundraising. Commercial fundraising activities by nonprofits, those involving businesslike promotion and sales, accounted for approximately $60 billion in 2001 (Strom, 2002). This comprised more than half of all nonprofit fundraising, far outstripping fundraising in government grants (36%) and private donations (10%). Internet fundraising has served as a natural extension to traditional nonprofit fundraising. Many nonprofits are landlocked in a way, meaning that they have no effective mechanism for reaching contributors outside their local sphere of influence. The Internet has changed that dramatically. Following are some examples of how individual nonprofits have used the Internet to enhance fundraising activities.

The Nonprofit Technology Enterprise Network (N-TEN) is an excellent national resource for nonprofits seeking information on the use of technology. N-TEN's Capacity Map is a searchable, online directory of individuals and organizations that provide technology services to the nonprofit sector. The map's purpose is to help nonprofits find the right provider to meet their needs and to help nonprofit technology professional find their peers and colleagues in the field. (N-Ten, n.d.).

As of December 2001, $110 million of the $1 billion donations to various September 11, 2001, funds came via the Internet (Souccar, 2001). The percentage of Internet-raised funds has remained stable since September 11 (through May of 2002). A Chicago-based, interactive agency, Two Way Communications, has adopted several local nonprofit groups and assisted them in creating Web sites. For Children's Memorial Hospital, a pledge site was created (http://www.forchildrens.org) and banner ads designed to raise money in conjunction with the hospital's annual telethon (Reidman, 1998). The Web site of Food for the Hungry (http://www.fh.org/learn/aboutus.shtml) indicates that its Web site is its second-largest source of new child sponsors, with sponsors found via the Internet nearly doubling to 7% over 12 months and generating annual income near $190,000 (Food for the Hungry, n.d.).

Artlinx (http://desktoppublishing.com/artlinx.html) includes approximately 400 links to arts or arts-supporting organizations. The Nonprofit Resource Center (http://not-for-profit.org) includes a comprehensive list of nonprofit Web sites, support organizations, fundraising and grant writing assistance, and marketing support. Fundsnet (http://www.fundsnetservices.com) is another comprehensive nonprofit Web site supporting nonprofit fundraising and development capacity. It has a detailed listing of foundations and funders Web sites, compiled since 1996.

Geneforum.org is a nonprofit organization whose aim is "to promote dialogue and educate people about genome science and its impact on their lives. Using proven strategies, experienced science educators create environments for learning and facilitate forums for discussion" (Geneforum, 2002). They are interested in protecting the privacy of an individual's genome and have identified the Web as the primary mechanism to promote their message. Gene Fowler, executive director of Geneforum, believes that the Internet is the most appropriate cost- and information-effective medium for his organization. He says that Geneform uses the Web to foster what is known as "deliberative democracy."

Research has begun to assess systematically the utility of Internet technology in nonprofits. In 2001, the Greater Detroit Chapter of Fundraising Professionals conducted a technology, media, and fundraising survey of 500 southeast Michigan nonprofit organizations (Fike, 2002). The survey examined a number of issues, including use of the Internet by nonprofits. Of interest, 38% of respondents used Web sites to solicit donations, whereas 22% accepted donations online. Half of the respondents felt that online fundraising was not successful for them, whereas 10% felt that it was moderately successful. Eighty-five percent of nonprofits maintained a Web site, and 45% had had a Web site for 1–3 years. Nearly all Web sites describe the organization's mission, and about half offer visitors an opportunity to register.

In summary, many nonprofits are using the Internet as an effective tool for raising funds and accessing potential donors. Most maintain Web sites and use them to raise funds and advertise. What is the impact of the Internet on volunteerism? The next section will examine this issue.

THE INTERNET AND NONPROFIT VOLUNTEERISM

Like fundraising strategies, volunteerism in nonprofit organizations is being impacted by Internet technologies. Nonprofits are beginning to understand that they can use their Web sites to market effectively to potential volunteers (Oehler, 2000). For example, Kidsonline, a nonprofit education organization, is using broadband Internet to train its volunteers and teach young at-risk students (Cruver, 2001). Making Opportunities for Upgrading Schools and Education (MOUSE) is a New York City–based nonprofit that connects the public schools with high-tech companies (Vargas, 2001). MOUSE provides hardware, software, expertise, and mentorship to participating schools. NetAid, a collaborative effort between the United Nations Development Programme and Cisco Systems Corporation, brings together musicians to help eradicate poverty and hunger in developing nations (Waldron, 1999). Among the pages on the NetAid Web site are volunteer recruitment, fundraising, and information dissemination.

Many volunteers are relatively young and technology savvy. Nonprofits are taking advantage of this by using Internet technology in the organization as well as external Internet tools. Not surprisingly, volunteer recruitment Web pages are becoming more common on Web sites. Nonprofits are advertising for staff positions on job search Web sites. A search of Monster.com (http://jobsearch.monster.com) using the term "nonprofit" yielded 162 hits for jobs open within the past 30 days.

Other nonprofit organizations use the Web as an integral part of their recruitment, communication, and marketing strategy. Blue Lake Fine Arts Camp (http://www.bluelake.org) is a nonprofit arts organization located in Twin Lake, Michigan. Blue Lake serves approximately 4,500 students in its summer camp, more than 1,000 American and international students in its international exchange program, an additional 1,500 students in its adult and youth camp. It is the home of a 100,000-watt FM fine arts radio station serving western Michigan. Blue Lake maintained a basic Web site through 2000 and recently added a number of more advanced features. Students can now print registration materials and job applications directly from the Web, volunteers can print application materials, and donors can pledge to the radio station's fund drive.

CASE STUDIES

The following section examines the Web and e-commerce portfolios of three nonprofits, including the Big Brothers Big Sisters of Metropolitan Detroit, the Ad Council, and the National Multiple Sclerosis Society. These organizations vary with respect to size, mission, and structure, but all have a Web presence. Telephone or personal interviews were conducted with staff from each organization with follow-up discussions concerning issues that arose during the original discussion.

Big Brothers Big Sisters of Metropolitan Detroit

Big Brothers Big Sisters (http://www.bbbs-detroit.com) is a national nonprofit organization whose mission is to strengthen children in need through one-to-one relationships with volunteer mentors (BBBS Detroit, n.d.). This is accomplished by identifying potential mentors and matching them with available little brothers and sisters. A detailed screening process helps to ensure that matches are successful. The BBBS operates numerous chapters around the country, one of which is located in metropolitan Detroit. The Detroit chapter is one of the largest in the country and maintains four locations in and around Detroit. David Lingholm, development director with the BBBS-Detroit, served as primary contact.

The BBBS Web site was first established in 1998. The site was developed internally and remains internally supported. The Web site's purpose was and is for marketing, advertising, volunteer recruitment, fundraising, and information dissemination. It currently does not have interactive e-commerce capacity but does have limited communications capacity via e-mail. The Web site includes several pages for volunteers, program descriptions, donations, news calendar, success stories, and contact pages. The Web site does not provide intranet capacity.

Lingholm identifies a number of potential enhancements to the site. He says that would like to see the Web site allow volunteers to register via the Web, as well as to provide a mechanism for mentor background checks. Also, the current donations page does not allow for interactive (direct) donations via the Web site. The donations Web page must be printed, completed, and then mailed to the BBBS-Detroit office. These are all time-consuming tasks that could be done in significantly less time via a forms-based Web page. "If I could be freed up another hour a week, that would be wonderful," says Lingholm. He says, "We have plans to do it, but the funding is the difficult part. Donors will ask, 'Is this a good use of my donation?' It is hard to make a connection between the benefits of technology and more time to spend on programs." Overall Lingholm believes that the BBBS-Detroit Web site is useful and helpful but needs updating and the capacity to accept donations and volunteers directly. In summary, the Web site was originally designed as a communications device with the community and as a public relations tool for potential mentors and patrons.

The BBBS National Office produced a public service announcement (PSA) in conjunction with the Ad Council. This PSA was released in late 2002 and includes a link to the BBBS national Web site. This use of technology is exactly the kind of imitative needed to ensure that nonprofits are able to compete effectively for funding and resources in a competitive 21st-century environment.

The Ad Council

The Ad Council (http://www.adcouncil.org) is a private, nonprofit corporation whose mission is to create PSAs addressing issues of importance, generally addressing quality-of-life issues. Each year the Ad Council conducts approximately 40 campaigns in issue areas such as improving the quality of life for children, preventive health, education, community well-being, environmental preservation, and strengthening families (Ad Council, n.d.). Originally, the Ad Council served as a nonprofit mechanism for supporting the war effort in World War II, creating many memorable public service announcements that raised $35 billion in war bonds and planting 50 million victory gardens. After World War II, President Truman asked the War Advertising Council, as it was known then, to continue with its mission. The newly constituted Ad Council went on to create well-known advertising campaigns such as Smokey the Bear, the Crash Test Dummies, and McGruff the Crime Dog.

The Ad Council originally established its Web site in 1995. The purpose was, and still is, to disseminate information and to enable media companies to preview and order public service advertising materials at no cost. The site has grown to the extent that today it provides detailed information on roughly 50 advertising campaigns. The Ad Council developed and maintains its Web site internally. It is fully interactive, allowing visitors to communicate with staff via e-mail link and Web-based form. This interactivity allows for visitors to perform a number of functions online, including the following:

- Donate online via secure Web-based form as well as by phone, fax, or regular mail
- Link Web forms to databases, giving managers the ability to use collected efficiently and effectively
- View current contributors based on contribution amount or name of contributor
- Order television, radio, newspaper, magazine, Internet, and out-of-home advertising directly through the Web site via Web-based form
- View the Ad Council's most recent ads via streaming video

The Ad Council Web site includes more than 100 pages and is organized into several broad categories associated with its internal organizational structure. These categories are links on each Web page and include the following:

- The Issues—Social issues important to the United States as defined by the Ad Council
- Campaigns—Advertising campaigns available for distribution to media and other outlets
- Organizations—Government or nonprofit organizations that support the Ad Council in its efforts
- Nonprofit Resources—Help for those who have developed or wish to enhance an existing campaign
- Make a Difference—Contributions and contributors pages
- About the Ad Council—An introduction to the Ad Council

The Ad Council Web site was developed by a contractor but is now maintained internally. The site has some unique features, the most interesting of which is the ability of users (media companies) to preview and order public service advertising materials at no cost. This means that a potential user may visit the site, preview examples (digital images, text, video) online, and immediately order a particular service via secure Web form.

The National Multiple Sclerosis Society

The National Multiple Sclerosis (MS) Society's (http://www.nmss.org/) mission is "to end the devastating effects of MS" (National MS Society, n.d.). The MS Society, through its nationwide network of chapters, supports research efforts, educates, provides a variety of programs, organizes fundraising events, and advocates for people with multiple sclerosis. The MS Society serves approximately 780,000 U.S. residents with MS and counts approximately 2 million volunteers, friends, and supporters among those who assist in the cause. It has chapters across the United States and is an affiliate of the Multiple Sclerosis International Federation.

The National MS Society Web site, established in 1995, is a comprehensive and detailed site with more than 200 Web pages. It is a fully functional e-commerce site, allowing the user to donate directly, pledge support for the MS Society via its MS Walks, volunteer, become a member, or learn about multiple sclerosis. Site visitors may communicate with the MS Society via e-mail, Web-based form, or secure payment form.

The Web site is organized around primary MS Society functions, including the following:

- Treatments—Describes existing treatments for MS
- Research—Provides emerging research regarding treatments and information for professionals
- Media—Discusses MS Society activities, news releases, and methods for the media to contact staff
- Education—Examines MS Society and affiliate education programs
- Donations—Provides, through secure server, mechanisms for visitors to donate directly
- My Community—Includes links to state and regional MS Society Web sites and contact information
- Member Services—Allows visitors to join the MS Society
- Special Events—Provides comprehensive information about MS Bike, MS Challenge Walks, MS Walks, including the ability to search by zip code for the nearest MS walk
- Advocacy—Outlines advocacy programs provided by the MS Society

The Web site also includes links to MS Society publications and offers content in Spanish. Intranet services are offered in the national office and in a number of the chapter offices.

In a conversation on May 31 2002, Abe Eastwood, information technology director of the MS national office in New York City, explained that the Web site is under constant revision. "We're working on several things, such as enabling 'members' to edit their own information, and integrating data from the Web site with our central CRM database."

The MS Society main Web site was developed by a contractor and is currently maintained internally. Many states and regions maintain their own Web sites, developed internally. These Web sites range from those with very few pages to comprehensive sites with nearly 100 pages. According to Eastwood, the Web site is undergoing a major revision and will be standardized along with the Web sites of all state and regional chapters. A vendor has been selected to accomplish this task and is currently working with some of the states to convert their sites.

RECOMMENDATIONS FOR SUCCESSFUL TECHNOLOGY IMPLEMENTATION

Developing a plan for implementing a Web and e-commerce strategy is an ongoing process. The plan should be an integral part of an organization's strategic plan and relate closely to an organization's communications plan. A comprehensive strategic plan is important whether or not the organization is developing a marketing strategy, embarking on a capital campaign, or launching a Web site. The following section outlines a strategic plan for implementing technology in a nonprofit organization.

Concept Evaluation

In any organization, it is important to have congruence between organizational mission and new and emerging ideas, particularly with respect to Web sites. As mentioned earlier, the Web site must serve the needs of the organization, not the other way around. When a Web site is developed without a direct connection to organizational mission (for example, fundraising or volunteerism), the Web site may detract from that organizational mission (Alsop, 1999, 167). Alsop notes, "it isn't enough that a Web site be well designed and treat my money appropriately. It has to provide actual, real-world service. I want my main vendors to have customer support just like the support you would expect from an airline or a courier service, and I want them to be competent about delivering products to me. That means the Web site can't be isolated from other parts of the business" (p. 167).

Define the Project

Project definition is a key piece of the puzzle in developing a technology plan. Although having a clear project definition may appear to be self-evident, often an organization will begin a process without completely understanding what it is that they want to accomplish and why they want to accomplish it. This is particularly important with respect to the Web and e-commerce. Web sites and e-mail servers are not to be constructed in a vacuum. In a nonprofit organization, Web sites are launched for two primary reasons: fundraising and volunteer recruitment. If these factors are given short shrift, the real reason for launching the Web site may be obscured.

Set Realistic Project Goals

Concurrent with defining the project, projects goals must be identified as well. What exactly is the Web site going to produce? Are increased revenues a primary goal of the Web site? Does the organization seek to increase the number of volunteers through online recruitment? Is more effective client communication a reason for adding an enhanced Web server? Each of these goals may be accomplished without upgrading or launching a Web site. Perhaps contributions are low because of staff misallocation or a lack of communication between the organization and contributors. The organization must be certain that e-commerce is the best way of increasing contributions or that a Web site is the best method of increasing volunteerism.

Specify Project Parameters

The scope and parameters of a project must also be identified. Once it is determined that Web enhancement or e-commerce are the best ways to accomplish organizational goals, project parameters must be set clearly. Parameters may be fiscal (fundraising goals, budget), human (staff, consultants, volunteers), technological (Web access, hardware, software), or knowledge-based (information availability, validity, and reliability). Fundraising goals are among those most important.

The organization must understand what client needs exist before identifying specific fundraising targets. Key to this a comprehensive understanding of the organizations' market. For example, as mentioned earlier, there exists a wide variety of nonprofits based on size, scope, and mission. The National MS Society has a national client base and would naturally seek funding from that base. According to Connie Nesbary, manager of clinical programs of the National Multiple Sclerosis Society Michigan Chapter, the society has clear knowledge of both the number of persons in the United States with MS and the levels of severity of MS that generally exist (June 3, 2002, interview). Combining this knowledge with an understanding of the cost of providing programs to people with MS and the role that the National Multiple Sclerosis Society plays in research, the MS Society can produce a relatively specific internal budget and a relatively specific fundraising target.

A local nonprofit has a smaller client base and would naturally seek funding from its community or region. The Spokane (Washington) Symphony Orchestra has a series of fundraisers and memberships designed to cater to and seek funding from patrons and local organizations (Spokane Symphony, n.d.). It maintains a Web site (http://www.spokanesymphony.org) on which a contribution page is located. According to Lori Mello, director of community education, HAVEN (Help Against Violent Encounters Now), raises money from several local resources as well as grants from state and federal government resources. HAVEN maintains a Web site (www.haven-oakland.org) and receives some funding through it (September 30, 2002, interview).

The organization must have a clear assessment of its human and technological resources before embarking on a major information technology (IT) initiative. Many small to medium nonprofits have limited or no dedicated information technology staff. All the goodwill in the world cannot overcome a lack of technical knowledge regarding constructing a Web site or adding e-commerce components to an existing Web site. These skills usually cost time and money. None of the small nonprofits interviewed for this chapter staffed a full-time technology employee, and the larger organizations maintained limited technology staff.

Establish Leadership and Management Structure

Regarding any project, someone must be responsible for project management and implementation. A Web site or e-commerce strategy is no exception. It is important to understand that implementing an Internet or e-commerce strategy is as much a management issue as it is a technology issue. As mentioned earlier, nonprofit technology staffing is often limited or nonexistent, so it can be difficult to identify appropriate staff to take a lead in implanting technology. In many organizations, the most well-respected, well-trained managers may have little or no technology experience (Kotrlik, Harrison, & Redmann, 2000), making it even more essential to identify internal or external support in identifying appropriate information technology management staff. Clearly, having a strong leadership portfolio and coherent management structure is essential.

Identify Stakeholders (Internal and External) and Procure Their Support

A whole host of stakeholders may be interested in a nonprofit Web site, including employees, managers, volunteers, contributors, patrons, participants, governmental agencies, students, researchers, and many others. It is important to determine which of these stakeholders may be of assistance in launching a Web site and ensuring its success. A comprehensive and relevant Web site is built on recommendations from all stakeholders.

Evaluate Project (Formative)

Continuous evaluation of the project is vital. Evaluation provides a useful exercise to ensure that the intent of the project is maintained and provides funding agencies (particularly governmental) confirmation that a thoughtful approach to the planning and execution of the project is being pursued. As part of formative evaluation, an initial budget must be developed. Understanding that the project is not fully developed, there are several ways to estimate a budget, including estimating the following:

- Which services will have an Internet presence
- The cost of contracting and paying a Web developer and Web host
- Internet connection fees
- Potential staff costs
- Balanced Scorecard approach

Of these tools, the Balanced Scorecard is a recent evaluative innovation and likely to hold promise for nonprofits because it "is comprehensive in that it considers the customer perspective, the internal business perspective, and the innovation and learning perspective, as well as the financial perspective" (Olson & Slater, 2002, p. 12). Generally, balanced scorecard approach may be tailored to a business's strategy, communicate strategic objectives, enhance feedback and learning, and, most intriguing to nonprofits, lead to enhanced revenues (Dalton, 2002).

Examine Primary Uses of a Web Site

Making the Web Work (2002) outlines a comprehensive list of reasons for nonprofits to develop a Web site, including the following:

- Membership development
- Fundraising, advocacy campaigns
- Getting feedback on programs and services
- Publicizing an event,
- Online registration for conferences
- Sharing best practices
- Updating mailing lists
- Freeform discussion and facilitated dialogue
- Managing collaborations
- Distributed publishing of calendars and new resources
- Highlighting successes
- Building awareness around an issue
- Displaying community assets through the use of online mapping
- Providing public access to searchable databases
- Cost-effective distribution of newsletters and other publications
- Marketing fee-based services
- Creating organizational photo or art galleries
- Building support through multimedia storytelling
- Addressing the needs of niche constituencies
- Enhancing media relationships
- Archiving frequently asked questions
- Creating the control center for an integrated internet strategy

Clearly, many nonprofits do not have the staff or funding to implement all of these tools and techniques, but understanding the universe of uses of a Web site is essential for getting the most "bang for the buck."

Create an Action Plan

Having a written strategy for implementing e-commerce in a nonprofit is solid management technique. An action plan usually asks the following questions: Who is responsible? For what are they responsible? When will the project be completed? and How much will it cost? An effective action plan may take also the form of a "task list," an annotated list of all tasks needed to accomplish or implement an action plan. A task list would include a time schedule for implementation that includes tasks, responsibilities, resources, and measures.

Environment Australia advises the Commonwealth Government on policies and programs for the protection and conservation of the environment, including both natural and cultural heritage places. It also manages a number of major programs, the most significant of which come under the umbrella of the Natural Heritage Trust. Environment Australia maintains an online action plan (Environment Australia Online, 2000), discussing in detail how it plans to implement its mission. Although this plan is specific to Environment Australia, it is general enough to give the reader an understanding of how an online action plan may be structured. The core components of their action plan are as follows:

- The nature of the agency and the environment in which the agency operates
- IT environment
- IT clients
- Services that the agency believes are appropriate to provide online
- Services that the agency believes are inappropriate to provide online
- Implementation strategies and timeframes
- Approaches to overcoming impediments to providing online services
- Legislative framework
- Evaluation and continuous improvement
- Strategies to ensure that the organization achieves agreed standards

Each step in the plan includes a description of how the step is to be implemented. It is concise and detailed enough to give the casual reader a clear understanding of what Environment Australia intends to do and how they intend to do it.

Finalize Funding

Once the action plan is completed, a final budget must be developed, accounting for the complete e-commerce and Internet strategy of the nonprofit, including increases in existing costs and new cost categories. Using a performance-based budgeting strategy is an effective tool to ensure that these costs are captured. Performance-based budgets require that a program provide some level of justification for base and increased funding. This is accomplished through the inclusion of performance measures in the budget document. Such systems require stronger roles for planning and goal setting and stipulate the conduct of performance measurement and reporting about program results (Willoughby & Melkers, 2001).

Implement the Action Plan

Once funding is finalized, the plan may be implemented. Implementation of an action plan needs to include a regular review of the action plan versus what is actually occurring. To the extent that the action plan implements e-commerce in a nonprofit organization, it is likely to be new or substantially revised, thus requiring significant oversight. Part of successful plan implementation is ensuring that minor problems do not become crises.

Another important piece of action plan implementation is ensuring that controls (fiscal and programmatic) are kept in place. This is of particular importance in the case of e-commerce- and Internet-related programs. Although implementing a Web site is ordinarily cost-effective, add-on pages or Web initiatives unrelated to the implementation of e-commerce may take place. If these initiatives are not part of the overall strategy, management must take a strong position to ensure that they are relevant, otherwise these add-on costs items must be rejected or pared back to the point that they are cost-effective.

Evaluate Project (Summative)

A summative evaluation is an important part of program implementation and review. Specific milestones need to be set for the evaluation of any new or revised program, and it is especially important to set evaluative milestones for an e-commerce initiative. Evaluating the extent to which the organization achieved these milestones is key to implementing a sound summative evaluation.

Carter McNamara (n.d.) of managementhelp.org provides an excellent synopsis of summative evaluation and why it is important to nonprofits:

> Program evaluation with an outcomes focus is increasingly important for nonprofits and asked for by funders. An outcomes-based evaluation facilitates your asking if your organization is really doing the right program activities to bring about the outcomes you believe (or better yet, you've verified) to be needed by your clients (rather than just engaging in busy activities which seem reasonable to do at the time). Outcomes are benefits to clients from participation in the program. Outcomes are usually in terms of enhanced learning (knowledge, perceptions/attitudes or skills) or conditions, e.g., increased literacy, self-reliance, etc. Outcomes are often confused with program outputs or units of services, e.g., the number of clients who went through a program.

Clearly, McNamara's typology fits well with the evaluation of an e-commerce initiative. Developing a Web site must comport with the nonprofit's organizational goals, as well as relate to the societal benefit that the nonprofit promotes.

CONCLUSION

Nonprofit organizations are moving gradually with respect to the implementing Internet technologies. The speed of Internet implementation is a function of a number of factors, including fundraising capacity, relevance, budget, management and staff capacity, and technology capacity. All of these factors are interrelated. For example, clearly Internet implementation has to be relevant to the function of the organization. Examples of this relevance are the many nonprofit Web sites with fundraising and volunteer recruitment Web pages. Without funding, these organizations cannot exist, and the Internet is a useful mechanism for raising funds. Furthermore, many young people use the Internet (e-mail, Web browsing, instant messaging) as their primary means of communicating with their peers. They are as likely to browse a volunteer recruitment Web page as they are to hear read or see a recruitment advertisement via traditional means. Hence, raising funds and recruiting volunteers relates directly to the relevance issue. Without a budget, however, the implementation of a Web site and e-commerce technologies cannot happen. Management has to be sufficiently knowledgeable regarding the Internet's relevant to the organization for a decision is made to fund a Web site or to add a fundraising Web page. Having internal staff with the capacity to create a new Web site or to enhance an existing Web site also is key regarding the likelihood that Internet technologies will be pursued by a nonprofit organization. Finally, nonprofit organizations must plan the implementation of Internet and e-commerce technologies carefully. Any significant change in an organization can be disruptive for employees, clients, volunteers, and stakeholders. Careful planning, implementation, and evaluation are key to successful organizational change.

GLOSSARY

E-commerce The application of communication and information technology to further the objectives of nonprofit, not-for-profit, business, and governmental organizations.

Foundation An organization created from dedicated funds from which income is distributed as grants to nonprofit organizations and, in some cases, to individuals.

Homepage the Web page designated to be the first page seen upon entering a Web site.

Intranet A method of connecting multiple computer systems over a relatively small area using Internet technologies; an Intranet can be as small as two computers in one room to as large as 1,000 or more computers in a large organization.

Nongovernmental organization Any private or nonprofit organization, usually referring to those providing support to citizens ordinarily provided by governmental units.

Nonprofit corporation A corporation not intended to operate for profit and that is barred by law from making a profit.

Not-for-profit An activity conducted without intent to raise money; the term may refer to hobbies, informal groups, and organizations not organized to make income.

Pledge A promise to make future contributions to an organization; for example, some donors make multiyear pledges promising to grant a specific amount of money each year.

These definitions were generated from the following source documents:

IUPUI Payton Philanthropic Library. Retrieved June 11, 2002, from http://www.ulib.iupui.edu/special/dict.html

Nesbary, D. (2000). *Survey research and the World Wide Web*. New York: Allyn & Bacon.

Dewitt, J. D. (2001). *Volunteer legal handbook* (6th ed., rev.). Guess & Rudd P.C. Retrieved June 12, 2002, from: http://iciclesoftware.com/vlh/VLH6Glossary.html

Council on Foundations. (n.d.). Retrieved June 11, 2002, http://www.cof.org/glossary/index.htm

CROSS REFERENCES

See *Digital Divide; Internet Literacy; Intranets; Politics.*

REFERENCES

Ad Council. (n.d.). About Us. Retrieved June 3, 2002, from http://www.adcouncil.org/fr_about.html

Allen, D. (2002). Lesson 169: Edge side includes. *Network Magazine, 17,* 22, 24.

Alsop, S. (1999). How I judge if a Website deserves my business. *Fortune, 40,* 67–68.

Big Brothers Big Sisters of Metropolitan Detroit. (n.d.). About us. Retrieved May 2, 2002, from http://www.bbbs-detroit.com/aboutus.html

Blau, A. (2002). Access isn't enough. *American Libraries, 33,* 50–52.

Cruver, P. (2001). Broadband the future of learning. *Multimedia Schools, 8,* 28–29, 31.

Dalton, J. (2002). Strategic score-keeping. *Association Management, 54,* 53–57.

Donaldson, T. (2001). Ethics in cyberspace: Have we seen this movie before? *Business and Society Review, 106,* 273–291.

Environment Australia Online. (2000). Action Plan. Retrieved June 3, 2002, from http://www.erin.gov.au/about/publications/online3oct.html

Fike, J. (2002, Spring). Technology, media and fundraising survey of southeast Michigan nonprofit organizations. *Detroit Chapter of Association of Fundraising Professionals Newsletter, 1,* 4.

Food for the Hungry. (n.d.). Retrieved April 27, 2002, from http://www.fh.org/learn/aboutus.shtml

Gardner, C. J. (1998). Nonprofits tap new donors with Internet fundraising. *Christianity Today, 42,* 26–27.

Geneforum. (n.d.). About us. Retrieved March 20, 2002, from http://www.geneforum.org/organization/

Independent Sector.Retrieved May 14,2002, from http://www.independentsector.org/e_philanthropy/schedule.html

Internet infatuation filters down to Italy: Europe's Internet capitals. (2000). *Advertising Age International, 29,* 40.

Kotrlik, J. W., Harrison, B. C., & Redmann, D. H. (2000). A comparison of information technology training sources, value, knowledge, and skills for Louisiana's secondary vocational teachers. *Journal of Vocational Education Research, 25,* 396–444.

Making the Web Work. (2002). MTNW toolbox. Retrieved April 3, 2002, from http://www.makingthenetwork.org/tools/index.htm

McNamara, C. (n.d.). Evaluation activities in organizations. Retrieved May 21, 2002, from http://www.managementhelp.org/evaluatn/evaluatn.htm

Mertler, C. (2002). Demonstrating the potential for Web-based survey methodology with a case study. *American Secondary Education, 30,* 49–61.

Montagu-Pollock, M. (2000–2001). A landmark year—but not a smooth one. *Asiamoney, 11,* 31–32.

National Multiple Sclerosis Society. (n.d.). Frequently asked questions. Retrieved May 3, 2002, from http://www.nationalmssociety.org/NMSS-FAQs.asp

Nesbary, D. (2000). The taxation of internet commerce. *Social Science Computer Review, 18,* 17–39.

Notovny, P. (2002). Local television, the world wide Web and the 2000 presidential election. *Social Science Computer Review, 20,* 59–72.

N-Ten. (n.d.) Retrieved October 1, 2002, from http://www.nten.org/capacity-map

Oehler, J. E. (2000). Not-for-profit organizations can profit by investing in the Internet. *The CPA Journal, 70,* 65.

Olson, E. M., & Slater, S. F. (2002). The balanced scorecard, competitive strategy, and performance. *Business Horizons, 45,* 11–16.

Pearson, T. (2002). Falling behind: A technology crisis facing minority students. *TechTrends, 46,* 5–20.

Pliskin, N. (1998). Explaining the paradox of telecommuting. *Business Horizons, 41,* 2.

Riedman, P. (1998). Nonprofit ads find a home on the Internet. *Advertising Age, 69,* 14.

Souccar, M. K. (2001). Fund-raising techniques see fundamental changes. *Crain's New York Business, 17,* 7–12.

Strom, S. (2002). Nonprofit groups reach for profits on the side. *New York Time March, 17, 2002 Late Edition Final.* Retrieved February 28, 2003 from http://80-web.lexis-nexis.com.huaryu.kl.oakland.edu/universe/document?_m=5d988f49139db6a7836b9c1314dfb879&_docnum=1&wchp=dGLbVzz-lSlAl&_md5=03e5f5ab0ba802913253111b302bc36d

Vargas, A. (2001). Providing the connections to make MOUSE roar. *Crain's New York Business, 17,* 50.

Waldon, C. D. (1999). NetAid: Music and technology join forces to battle poverty. *Choices, 8,* 13–16.

Westen, T. (2000). E-democracy: Ready or not, here it comes. *National Civic Review, 89,* 217–227.

Willoughby, K. G., & Melkers, J. E. (2001). Assessing the impact of performance budgeting: A survey of American states. *Government Finance Review, 17,* 25–30.

Witte, J. C., Amoroso, L. M., & Howard, P. E. N. (2000). Method and representation in Internet-based survey tools: Mobility, community, and cultural identity in Survey 2000. *Social Science Computer Review, 18,* 179–195.

FURTHER READING

Blazek, J. (2000). The Internet and tax-exempt organizations. *The Tax Adviser, 31,* 344–353.

Davis, J. (2000). Dot-com casualty list mounts as holiday season approaches. *InfoWorld, 22,* 32.

Toffler, A. (1980). *The third wave.* New York: Bantam Books.

Yetman, R. J. (2001). Tax-motivated expense allocations by nonprofit organizations. *The Accounting Review, 76,* 297–311.

The following is a short list of Web links to nonprofit Web directories and search engines. These Web links are useful to anyone interested in examining nonprofit Web sites or simply searching for nonprofits by category in the U.S. or around the world.

Action Without Borders—Idealist.org (http://www.idealist.org/). Includes more than 28,000 organizations and 6,000 volunteer opportunities

Independent Sector (http://www.indepsec.org/). A coalition of nonprofits, foundations, and corporations providing information dissemination and research services

IT Resource Center (http://www.itresourcecenter.org/). A national resource for nonprofit training, jobs, and research

The Nonprofit Resource Center (http://not-for-profit.org/) Includes links to nonprofits, nonprofit boards, nonprofit support organizations, fundraising resources, and marketing information

Philanthropysearch.com (http://www.philanthropysearch.com/). Includes searchable database of Web sites including fundraising, education, social services, publications, and charitable giving

Planet Volunteer (http://www.planetvolunteer.com/). Allows user to search for volunteers or volunteer organizations

Servenet.org (http://www.servenet.org/. Primarily a volunteer organization focusing on youth and young adults

Union of International Associations (http://www.uia.org/). Provides information on intergovernmental organizations and nongovernmental organizations

Yahoo Nonprofit Resource (http://dir.yahoo.com/Society_and_Culture/Issues_and_Causes/Philanthropy/Nonprofit_Resources/Web_Directories/)

Yahoo Volunteerism Links (http://dir.yahoo.com/Society_and_Culture/Issues_and_Causes/Philanthropy/Community_Service_and_Volunteerism/)

O

Online Analytical Processing (OLAP)

Joseph Morabito, *Stevens Institute of Technology*
Edward A. Stohr, *Stevens Institute of Technology*

INTRODUCTION

Management reporting is the earliest form of wide-scale decision support application and typically employs the data structures used for the organization's transaction processing. Management information systems (MIS) provided the necessary technology in early batch processing systems. As online data entry and access became available in the 1970s, transaction processing systems became known as online transaction processing (OLTP) systems and represent the bulk of a typical organization's systems. To maintain data consistency during transaction processing, OLTP designs emphasize the notion of data normalization in relational databases. Normalization of a company's data means essentially that each database table stores data about instances of a single entity (e.g., all customer instances) or of a single event (e.g., all purchase transaction instances.) The objective of normalization is to reduce data redundancy and streamline update transactions. In satisfying these objectives, OLTP systems sacrifice ease of reporting.

Generating reports from a relational database may involve joining information in multiple tables at a huge computational cost for databases of any size. This type of reporting concentrates on detailed listing of records and is well-suited for managing the bottom line and may include, for example, performance ranking of sales people and other such reports. The majority of reports in such systems are preformatted (sharing this characteristic with MIS systems). Sophisticated users may generate ad hoc reports using the database's query language.

With the advent of integrated transaction processing systems (e.g., enterprise resource planning; ERP), the back-end database design did not change in philosophy; it only became larger and more complex. Such designs are sometimes referred to as operational data stores (ODS) and represent integrated data stores of several applications (i.e., an integrated set of separate OLTPs). This environment is well suited to support operational queries. Such queries cut across functional applications and give an end-to-end readout of an event or situation. For example, a brokerage house may want to resolve a problem trade with a client and would need to run a report that displays the life history of a given trade: the order, the trade, its clearance, its settlement, and any final positioning information for the client. With traditional OLTP applications this would be a series of reports run and manually integrated by operations personnel covering hours or even days; with an integrated operational data store, this would be a single query executed by the client's broker or a centralized help desk.

The definition of an operational data store is somewhat slippery. We have seen the term used with reference to a data structure supporting a single OLTP application as well as an integrated operational data structure supporting integrated applications, including operational queries (e.g., the brokerage example). In a data warehouse environment, an ODS means some set of data that is a consequence of data extraction from OLTP data stores. The ODS typically is not accessed directly by users but serves as a staging area prior to making data available to users through data marts. In this sense, the ODS may represent a kind of enterprise data warehouse. An ODS may also serve as a staging area in the development of what Bill Inmon has called an "atomic" data warehouse (Inmon, 1996). This is a large data warehouse of detailed data that is only lightly categorized and summarized. In either case, an ODS is a staging area that supports data cleansing prior to data delivery to users. Data cleansing includes activities such as removing duplicate records, correcting invalid attribute values, and so on. Finally, the structure of an ODS in a data warehouse environment usually takes one of two forms: the first is as an integrated and normalized data model, and the second is merely a series of file structures corresponding to source data structures. The first implementation may require extensive data transformation and integration, while the second requires little if any transformation. The second implementation is more common than the first.

Systems that rely primarily on preformatted reports are inadequate for a world dominated by the need for speed and characterized by constant change. With the advent of the Web, users' demands for a "business intelligence" (BI) capability in organizations have increased dramatically. By BI, we mean the full set of activities required to analyze and capture, transform and derive, store and distribute data through applications and specialized tools to support analysis and decision making. (The forerunners of BI systems were called either MIS or executive information systems [EIS] depending on the level in the organization of the target user.) Please note that some organizations separate data warehousing and BI activities: the former refers to the development of data warehouses and marts, while the latter refers to the development of BI tools and applications that run against the data warehouses and marts to perform analyses that are critical to the firm.

For BI to succeed, users need to examine data at various levels of aggregation and to take different views of the data along multiple dimensions such as product sales by district over time or sales to customers by customer type and geographic region. Moreover, users should be able to "drill down" seamlessly from high-level aggregate data to more detailed levels of data, perhaps even to the transaction level. Online analytic processing (OLAP) systems provide these capabilities and more, thus providing analysts with a flexible means to examine their organization's data and to discover relationships and trends that less flexible tools could not reveal (Inmon, 1996). OLAP tools are therefore an indispensable component of any data warehousing and BI strategy.

OLAP represents the latest evolution in decision support. While traditional reporting is positioned to manage the bottom line (with OLTP data structures) and operational queries to enhance customer service and save costs (with ODS data stores), OLAP is designed to promote the top line by improving strategic decision making and customer relationships. In contrast to OLTP and ODS data schemas, OLAP employs multidimensional data stores. Traditional reporting and operational queries reference data structures designed primarily to support other types of processing (i.e., transactions) and hence support decision making incidentally. It is not surprising that these reports support simple analyses and decision making. In contrast, multidimensional data stores are designed specifically for OLAP, making analyses and decision support far richer than otherwise. Following is a brief review of selected differences between OLTP and OLAP systems.

OLTP Systems

- Each transaction execution involves a small number of logical records (usually one).
- Small number of transactions types.
- Large number of transaction executions.
- Comparatively static requirements and data schema.
- Involves primitive database operations (i.e., create, read, update, and delete; CRUD).
- Contains atomic, detailed data.
- Usually restricted to the current time period (e.g., present point in time going back for some short period of time such as three months).
- Performance and throughput comparatively important.

OLAP Systems

- Each transaction execution may involve a very large number of logical records.
- Large number of different types of queries (including ad hoc queries) and applications.
- Small number of long-running transactions.
- Comparatively dynamic analytical requirements and data.
- Generally read-only, but increasingly supporting dynamic calculations on the fly (e.g., supporting what-if scenarios).
- Contains value-added data (e.g., categorized, summarized, and derived data).
- Includes a wide range of historical time periods (e.g., often 7 years for legal applications).
- Performance and throughput relatively less important.

Before concluding this section, it should be noted that OLAP and data mining tools are complementary. Both tools are enabled by the existence of data warehousing capabilities in the organization, because both tools draw their data from sources external to themselves. Data mining involves discovering hitherto unknown patterns or relationships in data using an array of techniques from the field of statistics as well as from computer science and artificial intelligence. Often, the results of a data mining investigation are expressed as rules, for example,

IF Customer Visits to Product Page > 3 THEN
probability of purchase > 0.5.

This rule conforms with common sense and is also "actionable"—the company might offer such visitors to its Web site a discount on the particular product of interest.

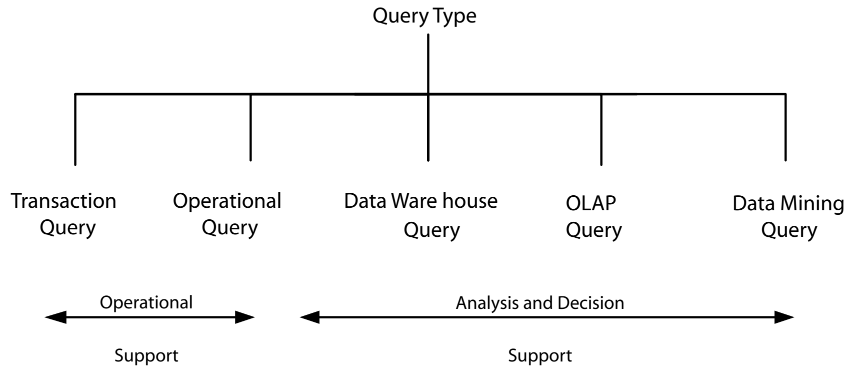


Figure 1: Varieties of query types.

Both OLAP and data mining tools are used to discover patterns in data in a broad sense. The differences are great, however. The OLAP philosophy views data as structured in terms of multiple dimensions such as time, product, or customer. The patterns that may be discovered in the data by the astute OLAP analyst may be useful, but they are somewhat constrained by the overall multidimensional structure. OLAP tools can be used by anyone possessing sufficient business knowledge and the desire to explore the data. In their present state of development, however, data mining tools are best used by data mining specialists in conjunction with business experts.

Generally, there is a continuum of queries that cover the range of OLTP, OLAP, and data mining (see Figure 1). The difference between the queries is not always clear-cut (Hand, Mannila, & Smyth 2001). For example, we may be interested in the address of a client or in uncovering the audit trail of a lost stock trade. The first is a transaction query, and the second is an operational query that, in the absence of an ODS or a data warehouse, would cover several separate operational databases. The distinguishing characteristic of these two queries is that they involve logically a single record and fall into the category of an OLTP query. At the other end of the continuum, an example of an OLAP query may be finding the sales of all products by month and region, while a data mining query may be finding what products have similar sales patterns or finding rules that predict the sale of a certain product using customer segmentation or clustering (Hand et al., 2001). An OLAP query is distinguished by aggregation, whereas data mining involves pattern discovery and rule development. Both OLAP and data mining involve a large number of logical records. The data warehouse query lies in the middle of the continuum and best exemplifies the unclear and rapidly evolving query environment. An example of such a query would be finding the trend in sales of a particular product in a particular region. This involves only light aggregation, and although it may be implemented as an OLAP query with a restricted data view (e.g., virtual OLAP), optimum implementation is with a reporting and query tool against a data warehouse. The variety of query types represent the changing nature of query requirements, query tool capabilities, and the emerging role of an integrated business intelligence environment as a support, even an "operational" system for the knowledge-based work that defines the 21st-century organization (Kochan, Orlikowski, & Cutcher-Gershenfeld, 2002).

Figure 2 illustrates a full-capability business intelligence architecture. It shows the relationship of the various constructs we have been discussing: operational data stores, data warehouses, data marts, data mining, and OLAP technologies.

There are several observations that we can make from this illustration, among them the following:

The BI environment includes an interrelated mix of platforms, data structures with different schemas, end-user tools and technologies, and applications.

The BI architecture is characterized structurally by a kind of back-end data capture, a central storage area, and front-end data distribution. This structure supports the following functions in the life history of a data element as it moves through the architecture: data identification and access; data capture, cleansing, rationalization and transformation; data derivation, categorization, and calculation; development of supporting meta-data; data distribution, reporting, OLAP, and mining. It should be noted that not all functions fit neatly in one particular place in the architecture. For example, data elements may, and often are, derived and calculated anywhere in the architecture. In fact, as discussed subsequently, one of the key characteristics of OLAP is the ability of an end user to derive and calculate data on the desktop as he or she sees fit.

Organizations tend to be data rich but information poor. The BI environment addresses this issue head-on by increasing the richness of data, and thus creating information, as it flows through the architecture. Furthermore, an end user may access data at any level of richness (i.e., specific information), depending on the analysis.

The BI architecture and technologies support several ancillary functions, such as the following:

- Identifying data inconsistencies within and between source systems.
- Serving as a bridge to combine data across legacy systems and create enterprise-level information.
- Identifying, recording, and operationalizing business rules.
- Empowering workers and decentralized decision making through the widespread, even organizationwide, distribution of information, made possible by the combination of OLAP and reporting technologies with the Web.

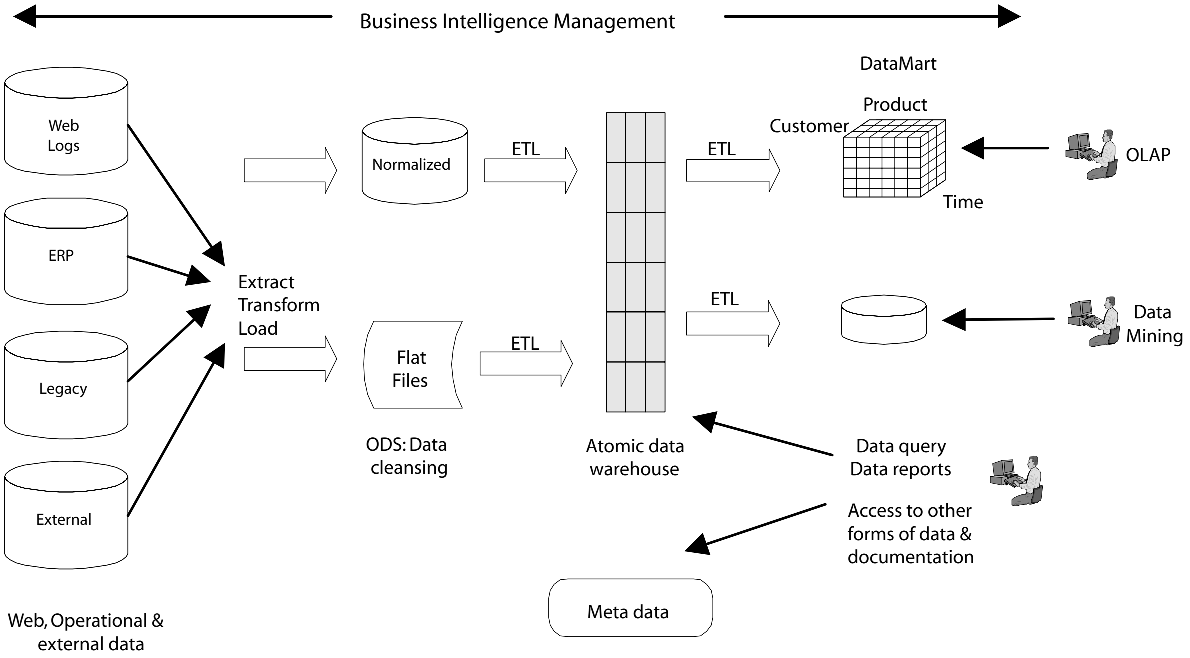


Figure 2: Relationship between ODS, data warehouse, data mart, OLAP, and data mining.

- Promoting knowledge work by capturing and distributing information and giving each worker the ability to further manipulate data and create additional information.

These are concerns for most organizations. The BI architecture is an opportunity to address each of these issues in an integrated and systematic manner.

OLAP TECHNOLOGY

Observe from Figure 2 that as the data moves from left to right, it increases in value. In classical information-processing theory, this is known as increasing the "information carrying capacity of data." This is accomplished through the application of explicit knowledge: data is turned into information with increasing richness as we progress along the "information continuum," from raw, atomic data to derived and transformed multidimensional data (Morabito et al., 1999). Data has its information-carrying capacity increased through semantic manipulation (e.g., derivation, categorization, summarization, etc.) and data visualization (e.g., graphics, traffic lighting, etc.), supported with multidimensional data structures. A multidimensional data structure is one in which data are stored in a form that makes interactive, spreadsheet-like analysis possible. Thus, an OLAP capability takes shape as we enrich data and then put it in a form amenable for interactive analysis and manipulation.

Data Capture and Cleansing

Data is initially accessed from "certified" sources of operational data, including weblogs and external data, extracted and stored in an ODS, either as an integrated and normalized database or as flat files (e.g., comma delimited, of fixed field size files). In the ODS, it is cleansed and then extracted, transformed, and loaded (ETL) into the first-level, or atomic-level, data warehouse. Typically this data warehouse is large and contains detailed data that is only lightly categorized and summarized. The atomic data warehouse is usually stored in a relational DBMS and is organized as facts (i.e., measures) and dimensions in a "star schema." A star schema organizes data as facts and corresponding dimensions in a kind of hub-and-spoke configuration. Each dimension is denormalized and captures all the attributes associated with that dimension. Variations to the star schema include a snowflake design: a snowflake schema is a star schema with selected dimension attributes normalized, particularly a dimension with an embedded hierarchy—it's the repeating values of the hierarchy that are normalized. Normalized dimension tables, whether or not part of a snowflake, are sometimes known as "outrigger" tables (Kimball, 1996). The star schema is suitable for reports and queries and also supports simple OLAP (virtual) activities as well as ROLAP (OLAP technology with a relational database).

It should be noted that the ETL function is applied between data layers or stores within the BI architecture—between source data stores and an atomic data warehouse, between an atomic data warehouse and a data mart, and so on. ETL functions are facilitated with tools; the primary tool is an ETL tool, but this technology usually needs to be supplemented with CASE tools and data dictionaries. Most ETL paradigms and corresponding tools are meta-data driven: the source and target data elements are identified at the logical (optional) and physical (required) level, while transformations and derivations between source and target data stores are defined with rules as meta-data. The actual implementation may be in the form of code generated by an ETL tool (early generation) or in an "ETL-engine" that moves and transforms data between data stores. Also, ETL tools may be either batch- or real-time oriented; in the former case, the load function is usually separated from extract and transformation and applied with the batch-load utility of the target database. This is done for performance reasons.

Data Distribution

To implement a full OLAP capability, we need to add more richness to the data than is typically contained in an atomic data warehouse. Therefore we execute an ETL process between the atomic data warehouse and a target data mart. A data mart is a small, focused data warehouse usually deployed at the departmental level, which usually involves more data transformations and derivations than its source data warehouse. The data mart may be organized as a star schema in a RDBMS, or organized along a "cube" metaphor in a proprietary database known as a multidimensional database (MDDB). The data mart contains additional categorization and derivations—again, its richness has been increased.

The combination of OLAP and reporting technologies with the Web allows end users easy access to the atomic data warehouse, data marts, and meta-data. The structure of the data largely dictates the type of OLAP access. For example, the OLAP technology that is used with a star schema in a relational database is known as ROLAP, whereas that used with an MDDB is known as MOLAP. Each of these areas is elaborated in the following section.

Varieties of OLAP

OLAP comes in a variety of flavors (see Figure 3). Each type will have an impact on its position as an e-commerce tool and its implementation in the BI architecture. We now briefly review the different types of OLAP.

OLAP

OLAP is the generic software technology associated with multidimensional analysis. OLAP involves aggregate data, and therefore multidimensional databases are sometimes deployed as data marts derived and aggregated from a source data warehouse. Although traditional reporting and queries are static and use a list or table metaphor, OLAP is distinguished by its interactive nature and by use of a spreadsheet metaphor. The spreadsheet metaphor is a simple data visualization technique that is highly effective in manifesting data patterns and graphics. The spreadsheet metaphor does not imply two-sided analysis, but rather an *n*-dimension analysis. Furthermore, each "cell" in the spreadsheet may contain any number of measures or "facts" for the same set of dimensions, for example, assets, commissions, number of trades, and so on. OLAP is associated with a number of advanced analytical and graphical capabilities as discussed later.

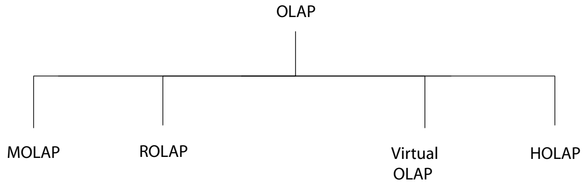


Figure 3: Varieties of OLAP.

MOLAP

MOLAP refers to OLAP technology with a proprietary database known as a multidimensional database (MDDB). The metaphor of an MDDB is a cube—an *n*-sided spreadsheet data structure. Each "side" of the cube is a dimension (e.g., time, product, location), and each cell is a measure, such as sales. The primary advantage of an MDDB is that it is extremely fast. Moreover, MDDBs often support dynamic derivations and calculations—an OLAP reference to a dynamic-defined fact will invoke the calculation. The disadvantages of MOLAP include a relatively small number of dimension attributes, an unfamiliar technology, and size limitations. MDDBs are typically deployed as data marts and often include complex derivations and aggregations as measures.

ROLAP

ROLAP refers to OLAP technology with a relational database. The back-end database is designed with a multidimensional schema for relational databases known as a star schema and its variations such as a snowflake. The star schema facts correspond to the measures in the cells of an MDDB and its dimensions correspond one-to-one to the sides of an MDDB cube. ROLAP's primary advantages include practically no size limitation, a large number of dimension attributes, and familiar relational technology. Its disadvantage is that it is significantly slower than MDDBs.

Virtual or Desktop OLAP

Virtual OLAP has a narrow focus, encompassing relatively few dimensions and measures. Virtual OLAP is operationalized through a simple reportlike query against a relational database (i.e., star schema) that returns a "micro cube" in memory. The user then has a subset of OLAP features available such as drill, pivot, and rotation. Other features are not available; for example, bookmarking only makes sense with "real" data structures. When the user is finished with his or her analysis, the micro cube is simply removed from memory. This form of OLAP is sometimes known as "desktop OLAP." One advantage of virtual OLAP is that the user receives only those facts and dimensions in which he or she is interested.

HOLAP

Hybrid OLAP refers to a technology in which a query optimizer references two data structures: one is a star schema and the other is an MDDB. In essence, HOLAP exploits the benefits of both MOLAP and ROLAP. The user submits a query request and the optimizer selects the most suitable data source, the relational star or the MDDB. The purpose of the hybrid approach is to optimize the data source and to promote ease of use. This rationale stems from the fact that most large-scale decision-making environments include both relational star schemas (in the form of a data warehouse) and MDDBs (in the form of derived marts). The hybrid approach optimizes the utility of both types of data structures.

A Word on Computer Platforms

Operational data sources may reside on a variety of platforms, but in practice most large organizations have their core operational systems on mainframes. The ODS (as either an integrated data base or as a series of flat files) may reside on either mainframe or client servers. If most of an organization's source data reside on the mainframe, however, it makes sense to implement the ODS on the mainframe as well. In this way, one may harness the power of the mainframe for data cleansing. The atomic data warehouse gives us an opportunity to implement on a client-server machine, but the mainframe may also be used. With a MDDB, the target platform is virtually always client-server. Finally, the users may access data (reports, virtual OLAP, ROLAP from the atomic data warehouse, and MOLAP from MDDBs) from a Web server. Thus, as with the enrichment of data along the information continuum, the platforms also shift in emphasis, from mainframe to client-server to the Web and the user's desktop.

THE OLAP DATA VIEW

Regardless of their underlying physical data structure, OLAP tools project a multidimensional view of data. The user is able to visualize the data as consisting of cells in a multidimensional cube. The dimensions of the cube (such as product, customer, and time) serve to identify, qualify, and describe facts in the multidimensional space. As mentioned earlier, a cell in the data cube might represent sales of a particular product in a particular region at a particular time period. The content of a particular cell is often called a "fact." Actually, each cell in the data cube may contain multiple facts. For example, each cell in the Product × Customer × Time cube in Figure 2 might contain data concerning sales, cumulative sales to date, contribution, margin, price, cost, and so on. A powerful feature of many OLAP tools is that the dimensions can support hierarchical views of the data, with each level in the data structure representing a particular level of aggregation. For example, the Product dimension might consist of All Products, All Product Classes, and All Products Within Each Class. This obviously facilitates viewing the data at multiple levels of integration. As the user picks a different level in each dimensional hierarchy, the values in the cells of the cube are automatically adjusted. For example, if the user wishes to perform an analysis at the highest level of aggregation representing all products, the cells of the cube will reflect total organizational sales, costs, etc. for the particular customer and time periods. As we have implied, it is possible to aggregate at different levels on each dimension of the cube. The hierarchical view of the cube's dimensions naturally facilitates "drill-down" capabilities. For example, if total sales of a particular product in some time period are unexpectedly high or low, the user can drill down to lower levels on the customer dimension to discover the particular customer classes or even customers that are responsible for the unexpected result.

OLAP tools also handle time in a sophisticated manner. For example, in the time dimension, most tools can manage calendars and calculate durations automatically. Mechanisms are also provided to handle automatically accumulations such as "sales to date" and common comparisons such as "sales this month last year" versus current sales, and budgets versus actual.

Meta-data

Of particular importance is the capability of the OLAP tool to handle meta-data. For example, it is essential for proper decision making for the analyst to understand the definitions of data: When is revenue recognized by the firm's accounting system? When was the data generated? By whom? and etc.

This concept of meta-data is particularly important in a BI environment. In the past, meta-data was meant primarily for the database administrator. It served to help the DBA design a database. We call this technical meta-data, which includes items such as data length and type.

In a BI environment, however, the primary users of meta-data should be the business user. For analysts and decision makers to be effective, not only must the data they manipulate be accurate and timely, but their understanding of the data must also be accurate. For example, the concept of a "household" in a bank can take on many meanings—a billing versus a marketing household for example. Do we include everyone living in a dwelling as belonging to a single household? What of children away at college? As another example, when calculating "assets under management," different departments will "roll up" assets differently for each product category—departments often define products differently. In fact, for most every data element there are different and equally correct perspectives; what distinguishes one from the other is the business context. Thus, a business user must be able to not only see meta-data but also manipulate it. A given business user should be able to select the appropriate context (e.g., a marketing perspective of household) and then see the appropriate breakdown and rollups.

In addition to manipulating meta-data, a user should be able to see the complete audit trail of every data element to which he or she has access. This would include source systems, data steward, transformation algorithms, associated business rules, date of last refresh, and so on. A user should also be able to access a library of existing reports, queries, and OLAP "bookmarks" for every data element; in this way a user has an opportunity to reuse existing analyses or OLAP navigation paths. This is all meta-data.

Finally, meta-data also includes other forms of supporting documentation for a particular perspective. This includes text documents, presentations, and navigation to internal Web sites for additional information. Thus we see that meta-data is a complex construct: It includes traditional technical meta-data as found in a data dictionary and business meta-data found in a variety of places and forms. The former concerns issues of data structure, and the latter concerns data context. To be effective, each kind of meta-data should be a "click away."

Sparsity

MDDB cubes tend to grow asymmetrically. By this we mean, for example, that not every product is sold in every store every day, every student does not take every course every semester, and so on. Thus, we often discover that the cells of a cube are empty as the cube grows in size. In a star schema, the fact table is said to be sparse in that we do not add rows in a fact table if there is no activity.

A second variation to the concept of sparse data concerns repeating values; for instance, the price of a product may be the same in every store every day for some period of time, and then the price may change everywhere simultaneously. In each case, sparsity is not an issue as such but may have an impact on size, performance, and certain types of calculations such as "average." The better MDDBs and corresponding OLAP tools will have some sort of compaction routine that avoids repeating duplicate values; instead, they record a given value and the range in which the value is repeated. In a star schema, the solution is not so straightforward. If we no not add rows to a fact table when there is no activity, the SQL AVERAGE function will yield incorrect results. The solution is to use the SQL SUM function and divide the results by the appropriate number of instances (i.e., compute SUM and cardinality separately and then combine to calculate the average). Unfortunately, this must be programmed and does not support end-user report development. One solution is to add rows containing "zero" values; this, in effect, eliminates sparisty and solves the "average" computation problem, but most likely creates a huge fact table—affecting both size and performance. Another solution is to not add rows containing zeros, but to choose an appropriate end-user query tool that meets this and other computing criteria beyond supporting structured query language (SQL) queries. (See Adamson & Venerable, 1998, for a discussion on sparsity and computational issues.)

FUNCTIONALITY OF OLAP TOOLS

OLAP tools combine ease of use with sophisticated functionality. This functionality was implied in the previous discussion but is discussed in more detail here. Generally, report formatting is secondary to aggregation and data visualization. OLAP tools do not emphasize traditional list reports, but rather charting and the display of tabular data.

Report Capabilities

The most basic and necessary functionality involves report formatting capabilities. Most OLAP tools will automatically format tabular views showing titles and row and column labels and display multiple dimensions on each axis in a visually clear manner. Moreover, they will automatically fit the report display to the user's screen or provide simple horizontal and vertical scrolling capabilities. OLAP tools will similarly format and paginate printed reports.

Pivot or "Slice and Dice"

One way to think about this class of functionality is to first imagine the underlying cube data structure. OLAP tools permit the user to navigate up and down (drill up or drill down) and around selected dimensions (pivoting)—in effect, a user may select which portion of the cube to view by merely rotating dimensions. Pivoting means that a user selects which dimension to display along the x-axis, which dimension to display along the y-axis, while selecting and holding constant the remaining dimensions. For example, in a cube that contains sales as a measure and product, region, and time as dimensions, a user may select a given product category and display, for that product category, regions (along the y-axis) versus time (along the x-axis). If the user then wanted to change perspectives, he or she would, for example, swap product and region: In this case, the user would have to select a given region and display, for that region, products (along the y-axis) and time (along the x-axis).

Nesting Multiple Dimensions

One of the more popular extensions of the pivoting capability is to nest dimensions along one axis. Although pivoting requires, for example, a user to swap dimensions and display one dimension along each axis, nesting permits a user to display several dimensions along a given axis. For example, in the scenario just described, rather than swapping product and region, a user may drag-and-drop product underneath region: In this case the y-axis would display product categories within each region while the time dimension remains displayed along the x-axis. This is a particularly powerful feature, provided the user doesn't get carried away and overload a display with too many nested data points.

Calculated (Virtual) Data

A powerful feature of OLAP is user-defined, automatic calculation of dates and durations as well as other arithmetic operations. For example, users may define a dimension as actual–budgeted and then create "on the fly" a new attribute called "variance" and define it as the difference between actual and budgeted measures. Similarly, users can define a dimension as actual–budgeted–projected and engage in a variety of what-if analyses.

Graphical Capabilities

As one would expect, the spreadsheet display associated with OLAP lends itself to the automatic generation of multidimensional bar and pie charts, line plots, and so on. The key ingredient here is the selection of the appropriate graphical display; for a given spreadsheet display, one type of chart may be more effective than another, whereas others may make no sense. For example, in the scenario we have been discussing, a display of sales for region versus product lends itself to a bar chart; a line graph would make no sense. Similarly, a display of sales for region versus time lends itself to a line graph; a bar chart would be possible but not be nearly as rich as the line chart. Thus, the selection of the appropriate visual display adds richness to the underlying data.

Drill Through

This is not a form of OLAP but a technique to connect an MDDB with its source relational star. The central concept here is that a user engaged in OLAP with an MDDB may need to drill down to a lower level of detail (i.e., grain) than a cube typically contains. The technology supports "drilling through" from an aggregate or derived cell in an MDDB to its corresponding detail records in the relational star from which the data in an MDDB was sourced and aggregated or derived.

Multiple Hierarchies

A multiple hierarchy is one in which a "child" element has more that one "parent." For example, a location dimension may have two hierarchies: one may be geographic (salesperson to store to city to state), whereas the second may be regional (salesperson to store to district to division). In this example, district and division are independent of city and state boundaries. As another example, the time dimension contains at least two hierarchies for most organizations: day to month to quarter to year and day to week. In each of these cases the key is to recognize that each hierarchy is logically part of the same dimension. The better MDDBs and corresponding OLAP tools support multiple hierarchies for a single dimension. In the absence of this direct support, the OLAP designer will have to define each hierarchy as a separate dimension.

Aggregation Fact Tables

MDDBs build aggregations into the dimensional hierarchies. With star schemas, we may predefine certain aggregations and store the results in "aggregate fact tables." These fact tables may be thought of as "derivative" fact tables because they are derived from main, detail-level fact tables (Kimball, 1996). The presence of predefined aggregations improves response time by avoiding repetitive summarization activities. On the other hand, we cannot build aggregations for every possible combination. The key considerations in choosing the right aggregations are sparsity and usage patterns. Aggregation is more efficient if we aggregate along dense dimension elements (a large number of rows for that dimension element compared with other dimension elements) rather than among sparse dimension elements; in the former case preaggregation will save resources, whereas in the latter case aggregations may be done on the fly without seriously affecting performance (see Stanford Technology Group, 1995).

Communication With Other Tools

OLAP tools allow the user to export reports to word processing and graphical documents or to external spreadsheets for advanced or individual computing that cannot easily be handled by the OLAP tool.

INFORMATION ARCHITECTURE FOR E-COMMERCE

The OLAP technologies described in the previous sections each have their place in e-commerce applications. For example, if the analysis requirement were to study market penetration or customer profiling, a MOLAP design would be used. If the requirement were for analysis of specific customer behaviors in a customer relationship application (e.g., CRM), a ROLAP or a drill-through implementation would be preferable because both aggregate and corresponding detail data would be needed.

Generally, organizations will use OLAP (and other data warehousing and decision support technologies) to analyze clickstream data. This may be used to assess traffic over the Internet at an organization's Web site. The second utility is to combine this data and analytical results with other data in an organization, in particular marketing data. This will create a link between, say, marketing campaigns and promotions with Web traffic. Figure 4 illustrates the integration of a Web server with various

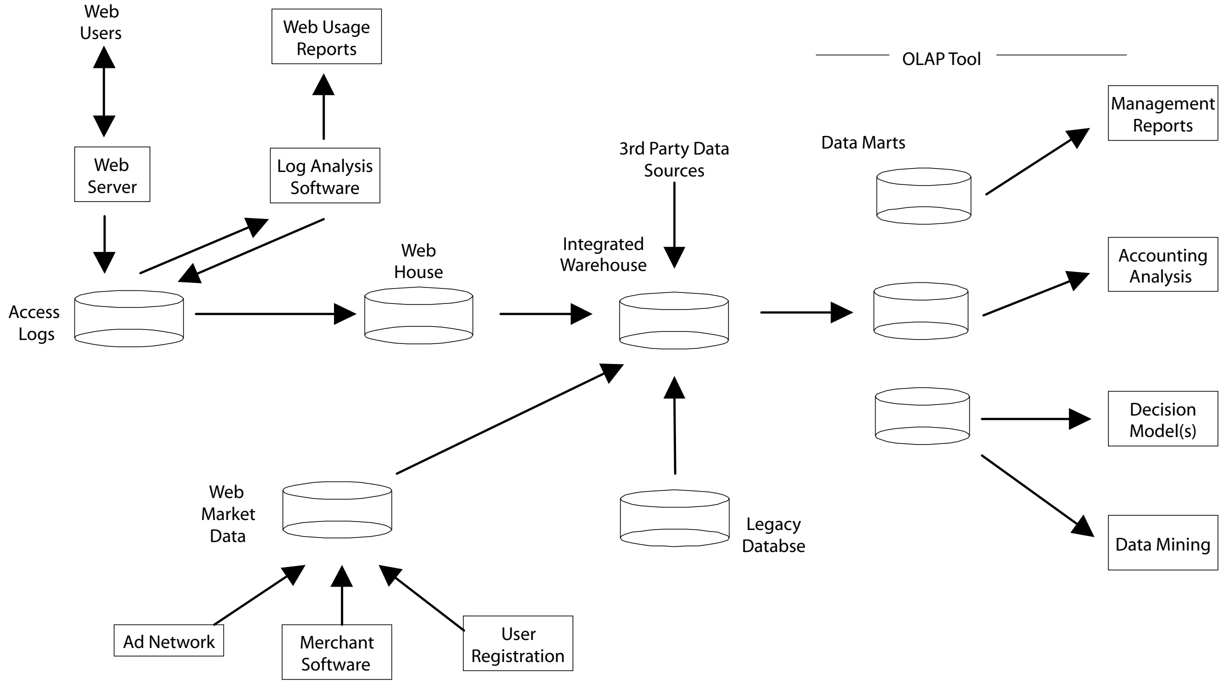


Figure 4: Data architectures for e-commerce.

other data sources and analytical tools including OLAP. Typical applications that result from these linkages are the following:

- Enhanced marketing and sales promotion activities (e.g., direct marketing campaigns),
- Support for customer relationship activities, such as providing possible sales leads, and
- Support for personalization and an enhanced customer experience at an organization's Web site.

Central to supporting a wide variety of e-commerce and OLAP applications is architecture and associated data-gathering activities. This environment cannot be single threaded for OLAP, but must encompass the full breadth of decision support activities and technologies.

In the remainder of this section, we explain each component of the architecture in Figure 4.

Access Log and Web Usage Statistics

Every access to a Web server is stored in real time in the server's "Transfer" or "Access" log (see Figure 4). This file contains a complete record of every event recognized by the Web server. For example, when a "hit" occurs, a record is generated containing the precise time, the IP address of the user, and the name of the requested page, file, or object. A "log analysis system" can then be used to provide detailed reports on, among other things, the number of "hits," "page views," and "visits" to the total site and to any page on the site (Stout, 1997).

The log files contain detailed data about all the events registered by the server. From this information, the Web site can learn valuable information about its users and about the features on the Web site that are of most interest and generate the most income.

Busy sites generate gigabytes of detailed data every day. To understand the data analysis steps, it is helpful to start with an understanding of the structure of weblog data (Stout, 1997). A typical Web server generates four "access log" files. These are the transfer, error, referrer, and agent logs, respectively. We now describe the data elements contained in a combined transfer, agent, and referrer log file (see Figure 5).

Host Field

The Host field contains the identifier of the remote host making the connection and requesting a document. The identifier can be the fully qualified domain name (the URL, or uniform resource locator, familiar to all Web users) or simply the Internet protocol (IP) address (a number that is the Web's identifier for a particular machine). The URL corresponding to an IP address can often be obtained by looking it up in the Internet domain name system (DNS), a hierarchical system of IP directories maintained on the Internet by various public and private authorities.

Ident Field

The Ident field was designed to record the actual user's name. For privacy and other reasons, however, it is seldom used; in its place, the server simply records a "-".

Combined Web Log File

Host-Field	211.56.155.44 (or www.acme.com)
Ident	-
Auth-User	(ID or password for protected area)
Time-Stamp	Date, Time
HTTP-Request	Method + page URL
Status-code	Error code (200 for success)
Transfer-Volume	File size (bytes)
Prior URL	Page user's browser was displaying
+ Search String	(if from search engine)
Path & File Name	Requested page (e.g.,/index.html)
Browser	User's browser & version no.
Server Cookie	User ID + data on interaction

Figure 5: Data fields from Web log files (combined).

Auth-User Field

The Auth-User field provides identification information for users accessing protected portions of the Web site. This can be used for billing users for their use of Web site features.

Time-Stamp Field

The Time-Stamp field provides the date and time (down to the second) of the transfer request.

HTTP-Request Field

The HTTP (hypertext transfer protocol) Request field contains a key word describing the type of request being sent to the server and the version number of the HTTP protocol used to make the request.

Status-Code Field

The Status-Code field records the success or failure of the transaction. In the latter case, the reason for the failure is also coded.

Transfer-Volume Field

The Transfer-Volume field records the amount of data (in bytes) transferred as the result of a successful "Get" operation.

Prior URL Field

The Prior URL field contains the URL of the current page or the URL of the page from which the user first came to the Web site (and perhaps also the actual search parameters used if the user comes from a search engine).

Path & File Name Field

The Path & File Name field contains the URL of the user's next requested page.

Server Cookie Field

The Server Cookie field contains a unique identifier for the user plus information relevant to the user's interaction with the Web site.

Web Server log files contain basic information about Web site usage but can be difficult to interpret. Note that the server will be recording data from multiple users in real time. This means that the records corresponding to a given user will normally be dispersed throughout the log.

Recognizing Visits or Sessions

From the previous discussion, it can be seen that the number of "hits" recorded by a Web site is not a reliable or meaningful statistic. The hits record events that can be accesses to HTML (hypertext markup language) pages, HTML page frames, gif image files, Java applets, and so on. It is much more useful to record the number of page views and visits by users to the Web site. For example, advertisers are often charged according to the number of page views (or "impressions") to which their ads are exposed. This information can be found from the Prior URL field. Visits and unique visitors can be even more meaningful statistics. A visit is recorded in the transfer log as a sequence of hits coming from the same IP address. The visit ends when the user either logs off from his or her browser or clicks on a link to another site. A problem arises if the user moves away from his machine for an extended period of time without clicking away from the site. To provide closure in these circumstances, it is customary to record the end of a visit after 30 minutes have elapsed with no response from the user. Thus, a new visit is recognized if the user resumes his or her session (say) 35 minutes after his or her last activity.

There are two other methods for determining visits—both of which are more accurate than the use of log files. First, if users are required to identify themselves on entry to the Web site their identity will be maintained throughout the visit in the auth-user field. Furthermore, user records can be generated to allow for comprehensive tracking of the user's behavior over multiple visits, which is the ideal situation. Many sites do not require users to sign in for fear of losing them. A second, somewhat controversial, method for tracking visits, is to identify users through the use of "cookies." A cookie is a small text file that is generated by the server on the first access by a user. The cookie contains, minimally, the identification of the server and a unique user id number that is dynamically assigned by the server to the user, or rather, to the user's computer. The cookie is sent to the user's machine, where it is accepted by the browser and stored on the local disk drive—often with other cookies in a file called "cookie.txt." When the user initiates the next request for a page from the server, the browser examines its file of cookies to see if a cookie exists for the server. If this is the case, the cookie information is included in the next request to the server. In this way, the server can identify the visitor and understand the current "state" of his or her interaction with the system.

Information about individual users is built up over time. If a unique identifier is assigned via a cookie when a user signs on to a site for the first time, the behavior of the user can be tracked over time. The user is still anonymous, however. If the user signs on to some service provided by the site, then the user can be identified by name. With this information, a match may be possible to the customer's records in the company's legacy files (see Figure 4).

Web Usage Summary Statistics

A wealth of information can be gained using appropriate software to analyze the data stored in Web Access logs. A number of commercial "log analysis" packages are available that perform this function. These packages are installed on the Web site's server and used to provide a wide range of useful reports. The following is a list of the typical statistics that are gathered and reported by these packages (adapted from Stout, 1997). Other technical statistics such as the browsers used to access the site are also recorded.

User information:	**Site activity information:**
User profile by regions	Top requested pages
Most active organizations	Top downloaded files
Organization breakdown	Top submitted forms and scripts
North American states and provinces	Top download types
Most active cities	Forms submitted by users
Sites accessed by proxy	Activity statistics
Top referring sites	Activity by day of the week
Top referring URLs	Activity by hour of day
	Most accessed directories

A number of companies specialize in tracking traffic on Web sites. To use these commercial services, it is only necessary to install a small program that periodically (e.g., hourly or daily) sends the tracking company a copy of the log file. The Web analysis reports, which are similar to those that can be produced in-house, can be read by accessing the service's Web site.

Analyses produced by the software described in this section are useful for ongoing measurements of the performance of the Web site and short-term decision making. The logic that such programs use to distinguish visits, visitors, and so on, is valuable. Log analysis software can therefore be used to produce files of clean, summarized data for subsequent, more sophisticated analyses. This data can be entered into the data warehouse. From there it can be used to estimate parameters for decision models and to provide inputs to data mining exercises in which the objective is to discover deep insights about users and usage patterns. Log analysis software can therefore constitute a vital link in the information chain leading to higher Web site profitability.

ORGANIZING DATA FOR OLAP AND E-COMMERCE

We now come to the point where data must be appropriately organized for OLAP in e-commerce applications. The key is to construct a BI architecture (see Figures 2 and 4) with Internet data. There are two general approaches depending on the e-commerce application. The first is to build an Internet-specific data mart for clickstream analysis; the second is to construct an integrated structure (data warehouse or mart) that contains data sourced from both the Internet and non-Internet sources (e.g., operational). Each of these areas is discussed in the following sections.

Data Webhouse

Up to this point we have been discussing clickstream data as a source of data. This data must be used to populate an integrated Data Warehouse as shown in Figure 4. There are two approaches. The first is simply to build the complete data warehouse illustrated in Figure 4. Another approach is to first build a data warehouse of Web data, which is known as a data webhouse (Kimball & Merz, 2000). This data webhouse may be used as a source for analysis (including OLAP) of Web traffic and as a staging area for integrated data warehouses or marts. These integrated data warehouses include selected data from the webhouse as well as specific organizational data, shown as the Legacy Database in Figure 4.

The key decision in designing the data webhouse is the selection of its grain. The grain is a representation of a single fact record. At the very least, the grain should be a completed visit or session. For more granularity, a grain of page event may be chosen. Each design results in a huge data webhouse that may be used to cluster customers and analyze their respective behaviors, such as time spent on the site, percentage of sessions that yield product sales, and so on. Other analyses, such as sales from demographic groups, are better implemented with aggregations of sessions. Each of these designs is illustrated in Table 1, adapted from Kimball and Merz (2000).

As shown in Table 1, the first two designs yield large, detail-oriented data structures best implemented as star schemas in a relational database employing ROLAP analytical tools. The session aggregation data structure is substantially smaller and can be implemented in a MDDB employing MOLAP technology. Because the aggregation session data structure is derived from the detailed session data structure, each may be linked through drill-through or HOLAP technologies.

Building the Integrated E-data Warehouse

As shown in Figure 4, advertising and sales information, beyond that recorded in the weblogs, comes from a number of different sources and contributes to the Web market data database. In particular, if the user registers at the site, valuable identification and demographic information is obtained. The data warehouse is populated by extracting Web transaction data from the Web server's transfer log (or the log analysis software) and market data from the Web market database. If the user can be identified (because he or she has registered or purchased goods or services), it may be possible to link the accounting, product and customer data from the firm's financial database (see Figure 4). Finally, again depending on precise identification of the user, the firm may purchase valuable demographic and financial information about the user from third-party sources such as mail-order houses. For example, these databases may contain information on the houses and vehicles owned by the user.

The integrated data warehouse combines data from the webhouse with data from existing data warehouses (or their legacy sources). Integration is possible only with common dimensions, sometimes called conformed dimensions (Kimball, 1996). For example, dimensions that are common to the clickstream datamarts shown in Table 1 and traditional marketing or sales data marts, for example, would most likely include calendar date, time-of-day, customer, product, and causal (inferred reason for user presence on the Web site) dimensions (Kimball & Merz, 2000). The technology used for an integrated data mart will follow that chosen for the clickstream data mart.

Linkage supplements traditional marketing and sales analyses with clickstream data. In effect, an organization's Web site becomes another channel and the integrated data mart is an extension of the marketing or sales data marts. With an integrated data mart an analyst will be able to determine sales percentage and profitability along each of an organization's channels, including the Web. Other

Table 1 Sample Clickstream Datamarts (adapted from Kimball & Mertz, 1999)

Dimension Name	Clickstream Session Fact	Clickstream Page Event Fact	Clickstream Session Aggregation Fact
Calendar date	X	X	
Time-of-day	X	X	
Customer	X	X	
Page	X	X	X
Session	X	X	X
Referrer	X	X	
Causal	X	X	
Event		X	
Product (page subject or event trigger)		X	
Calender aggregation unit such as month			X
Demographic aggregation			X

Table 2 Data Mart for Sales Channel Analysis

Dimensions				
Time	**Customer**	**Product**	**Channel**	**Facts**
Year Quarter Month Day of week Time of day Current month Prior month	Customer class Customer Size Demographics	Products lines Products	All channels Retail stores Catalog Web	Revenue Costs Profit Units sold

analyses would include the proportional contribution of the Web after a sales promotion or advertisement, most active locations (e.g., cities), and so on.

Developing Data Marts

An integrated data warehouse as described in the last section, is difficult and time-consuming to construct. It may also contain hundreds of gigabytes of data. To reduce complexity and increase computational speed, it is customary to develop a series of data marts using an OLAP or ETL tool (see Figure 4). Each data mart will reflect a particular view and will contain a subset of the total data in the warehouse. For example, one data mart might contain customer and clickstream data with the objective of using a data mining tool to discover patterns of Web site access that might distinguish different classes of customer. Given this information, the company might be able to recognize particular customers and their associated class and customize the Web site dynamically. Another data mart might contain product and customer data that could be introduced to an OLAP tool and used to discover trends by time, customer class, and geographic region.

As shown in Figure 4, the OLAP tool might also be used to extract data for routine management reporting and accounting purposes or for input into a decision support tool or spreadsheet that will support quantitative modeling and prediction.

USING OLAP IN E-COMMERCE APPLICATIONS

Recent experience has indicated that even within an organization only the serious analyst makes direct use of OLAP technology. This is primarily for two reasons.First, there is a long learning curve associated with OLAP; second, many of the attractive features of OLAP (e.g., drill-down, graphics) have been ported over to the more familiar and easier-to-use reporting technologies. Less expert users, however, make use of OLAP through preformatted and automatically generated reports. OLAP views may be set up with bookmarks and other MDDB features. The most common use of OLAP within an e-commerce environment involves Internet data, particularly Internet data that is integrated with marketing and sales data to form an integrated data mart as discussed in the previous section.

There are several ways in which OLAP concepts can be used in e-commerce. The following three scenarios are typical.

The Web Is Just Another Sales Channel

The multidimensional structure described in Table 2 will allow analysis of sales over multiple channels including the Web. For example, one could use this design to display monthly "high-end" (customer class) revenue by sales channel. To do this, in one commercial OLAP tool, one would simply click on Month, Customer class, All channels, and Revenue in each respective column.

Clickstream Analysis

The data cube described in Table 3 can be used to investigate many facets of the Web site usage. For example, one could easily plot a graph of number of page views versus customer zip code for a geographical analysis, number of page views versus time-of-day to identify peak usage, and so on.

Table 3 Data Mart for Click Stream Analysis

Dimensions			
Time	**Customer**	**Product**	**Facts**
Quarter Month Week Day of week Time of day Current month Prior month	Customer class Customer zip Customer	Products lines Products	Sessions Visits Page views Subscription $

Providing OLAP Capabilities to Customers

In the two previous scenarios, the OLAP cubes were designed for internal use. In the third scenario, the objective is to provide a service to customers by allowing them to analyze data pertaining to their own interaction with the company. This might be done by providing customers with many preformatted reports. For example, ARIBA provides this service to its customers, allowing them to view reports of summary and detailed transaction data. An alternative is to allow customers execution access to the OLAP cube itself. For example, at the request of its institutional investors, one large insurance firm in the Northeast plans to provide OLAP execution capabilities on its extranet. Analysts employed by the firm's customers will then be able to slice and dice the data in their "book," examining trends in billing versus claims or analyzing disability claims by class of employee, for example.

A firm, could, of course, employ several or all three of these modes of using OLAP.

CONCLUSION

OLAP is a technology that facilitates the advanced analysis of data along a set of predefined dimensions. It is often used to support both structured and unstructured decision making. OLAP comes in a variety of forms, depending on the back-end data warehouse or mart. Management of the modern BI environment does not separate OLAP from other technologies or their back-end data stores. As shown in Figure 2, the BI environment is distinguished by an architecture that serves to increase the information carrying capacity of data (i.e., increasing data richness) in a formalized and documented way. Each type of end-user tool (data warehouse query and reporting, OLAP and its varieties, data mining) is tied to the appropriate level of data richness on which it operates. From such an "architected" decision support environment, it is relatively easy to mix and match data and tools; the development of a data webhouse to analyze Web data, integrating it with other data marts, making the separate or integrated data available to internal or external clients, and so on follow from the BI architecture.

The 21st-century organization is characterized by knowledge-based work and the integration of the technical and social dimensions of technology (Kochan et al., 2002). OLAP is one key technology for knowledge-based work, but as we have discussed throughout this chapter, OLAP can no longer be separated from other related technologies. Thus, the second dimension of the BI environment is the intermingling of delivery technologies—the requirement for an environment that mixes query paradigms and makes each available to a user (illustrated in Figure 2) as well as complex analytical applications. For example, we have seen organizations apply data mining to large data sets and deploy the results through reports or OLAP. These are sometimes called "opportunity reports" and have been used in several industries: opportunities for client sales in the financial industry and for drug discovery in the pharmaceutical industry. For example, a large brokerage firm runs a series of complex analyses on its client base and deploys "opportunities" that are personalized: Each broker may access a report or OLAP presentation of sales opportunities, depending on the mining criteria (deployed as meta-data) for his or her specific clients. In this case, data mining, reporting, OLAP, personalization, and meta-data are seamlessly integrated and become opportunities for cross- or up-selling for the stockbrokers of this organization. We believe that this will represent the next generation of BI.

GLOSSARY

Aggregate fact A predefined aggregate measure along a dimension in a multidimensional database or in an aggregate fact table in a star schema.

Business intelligence (BI) The full set of activities required to analyze and capture, cleanse and store, transform and derive, and distribute data through analytical applications and specialized tools to support analysis and decision making.

Conformed dimension Common or shared dimensions (logical or physical) among two or more data warehouses. Conformed dimensions permit the incremental construction of an integrated data warehouse. Integration is achieved through conforming dimensions among the separate data warehouses and their different sources—a data webhouse and a marketing data warehouse, for instance.

Data cleansing A data warehousing function that is executed before data delivery to users. Data cleansing includes activities such as removing duplicate records, correcting invalid attribute values, and so on. Data cleansing is typically applied to operational data stores and occurs at the instance level (dealing with actual values) and not at the type level (technical meta-data such as data type or length).

Data warehouse A database system that supports analysis and decision making. The data warehouse stores data from various operational databases and is the central construct in the BI environment.

Data webhouse A term used to refer to a data warehouse consisting of Web data. The data webhouse may be used as a data source for analysis of Web activity or as one data warehouse to be integrated with others (e.g., marketing) through conforming dimensions (Kimball & Merz, 2000).

Executive information system (EIS) Strategic-level information systems that serve to support unstructured decision making through advanced graphics and communications (Laudon & Laudon, 1998). Also known as an executive support system (ESS).

Extract, transformation, and load (ETL) A set of data warehouse functions that embody the capture, transformation, and loading of data elements between data layers or stores within the business intelligence environment (e.g., between source data stores and an atomic data warehouse, between an atomic data warehouse and a data mart, etc.).

Hybrid online analytical processing (HOLAP) HOLAP is an optimized and integrated form of both MOLAP and ROLAP, designed to exploit the benefits of both OLAP technologies.

Meta-data Usually defined as data about data, meta-data is more properly defined as information of varying richness describing data. There are two types of meta-data, technical and business. The former is the traditional view and reflects the historical perspective of meta-data as something contained in a data dictionary for use by information technology professionals, particularly data or database administrators. Business meta-data is for use by an end user and describes the business context of a given data element or elements.

Multidimensional data An organization of data used to support analysis and decision making, particularly by end users. The key factors in modeling multidimensional data are simplicity and understandability. In practice, the key design criteria are aggregation and speed of retrieval.

Multidimensional database (MDDB) A proprietary database used to store multidimensional data. The MDDB is organized as an *n*-sided spreadsheet and is known as a "cube" structure. Naturally, the cube is a metaphor because a given MDDB may have many sides (through in practice more than a dozen is rare). The sides of an MDDB structure correspond to the dimensions of a star schema design while the cells correspond to its facts or measures.

Management information systems (MIS) Management-level information systems that serve the functions of planning, controlling, and decision making by providing routine summary and exception reports (Laudon & Laudon, 1998).

Multidimensional online analytical processing (MOLAP) MOLAP is OLAP applied to multidimensional data stored in an MDDB.

Online transaction processing (OLTP) A transaction processing system normally associated with an organization's operational processes. OLTP systems typically operate on a single, logical record of data and represent the bulk of an organization's systems.

Online analytical processing (OLAP) An interactive analytical system for use by analysts and other key decision makers in an organization. OLAP systems are distinguished by their projection of a multidimensional view of many logical records of data.

Operational data store (ODS) An ODS has several definitions, usually falling within one of two categories, representing operational and data warehouse perspectives, respectively. In the operational category, an ODS is the underlying data store that supports either a single OLTP or an integrated OLTP system. In the data warehouse category, an ODS refers to the data store that is a consequence of extraction from source, operational data.

Relational online analytical processing (ROLAP) ROLAP is OLAP applied to multidimensional data stored in a star schema in a relational database.

Star schema A multidimensional relational database design. The data is organized as facts and corresponding dimensions in a kind of hub-and-spoke configuration. Each dimension is denormalized and captures all the attributes associated with a given dimension.

Visit or session A sequence of "hits" from the same Internet protocol address. The visit ends when either the user logs off, clicks on a link to another site, or is timed out after a predetermined length of inactivity, often 30 minutes.

CROSS REFERENCES

See *Data Mining in E-Commerce; Data Warehouses and Data Marts; Databases on the Web; Knowledge Management.*

REFERENCES

Adamson, C., & Venerable, M. (1998). *Data warehouse design solutions.* New York: Wiley.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Cambridge, MA: MIT Press.

Inmon, W. H. (1996). *Building the data warehouse* (2nd ed.). New York: Wiley.

Kimball, R. (1996). *The data warehouse toolkit.* New York: Wiley.

Kimball, R., & Merz, R. (2000). *The data webhouse toolkit.* New York: Wiley.

Kochan, T., Orlikowski, W., & Cutcher-Gershenfeld, J. (2002). Beyond McGregor's theory Y: Human capital and knowledge-based work in the 21st century organization. Retrieved December 2002 from http://mitsloan.mit.edu/50th/papers.php

Laudon, K. C., & Laudon, J. P. (1998). *Management information systems: New approaches to organization & technology* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Morabito, J., Sack, I., & Bhate, A. (1999). *Organization modeling: Innovative architectures for the 21st century.* Upper Saddle River, NJ: Prentice-Hall.

Stanford Technology Group (1995). *Designing the data warehouse on relational databases.* White Paper. Menlo Park, CA: Stanford Technology Group.

Stout, R. (1997). *Web site stats: Tracking hits and analyzing traffic.* Berkeley, CA:.Osbourne McGraw-Hill.

Online Auctions

Gary C. Anders, *Arizona State University West*

ONLINE AUCTIONS AND PRODUCTS

As a consequence of the Internet, we are experiencing a technologically based economic revolution. One aspect of e-commerce is the growing popularity of online auctions. Through online auctions the Internet has become a global marketplace for hundreds of millions of people. Online auctions overcome many of the limitations associated with conventional markets. In contrast to traditional face-to-face transactions, online auctions provide an electronic format for buying and selling. Internet auction sites make it less expensive for buyers and sellers to meet, exchange information, and complete transactions. Online auctions are a portal where buyers and sellers separated by time and distance can make transactions for a great variety of items. Online auctions operate continuously—24 hours a day, seven days a week, 52 weeks a year. Once the terms of the auction are established, buyers and sellers from all over the world can participate. Internet auctions can range from common household items in business to consumer (B2C) auctions to million-dollar equipment in business to business (B2B) auctions. (Among the many advantages of online auctions is the ability to set sale quantities on the basis of existing inventory, and even offer special closeouts for damaged or defective items.) There are an increasing number of B2C online auctions selling a wide variety of goods and services. In addition, there are individual seller online auctions that make up a segment of online B2C auctions.

Estimates of e-commerce are quickly outdated because of the high rate of growth. Still, it is possible to see the emerging trends even though the actual forecasts are understated. For example, according to the Forrester Research, global Internet commerce including both B2B and B2C will be approximately $6.8 trillion in 2004 (up from approximately $657 billion in 2000; Forrester, n.d.). Of this, the B2B component of online auctions is expected to account for 29%. Although these types of auctions are briefly discussed, the primary focus here is on B2C auctions.

Table 1 presents a list of the most successful Internet auction sites. eBay, which owns and operates eBay.com, the largest online auction, has a huge customer base of approximately 22 million registered users from 200 countries. Established in September 1995, eBay has an estimated net worth of $1.5 billion. In 2000, eBay users completed transactions with an aggregate value of more than $5.4 billion. eBay features a powerful search engine that allows bidders to search more than 8,000 categories for a specific item. The main categories on eBay.com include art and antiques; books; business office and industrial; clothing and accessories; coins; collectibles; computers; consumer electronics; dolls and bears; home and garden; jewelry, gems, and watches; movies and television; music; photo; pottery and glass; real estate; toys and hobbies; and everything else.

eBay has recently added new specialty auctions and services to its product line. eBay Motors in association with AutoTrader.com offers auctions for automobiles, parts, and accessories. eBay Premier offers premium art and collectibles in conjunction with traditional auction houses Butterfields and Kruse International owned by eBay. In addition, eBay owns Half.com, a fixed price venue that permits sellers to list books, recorded music, movies, and video games for sale. eBay and its affiliates also operate online trading platforms in numerous foreign countries.

The online bookseller, Amazon, recently branched out to offer Internet auctions. Amazon features an international orientation with conversions of various currencies into U.S. dollar prices and its own highly developed search engine. Egghead, which merged with Onsale in 1999, subsequently filed for bankruptcy and was acquired by Fry's Electronics, which is owned by Kroger. The Egghead site, powered by Amazon.com, primarily specializes in computer and office equipment. Priceline allows bidders to set their own prices on travel-related services, but is quickly expanding into new areas including automobiles and home financing. Yahoo recently launched a shopping network site that combines new and used retail sales with auctions and classified advertisements. UBid.com has grown from an online vendor of refurbished items to a major auction site through strategic alliances with Microsoft Network (MSN), Excite, and other companies.

Online B2B auctions specialize in industrial markets for components and surplus equipment. The first online auction for industrial goods was Trade Out.com founded in October 1998. During its first 5 months it sold $22 million to 7,500 customers. FreeMarkets creates real-time virtual markets for industrial products and commodities for pre-qualified buyer and sellers (Vigorosa, 1999). Other industrial auctions include Equipp.com, DoveBid, iMark.com, bLiquid.com, and AsseTrade, which all handle industrial products and manufacturing equipment (Teschler, 2000).

Table 1 Major Business-to-Consumer (B2C) and Business-to-Business (B2B) Online Auctions

eBay.com	Founded in 1995, eBay is the largest online auction. It recently undertook a successful joint venture with AOL.
Amazon.com	Amazon began offering auctions on March 30, 1999, to its established base of more than 8 million customers.
Egghead.com	Egghead merged with Onsale in 1999 but was absorbed by Fry's Electronics in 2001. Now powered by Amazon, this site specializes in computers, electronics, and office products.
Priceline.com	This site offers travel, hotel rooms, rental cars, plus additional categories for long-distance telephone, automobiles, and home financing.
uBid.com	Started in 1997 to sell refurbished computer equipment, uBid has grown into a successful auction site offering name-brand consumer electronics.
Yahoo! Shopping Network	Started in 2001, this portal combines retail sales with auctions and classified adds.
TradeOut.com (B2B)	This online industrial auction was established in 1998 and sold more than $22 million in its first 5 months.
FreeMarkets.com (B2B)	Started in Pittsburgh in 1995, FreeMarkets.com established industrial auctions for components such as circuit boards. Gross revenues from 2000 were in excess of $8.3 billion.

Large manufacturing companies have established reverse online auctions to collect competitive bids from vendors. Visteon Automotive Systems (a division of Ford Motor Company) used a 90-minute online auction to collect bids on $150 million dollars in supplier contracts. As a result, Visteon was able to cut several weeks off the usual procurement process and realize considerable savings (Dalton, 2000). According to Bollinger and Madey (2000), online auctions reduce delays, inventories, cycle times, and operating costs while improving quality and profit. They also point out that online B2B auctions make it possible for global suppliers to compete on the same level as local suppliers. Other large companies, such as General Motors and DaimlerChrysler, are developing online auctions as profit centers (Teschler, 2000). There is a fairly strong possibility, for example, that some companies will implement internal online auctions that encourage employees to bid on jobs as well as employee teams to bid on projects.

Even such typically staid industries as banking and mortgage lending have been swept up in the fast-moving currents of online auctions. Consumers can now buy a house and arrange for financing and insurance through online auctions. Some banks such as Pittsburgh-based PNC Bank Corporation have started auctioning Certificates of Deposit online (Dalton, 2000). A newly formed investment bank, W.R. Hambrecht & Co., uses the Internet to collect online bids for public offerings (Dalton, 2000).

Placing an item for sale on the Internet is often cheaper that placing a classified advertisement, and the seller is able to reach a broader audience. According to Lucking-Reiley (2000b) online auction commissions are about one fifth of what traditional auctions charge. Moreover, a seller does not physically have to transport goods to the market. Quantities offered can range from one item to an entire inventory, and sellers are able to offer goods at differing initial prices. Overall, the ease of replicating the same format for multiple items makes selling on online auctions highly attractive. This dynamic pricing feature provides a mechanism for first-degree price discrimination for the complete extraction of consumer surplus based on what each buyer is willing to pay for an item. This is possible because of the large number of buyers and the difficulty of collusion. Online auctions reduce the costs of gaining market exposure while providing businesses with a mechanism for testing the sensitivity of demand to various prices or jointly sold items.

Here is a brief explanation of how an online auction works. To participate, both the seller and the buyer have to register with the auction site and agree to the auction rules. The seller submits an item for sale that often includes a photograph and a short description along with other relevant information including the accepted forms of payment. In this process the seller can choose the type of auction, length of auction, and include either a minimum bid, or a secret reserve price that must be met before the good can be sold. The buyer follows the auction and may make incremental bids, or use a bid proxy to automatically increase his or her bid. Bids are increased according to the current price with high prices requiring a larger incremental increase. Toward the end of the auction the bidders compete up to the closing moments when the highest bid is finalized. Immediately after the end of the auction, both the buyer and the seller are notified of the result and are provided with each other's e-mail address so that they can communicate with each other regarding arrangements for payment and shipping. Auction fees are charged against the seller's account or credit card. The participants of the auction, their bids, feedback about each other, and the winning bid become a part of the auction history and are available for review.

TYPES OF AUCTIONS

There are nine principal types of auctions, and most of these are replicated online:

1. In *Standard English ascending price auctions*, an item is listed with an opening bid that establishes the minimum price. Bids increase until the auction ends or until there is no further activity. This is the most common type of online auction.
2. *Reserve auctions* allow a seller to list an item with a low starting price, but the item will not be sold until the bidder meets the seller's reserve price. Reserve auctions provide the seller with protection from low bidding.
3. *Dutch auctions* In online Dutch auctions, the buyer specifies the quantity that is being bid on along with the offer price. Winners in online Dutch auctions pay the lowest price of the winner who established the quantity offer that clears the market. This term is also used in online auctions where multiple identical items are offered.
4. *Sealed bid auctions* attempt to reduce the "winner's curse" syndrome by not disclosing bids until after auction is over. In the "second-price" sealed bid auction, the highest bidder wins, but pays the second highest price.
5. *Descending price auctions* are a more traditional version of the Dutch auction. These auctions start at a high price and incrementally reduce the price until the good is sold.
6. *Double auctions* involve several buyers and sellers who publicly announce their demand and supply prices. Westland and Clark (1999) argue that double auctions are more efficient because their prices reflect all relevant information. An example of this might be trades on the floor of an exchange such as the Chicago Mercantile Authority.
7. *Commodity auctions* are used for oil, freight transportation, and electricity. These types of auctions have also been used to sell bandwidth to telecommunications companies and drilling rights on publicly owned lands.
8. For *Industrial auctions* of surplus manufacturing equipment, a date and time for the start of the auction and its duration are specified, and in many cases, bidders need to be to pre-qualified to participate.
9. *Reverse auctions* allow companies to collect online bids from prospective contractors. For example, a large manufacturing company will hold an online auction to collect bids from pre-certified suppliers to fill orders for parts.

Generally, online auctions run for a specified period of time, usually between 3 and 7 days (for more discussion, see Lucking-Reiley, 2000a). Some auctions such as Amazon have a "going, going, gone" feature that allows active bidding to extend the auction. In the most common English or ascending price online auctions, bidders can enter their reserve and a proxy will automatically increase their bids up to this amount. Online proxy bidding also tends to reduce the possibility of cheating that, according to Lucking-Reiley (2000b), is fairly common in second-price auctions. During the course of the auction, bidders are notified by e-mail of the status of their bids. Sellers and successful bidders are automatically notified when a transaction is completed. In this respect, eBay auctions come closer to Vickrey's (1961) "second-price sealed bid auction" because these auctions allow the prospective buyer to consider their reservation price rather than simply reacting to the bidding activity. There are auctions that legally provide for interstate sales of firearms and knives such as gunbroker.com. Transactions that involve modern weapons are subject to federal regulations and are shipped to and from licensed dealers, who do the required background checks and paperwork. There are, however, strenuous efforts to prohibit the listing of certain items on online auctions. The obvious examples are illegal items such as non-regulated firearms sales, pornography, and items that violate the copyrights of manufactures. Because of the liability issues online auctions sites are increasing their vigilance of both products and vendors. For example, eBay established the SafeHarbor program, which has a number of functions including guidelines for trading, dispute resolution, and investigations of possible misuse of eBay such as the sale of prohibited goods or shill bidding.

HOW AUCTIONS MAKE MONEY

Online auction.s offer a wide assortment of goods while bypassing traditional intermediaries. Online auctions are a robust and efficient market that brings a large number of buyers and sellers together. Although sites exist where it is possible to post an item at no cost, most online auctions charge fees for this service. TradeOut.com charges a flat fee of $10 per listing, plus a 5% commission on sales. In contrast eBay has a fairly complicated fee structure that varies with the type of auction, special options chosen, and the final selling price. The nonrefundable fee for posting an item is called the insertion fee (see Table 2). Insertion fees tend to vary by the type of auction including the following:

- Regular listings insertion fees, based on the opening value or minimum bid
- Reserve price auction insertion fees, based upon the reserve price of the item listed for sale. eBay charges an additional fee for reserve price auctions ranging from .50 to $2.00, but this fee is refunded if the item sells.
- Insertion fees for Dutch auctions, based on the opening value or minimum bid of the item listed for sale

Table 2 eBay Insertion Fee Schedule

MINIMUM BID, OPENING VALUE, OR RESERVE PRICE	INSERTION FEE
$0.01–$9.99	$0.30
$10.00–$24.99	$0.55
$25.00–$49.99	$1.10
$50.00–$199	$2.20
$200 and up	$3.30

Source: eBay, your personal trading community website, http:/w.w.w.eBay.com. These fees were in effect in April 2002. Other fees are applied to more expense items such as automobiles, motorcycles, and real estate and are sold in different eBay auctions.

Table 3 eBay Optional Fees

SELLER FEATURE	DESCRIPTION	INSERTION FEE
Home Page Feature	The item will be listed in a Special Featured section on eBay's Web site	$99.95
Featured Plus!	Special featured items are listed first each category	$19.95
Highlight	A colored band is placed around the item to draw attention	$5.00
Bold	The title of the item is bolded	$2.00
Gallery	Using a JPEG format, a small picture of the item is included with the description	$0.25
Gallery Featured	This is a larger sized picture included in the listing	$19.95
List in 2 Categories	Allows seller to list an item in more than one auction category	Double the insertion, plus optional feature fees
10-Day Auction Duration	This is eBay's longest listing	$0.10
Buy It Now	This feature allows a buyer to purchase an item instantly	$0.05

Source: eBay, your personal trading community website, http:/w.w.w.eBay.com. These fees were in effect in April 2002. Other fees are applied to more expense items such as automobiles, motorcycles, and real estate and are sold in different eBay auctions.

multiplied by the quantity of items offered. Sellers use Dutch auctions to sell a quantity of an identical item. Therefore, they are willing to pay a higher insertion fee because of the potential for multiple sales.

Additional nonrefundable fees are charged for different listing options. These options include featured auctions that showcase items, giving them more visibility; gallery listings that allow sellers to post small pictures of the item; and featured gallery with a larger size picture. Table 3 summarizes these auction fees.

Along with fees for listing an item, online auctions charge a final value fee if the auction is successfully concluded (i.e., the goods are sold and the seller collects money). These fees vary depending on the selling price and the type of auction. For example, if the item sold for more than $25, the final value fee on eBay is 5.25% of the first $25 plus an additional 2.75% for sales of between $25.01 and $1,000, and another 1.5% of anything over $1,000. For Dutch auctions, the final value fee is based on the lowest successful bid multiplied times the number of items sold. Table 4 presents the eBay sales fees. eBay does not charge a fee for listing an item on Half.com, but sellers are charged a commission equal to 15% of the final sale price.

There are other fixed fees for higher value items: $40.00 for automobiles and $25.00 for motorcycles. A different fee structure applies to real estate auctions. These fees apply only when the auction is successful. If there are no bids or no bids equal to the minimum reserve, eBay does not charge a final value fee. Except for real estate, final value fees (FVF) are charged at the end of the auction based on the item's final sale price, unless in a reserve auction the minimum price has not been met, or no bids were received on the item. Otherwise sellers are charged these fees regardless of whether the sale is consummated. A seller may report a nonpaying bidder and receive a credit for the FVF amount, but not for the insertion fee.

In addition to auction fees eBay and other companies make considerable revenue from selling advertisements, providing e-commerce services, and hosting online stores. For example, eBay's Application Program Interface (API) and Developers' Program allows other companies to use eBay content for their own businesses. In 2001, it earned approximately $700 million in net revenues from its auctions and subsidiaries. eBay attributes its success to a

Table 4 eBay Final Value Fees

CLOSING VALUE	FINAL VALUE FEE
$0–$25	5.25% of the closing value
$2–$1,000	5.25% of the initial $25 ($1.31), plus 2.75% of the remaining closing value balance
Over $1,000	5.25% of the initial $25 ($1.31), plus 2.75% of the initial $25-$1000 ($26.81), plus 1.50% of the remaining closing value balance

Source: eBay, your personal trading community website, http:/w.w.w.eBay.com. These fees were in effect in April 2002. Other fees are applied to more expense items such as automobiles, motorcycles, and real estate and are sold in different eBay auctions.

very large subscriber base, a large selection of goods, and name-brand recognition. Because of this, eBay has developed a critical mass of buyers and sellers that result in a high rate of successful auctions. Through major investments in Billpoint and PayPal, eBay earns profits from its electronic payment services that allow buyers to pay for transactions over the Internet.

The economics of online auctions are based on relatively high fixed costs, but decreasing average costs. Attracting more buyers and sellers to an auction increases the number of successful transactions. The greater the number of completed auctions the more the fixed costs are spread over a larger quantity of goods sold. This is known as economies of scale, and it has important implications for the online auctions industry. Economies of scope involve using the Internet portal to offer an increasing array of different types of goods and services. Because of the exponential growth, major online auctions have experienced lucrative returns on their investment. The potential for significant economic profits has given rise to increasing competition.

COMPETITION IN ONLINE AUCTIONS

Global online auctions now include such sites as eBay, Amazon.com, Yahoo!, uBid, Buy.com, AOL.com, and MSN. As these companies have broadened their product offering, they have incited competition from distributors, liquidators, catalog and mail order companies, and traditional retailers. (See eBay filings for the U.S. Securities and Exchange Commission at http://www.sec.gov/cgi-bin/srch-dgar?text = eBay&first = 1993&last = 2002&mode = Simple.) There are also a growing number of more specialized online auctions, such as Ariba, Bid4Assets, bLiquid.com, BizBuyer.com, CloseOutNow.com, Cnet.com, DoveBid, First Auction, Surplus Auction, Oracle, bLiquid.com, Overstock.com, Sabre, Ventro, and VerticalNet. Some of the new entrants into the online auctions industry such as MSN have substantial capital to support the development of a competitive platform, others are seeking a unique market niche. One such example is local auctions that specialize in items for a regional market. Some examples are the Web sites of local newspapers and Internet providers. There has been an attempt to establish a system of regional online auctions by CityBid.net. To meet this challenge, eBay launched eight regional sites for items of local interest or that are too bulky or expensive to ship.

According to traditional microeconomic theory, industry profits largely accrue to firms with the lowest costs of production. In the competition between established global auctions with a large customer base and extensive product offering, new entrants must be able to overcome the advantages of economies of scale. The existence of items particularly suited to local auctions does not insure that local auctions can compete against global auctions. The success of an auction is directly related to the exposure it receives from a large number of subscribers. Global auctions that have achieved a "critical mass" can provide many more choices to both buyers and sellers and have an established reputation, which means that they can more easily attract customers to their Web sites.

If local auctions are to be successful, they must be able to provide sellers with compensating advantages. Global auctions offer sellers maximum exposure to the largest number of prospective buyers, but for certain sellers of nontransportable goods and services, a local auction may be more useful. Online local auctions may provide the seller with an opportunity to generate additional revenue. Let's take a look at a hypothetical example. Suppose a recently graduated dentist starts a practice. While she or he may get referrals from family and friends, it is unlikely that this dentist will have enough clients in the first year to pay the extensive fixed costs of renting space and paying for the equipment. A local area auction could provide an opportunity to boost the number of clients, but it would not make sense to auction this service in the global auction. Other prospective local auction items include professional services, theater and movie tickets, restaurant meals, tours, and sporting events. All of these have a common nontransportable feature. Even though they can be sold nonlocally to enjoy these products and services, a buyer must be physically present. Local auctions must provide sellers with a way to attract bidders who are more likely to buy their nontransportable items. For instance, if I am interested in selling tickets to a sporting event or concert in Phoenix, global auctions are only relevant if they include some bidders who will be in Phoenix at that time. Global auctions offer sellers the ability to offer their wares to millions of potential bidders. The dominant firms like eBay believe that the size of their auction site provides sellers with a ready market. Because of this they compete without reducing their fees. But as global auctions increase their size and scope, it means that many auctions may fail due to a lack of bidders. Therefore, the success of local auctions largely depends on their ability to attract a loyal client base of higher probability buyers. As a result, local auctions will have to target a segment of the online market and concentrate their resources in building a comparable reputation, and so on. If they are successful, it may be possible to take away some of the market share from global auctions. In this respect, local online auctions will create new markets for items of local interest such as professional services. At the same time, they may lure customers away from their competitors despite switching costs that include seller reputation established through customer feedback. For example, eBay has fostered customer loyalty by providing buyers and sellers chat rooms, bulletin boards, and threaded discussion boards. But this service is easily replicated and enhanced, as in the case of Amazon.Com, which provides product evaluations and other information to buyers. Table 5 summarizes the differences between global and local auctions.

Given the ease of adding an online auction capability to an existing Internet-based company it is likely that the number of new entrants will continue to increase over time. For firms willing to dedicate the additional resources to add an e-business component to their existing operations, there could be considerable benefits. For example, in addition to selling their products, a business receives exposure that constitutes a form of extremely low-cost advertising. Building a database of repeat customers also allows firms to market their products directly through e-mail.

Table 5 Comparison of Online Auctions

GLOBAL AUCTIONS	LOCAL AUCTIONS
Established URL brand	Must rely on search engines to find site
Large customer base	Smaller, but targeted customer base
Large number of auctions	Fewer auctions, but focused offerings
Economies of scale and scope	Niche market player
Highly capitalized	May have an existing subscriber base
Possible diseconomies of scale	Potential for higher quality service
Increased buyer transactions costs	Reduced search time for local products

eBay's competitive strategy is based on growth through geographic expansion, increasing product categories, and now fixed-price sales. To counter upstarts, eBay has established local trading for 60 geographic regions in the United States and has launched a similar regional service offering in Germany, the United Kingdom, and other countries. This continued growth raises questions about the technological limits in online auctions. It remains to be seen whether the dominant firms may reach the point where having such a large number of auctions increases the possibility of system failures.

In traditional microeconomics, we use the term *diseconomies* in the physical production of goods. This means that at some point beyond the minimum efficient scale (MES), the costs of production will increase. In information-based businesses, a diseconomy might occur when an increasing number of items and auctions undermine the firm's ability to manage the increasing amount of data. In other words, eBay may become so large that search algorithms designed to find items will become overloaded, resulting in longer response times. These longer wait times increase transaction costs to bidders and may negatively impact sellers. In short, the possibility that such large-scale data management may eventually encounter technological limitations creates another niche for local auctions.

In addition to local auctions, global auctions must overcome the competition arising from intermediaries such as Dealtime.com that provide efficient searches to several online auctions. To develop a client base, auctions directly market to sellers, but new buyers must rely on search engines to take them to an auction Web site. As a result of the function played by auction consolidators and Web browsers, it seems likely that they may seek to capture a share of online rent. Eventually, increased competition may result in reduced operating margins, loss of market share, and diminished profits. Certainly we can anticipate a dynamic and multifaceted rivalry between all of these players and increasing sophistication of the online market place.

ACADEMIC RESEARCH ON AUCTIONS

The purpose of this section is to briefly review some of the important academic research on auctions. For a more extensive review, see Kagel and Roth (1995). Vickrey (1961) was one of the first economists to study auctions seriously. Much of the scholarly literature on auctions has been devoted to estimating the revenue outcomes from various types of auctions (Ward & Clark, 2000). A stream of literature began to emerge which found that first price Dutch auctions and second-price sealed bid auctions should result in equivalent revenues to the seller given differences in private valuation (Kagel & Roth, 1995). Milgrom (1989) pioneered research that compared various auction formats and bidding behavior. Bulow and Roberts (1989) further developed the theorem of revenue equivalence. McAfee and McMillian (1996) examined how open bidding encourages higher bids. Emmanuel, Perrigne, and Vuong (2000) demonstrated a procedure that estimates the underlying distribution of bidders' private values from observed bids.

The standard auction model postulates a single good sold at a competitively determined price. Biglaiser and Mezzetti (2000) extended this model to incentive auctions in which principals compete for the exclusive services of an agent. This is a variant of the first-price sealed bid auction in which bids are incentive contracts and private information exists on both sides. Examples of this type of auction include competition for sports stars or talented executives. Brannman and Froeb (2000) studied the effects of mergers and small business preference policies in Forest Service timber auctions.

Dasgupta and Maskin (2000) show that the Vickrey (one good) and the Groves-Clarke (multiple goods) auctions can be generalized to attain efficiency if the bidders' information is one-dimensional. Efficiency is defined as "auctions that put goods into the hands of buyers who value them the most" (Dasgupta & Maskin, 2000). But when bidders' information is multidimensional, in other words buyer's information cannot be reduced to one dimension such as private valuation, no auction is generally efficient. To appreciate the importance of this point we must understand that in the Vickrey auction, bidders reveal their true private valuation. In the multidimensional and more representative case, a buyer's valuation often depends on the signals they receive from other bidders. In other words, auctions with complex interplay between multiple buyers forming affiliated private valuations in response to the revealed valuations of competing bidders is "constrained efficient" subject to incentive constraints.

Until recently, little attention has been paid to understanding the differences between online bidding results and traditional auctions (Roth & Ockenfels, 2002). Bapna (2000) provided an excellent review of the literature on auction theory and online auctions. Empirical research based on actual participation in online auctions provides us with an understanding of the issues unique to Internet platforms. For example, it is generally understood that competition between buyers within the context of online auctions reduces the buyer's consumer surplus. Given the higher reservation prices of bidders with a higher income or stronger preference for a particular product, their participation can drive up prices and thereby reduce expected gains.

Bajari and Hortacsu (2000) examined the empirical relationship between the number of bidders and the corresponding decrease consumer surplus. They studied price determination and online bidder behavior using a data set from eBay coin auctions in which the average seller had previously conducted 203 auctions, and the buyers had previously won an average of 41 eBay auctions. They found that 85% percent of these auctions resulted in a sale with an average of three bidders per auction and that the average sale was 83% of book value with a mean bid spread of 32% of book value for coins ranging from $3 to $3,700. They found that in a sample of 516 auctions over a 5-day period, 20 auctions resulted in a winning bid of more than 50% of the item's book value resulting in the "winner's curse." In other words, the highest bid for the item is substantially greater than the book value or replacement cost.

Mathematically this can be modeled as

$$X_i = V_i + E_i,$$

where X_i, the bid, is equal to expected value V_i, plus a noise effect E_i common to all bidders; however, in winner's curse" situations,

$$X_i = U_i + V_i + E_i,$$

where U_i is the subjective utility associated with winning the auction that may cause overbidding and the winner's curse.

Bajari and Hortacsu also found that coins on eBay with a higher book value tended to be sold with a secret reserve price and low minimum bid, and that a low minimum bid, high book value, and low negative feedback increase the number of bidders. Thus, the number of bids submitted in the auction is a decreasing function in the ratio of minimum bid to book value.

Another aspect of some online auction is sniping, a term used to describe the placing of last-second bids. Usually, the sniper does not bid earlier in the auction and waits until the last possible instant before bidding. This can be an effective bidding technique unless the auction rules extend the time. Bajari and Hortacsu's analysis confirms that most bidding activity occurs at the end of the auction (also see Roth & Ockenfels, 2002). Furthermore, an increase in the number of bidders reduces buyer profit by about 3.2% per bidder, which increases the likelihood of overbidding.

Katkar and Lucking-Reiley (2001) found that the use of secret reserve prices in eBay auctions tends to lower the probability of a successful outcome and is less desirable than an equivalent minimum starting bid. Based on actual auctions for collectable cards public reserve auctions with a higher minimum bid ended successfully in 72% of the auctions as opposed to 52% for secret reserve auctions. This means that there is a significant difference in bidding behavior when items are placed on auction with a reserve versus a low opening bid and no reserve.

TRANSACTIONS COSTS AND RISK

Exchange always contains an inherent element of risk for the contracting parties. In the case of online auctions the level of risk is higher for the buyers for several reasons. First, even though feedback on sellers does provide evidence of their reliability, too little information may be available to warrant complete confidence. Nonetheless, McDonald and Slawson (2000) have statistically shown that a seller's good reputation increases both the number of bidders and the prices in eBay auctions. Most vendors work hard to establish an unblemished reputation for honesty and fairness—even going so far as to provide a money back guarantee. Yet some unscrupulous individuals can and do commit fraud—keeping the money, but never sending the buyer the item that they paid for. Now, with escrow account services and online payment systems such as PayPal, this risk has been somewhat reduced. With escrow capabilities, a third-party ensures payment after the buyer receives the item. Electronic payment systems provide fraud investigation and verify their customers. Electronic payment services open up new opportunities for fraud, however. By subscribing to services such as PayPal, members must provide bank account and credit information over the Internet. Second, even when the vendor is reliable, the possibility exists that the goods are damaged or inappropriate for the buyer's intended purpose. A seller's good reputation is reassuring, but it does not always provide the same measure of certainty as a physical inspection of the goods. Third, auction prices are not guaranteed to be cheaper than one could find at a local store. Numerous opportunists pander goods at inflated prices hoping to snag sales from inexperienced buyers. Before placing a bid, experienced buyers carefully review recent auctions to ascertain prices. Although prohibited by online auctions, some unethical sellers use shills to inflate bids. The Latin adage *caveat emptor* (let the buyer beware) is sound advice for online bidders.

The Internet provides a new and effective medium for buyers to collect up-to-the-minute price information. Most online auctions provide an extensive database of previous transactions that enable bidders to gather price information. However, collecting up-to-date price information can be time-consuming. Pricing standardized commodity items in new condition is relatively easy, but when the condition varies, it becomes more difficult to determine accurate values. The costs of finding the auction, of reviewing seller feedback, and estimating a bid price all constitute the transactions costs of participating in online auctions. In effect, online auctions shift a large portion of the transactions costs on to the buyer. Given the increasing popularity of online auctions, these costs are accepted by buyers partly because they can be combined with other online activities.

For comparative purposes, consider the following equation, which models online auction buying:

$$Pa = Pb + T + Sc + W + R + Tax$$

Pa is the total cost of the auction for the winner. *Pb* is the winning bid. *T* is the search costs of reviewing seller feedback and previous closing prices. *Sc* is the shipping cost charged buy the seller. *W* is the disutility (delayed gratification) that comes from waiting for delivery. *R* is the normally distributed risk for fraud or dissatisfaction. *Tax* is the sales tax, where applicable, from purchasing online. It is relevant to point out that some of these costs are incurred whether or not a bidder wins the auction.

This equation captures the most significant differences from over-the-counter sales. First, no sales taxes are collected on online auctions unless the sale is made to a buyer in the same state where online taxes are charged. As a result the auction price can be between 4–9% cheaper than physical retail stores depending on state and local taxes. (The application of sales taxes to sellers without a physical presence in the state where to goods are sold is currently under review.)

Second, even though online shopping can require extensive searching, these costs are not always lower in physical markets. The cost of visiting numerous brick-and-mortar stores to compare prices and quality during their regular business hours can be relatively more expensive than shopping at home. Shopping online shifts the time constraint and allows the buyer to ration more valuable prime time toward more highly valued activities. Third, in some instances, sellers require additional fees in excess of the actual cost of shipping the item. Fourth, once a buyer wins the auction she or he must wait until the item is delivered, which can take a couple of weeks. Fifth, the risk of loss or dissatisfaction is increased in online sales because the buyer never actually sees the item until after it has been purchased. That is why escrow services have been created. From the seller's perspective the auction cost is the cost of the item, the auction fees, plus the opportunity cost of their time to list the item, correspond with the buyer, as well as packaging and shipping cost. Dangers to sellers include colluding bidders who post low and unreasonably high bids. Toward the end of the auction, the high bidder withdraws his bid, allowing the low bidder to purchase the item at the low price. Given all of these considerations it is a statement of true market behavior that such a great percentage of auctions are successful.

CONCLUSION

Online auctions make it easier for buyers and sellers to interact without leaving their homes or offices. At the same time, it proves a form of recreation and entertainment. The Internet encourages comparative shopping but also facilitates random impulse buying. Even with the availability of pricing information, it is common to observe the "winner's curse" when a buyer gets caught up in the excitement of the auction or overestimates the value of an item. Furthermore, participating in online auctions can take up a great deal of time. Despite these drawbacks the popularity of online auctions is increasing.

In conclusion I offer the following observation: In general, the success of online auctions is based on four factors: name-brand recognition, Web site architecture, trust, and the combination of talents and business processes accumulated over time.

Name-brand recognition is a function of a number of factors including the length of time that the auction had been in existence, the size and the volume of the auction, and the quality of the service provided to both buyers and sellers. Offering higher quality in this industry means targeting bidders with a higher propensity to buy the products with faster searches and focused product offerings.

Architecture relates to the construction of the Web site, its overall ease of navigation, and the speed with which searches are conducted and transactions posted. It includes the ability to efficiently take bidders into databases for the retrieval of past auction prices, feedback, and even to other auctions for comparisons.

Trust includes the feedback rating on buyers and sellers, but also includes insurance and fraud protection for auction purchases.

Business processes are a function of the human talent accumulated over a range of experience. These factors result in the continuous innovation of new features while improving the security and quality of the basic service.

The online auctions industry is made possible because of the Internet and is therefore affected by changes in technology, government policies, and changing customer expectations. The future of competition will depend on a company's ability to adapt to future changes.

One of the biggest threats to online auctions is the possibility that state and local governments could apply sales taxes on transactions. The Advisory Commission on Electronic Commerce created by Congress recommended continuing a moratorium on any new Internet sales taxes that expired in 2001. But as online commerce increases, state and local governments, concerned that online vendors are eliminating many small retailers (Dalton, 1999) and thereby reducing state and local sales taxes, will move to apply Internet sales taxes. Various coalitions of states and brick-and-mortar retailers have started to lobby congress. Given the current fiscal crisis it is highly likely that states will soon start taxing internet transactions. Given the complexity of the many state and local tax structures, the movement toward the adoption of Internet sales taxes would likely require a plan for the simplification of the existing tax structures.

In any event, the decision to allow states to levy sales taxes on Internet sales on nonresidents would significantly reduce the difference in prices between virtual and physical shopping and undermine some of the competitive advantage that auctions currently enjoy. Goolsbee (2000) found that Internet sales are highly sensitive to sales taxes. His econometric analysis found that application of existing sales taxes to the Internet could reduce the number of online buyers by up to 24%. Likewise, it is possible that new laws and regulations will be adopted to protect consumers. Compliance with these laws could increase the cost of doing business. Foreign countries may also enact their own regulations that will make it more difficult to operate a single platform internationally. (See eBay filings with the Securities and Exchange

Commission http://www.sec.gov/cgi-bin/srch-dgar?text = eBay&first = 1993&last = 2002&mode = Simple.)

The preceding discussion deals with a facet of e-business that is still in its infancy. It is becoming clear that our traditional educational approaches will need to be reexamined in light of these developments. No doubt students fully integrated into the Internet will call into question much of what has been accepted by past generations. Goldfarb (2000), for example, described various class assignments he used to "raise the students' consciousness about the functions and mechanics of pricing in different economic settings." These include the classification of various types of auction, and the identification of the characteristics of auctionable items. The Internet provides a valuable opportunity for educators to redefine the research and teaching frontiers by developing curricula for teaching courses for the New Economy. As an experiment, I successfully participated in more than 50 auctions on eBay for a range of goods. In 50% of the auctions, identical goods from local merchants or catalog stores were available at a lower price. Given this, the commercial value of the online auction may be overstated. The entertainment value of following auctions, or competing with other potential buyers, may be in combination with the increased convenience a major reason for their continuing popularity. Certainly, we are living in an age where an Internet business is by definition global, and consumers have more choices that at any other time in human history.

ACKNOWLEGMENT

An earlier version of this paper was presented at the Fiftieth International Atlantic Economic Conference, Charleston, October 15–18, 2000. The author expresses his appreciation to Marilyn Greenstein, Pierre Balthazard, P. Krishnakumar, Robyn Stout, and the reviewers for their comments.

GLOSSARY

Common value auctions Auctions in which the item has the same value to all bidders.

Consumer surplus Savings realized by a buyer who pays less that his or her reservation price for an item.

Dutch auction Originally a term used by economists to describe a descending price auction, but that now describes an online auction for multiple units of an identical item.

Economies of scale A reduction in the average cost of producing a unit of output associated with increases in the quantity of output.

Electronic payments Companies such as PayPal and Billpoint that provide successful bidders or buyers the ability to pay for an item using credit cards or electronic checks.

Feedback Comments made by the buyer or seller after the conclusion of an auction that provide information on the transaction and establishes a reputation for each party.

First-degree price discrimination Extracting the consumer surplus from each and every buyer.

Insertion fee A fee paid by a seller to list an item for auction; this fee varies with the value of the item, the reserve price, and the type of auction.

Private valuation auctions Bidders know with certainty what the value of the item is to them.

Proxy bidding Using an automatic feature that will incrementally raise the bid until it reaches the bidder's reserve price or the auction is successfully completed.

Reserve price auction An auction in which the seller places a secret "reserve price" that is the lowest price for which the good can be sold; to win the auction, the bidder must place the highest bid that is equal to or greater than the seller's reserve.

Shill The use of a seller's accomplice to bid up the price in an auction.

Sniping A bidding strategy in which a buyer wins by placing a slightly higher bid in the closing seconds of an auction.

Transaction costs All of the costs involved in a transaction including the cost of collecting information, following the auction, and arranging payment.

Winner's curse A winning bid greater than the value of the item; this is generally associated with common value auctions, but it can occur anytime a buyer overvalues an item.

CROSS REFERENCES

See *Business-to-Business (B2B) Internet Business Models; Business-to-Consumer (B2C) Internet Business Models; Electronic Payment; Online Auction Site Management.*

REFERENCES

Bajari, P., & Hortacsu, A. (2000). Winner's curse, reserve endogenous entry: Empirical insights from eBay auctions. Retrieved August 2, 2000, from http://papers.ssrn.com/paper.taf?abstract_id = 224950

Bapna, R. (2000). IS perspective of research issues in electronic commerce and online auctions. In M. Chung (Ed.), *Proceedings of the Sixth Americas Conference on Information Systems* (pp. 926–928). Long Beach, CA: California State University.

Biglasier, G., & Mezzetti, C. (2000). Incentive auctions and information. *Rand Journal of Economics, 31,* 145–164.

Bollinger, A. S., & Madey, G. R. (2000). An analysis of electronic auctions as a mechanism for supply chain management. In M. Chung (Ed.), *Proceedings of the Sixth Americas Conference on Information Systems* (pp. 936–941). Long Beach, CA: California State University.

Brannman, L., & Froeb, L. (2000). Mergers, cartels, set-asides, and bidding preferences in asymmetric oral auctions. *The Review of Economic and Statistics, 82,* 283–290.

Bulow, J., & Roberts, J. (1989). The simple economics of optimal auctions. *Journal of Political Economy, 7,* 1060–1090.

Dalton, G. (1999, October 4). Going going gone. *Informationweek,* 45–50.

Dalton, G. (2000, February 15). Online auctions pick up. *Informationweek,* 37.

Dasgupta, P., & Maskin, E. (2000). Efficient auctions. *The Quarterly Journal of Economics, 115,* 341–388.

eBay, your personal trading community website. Retrieved from http:/w.w.w.eBay.com

eBay filings for the U.S. Securities and Exchange Commission. Retrieved on October 20, 2000, from http://www.sec.gov/cgi-bin/srch-dgar?text = eBay&first = 1993&last = 2002&mode = Simple

Emmanuel, G., Perrigne, I., & Vuong, Q. (2000). Optimal nonparametric estimation of first-price auctions. *Econometrica, 68,* 525–574.

Forrester Research. (n.d.) Forrester findings: Internet commerce. Retrieved on December 17, 2000, from http:// www.forrester.com/ ER / Press / ForrFind /0,1768, 0, 00. html

Goldfarb, R. S. (2000). An Onassis retrospective: What products are auctioned, and why? *Journal of Economic Education, 31,* 157–168.

Goolsbee, A. (2000). In a world without borders: The impact of taxes on Internet commerce. *Quarterly Journal of Economics, 115,* 561–576.

Kagel, J., & Roth, A. (1995). *The handbook of experimental economics.* Princeton, NJ: Princeton University Press.

Katkar, R., & Lucking-Reiley, D. (2001, March). *Public versus secret reserve prices in eBay auctions: Results from a Pokemon field experiment* (NBER Working Paper No. w8183). Cambridge, MA: National Bureau of Economic Research. (Available online at http://papers.nber.org/papers/W8183)

Lucking-Reiley, D. (1999). Using field experiments to test equivalence between auction formats: Magic on the Internet. *The American Economic Review, 89,* 1063–1080.

Lucking-Reiley, D. (2000a). Auctions on the Internet: What's being auctioned, and how? *Journal of Industrial Economics, 48,* 227–252.

Lucking-Reiley, David. (2000b). Vickrey auctions in practice: From nineteenth century philately to twenty first century e-commerce. *Journal of Economic Perspectives, 14,* 83–92.

McAfee, R., & McMillian, J. (1987). Auctions and bidding. *Journal of Economic Literature, 25,* 699–738.

McDonald, C. G., & Slawson, V. C. Jr. (2002). Reputation in an Internet auction market. *Economic Inquiry*, 40, 633–650.

Milgrom, P. (1989). Auctions and bidding: A primer. *Journal of Economic Perspectives, 3,* 3–22.

Roth, A., & Ockenfel, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions of the Internet. *American Economic Review, 92,*1093–1103.

Teschler, L. (2000, March 9). Let's start the e-bidding at $50. *Machine Design,* 148–151.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance, 18,* 8–37.

Vigorosa, M. (1999, January). On the Internet. *Purchasing, 126,* 85–86.

Ward, S. G., & Clark, J. M. (2000). Bidding behavior in virtual versus "live" auctions: An examination of the eBay collectibles market. In M. Chung (Ed.), *Proceedings of the Sixth Americas Conference on Information Systems* (pp. 945–946). Long Beach, CA: California State University.

Westland, J. C., & Clark, T. H. K. (1999). *Global Electronic Commerce.* Cambridge, MA: The MIT Press.

FURTHER READING

Klein, S., & O'Keefe, R. (1999). The impact of the web on auctions: Some empirical evidence and theoretical considerations. *International Journal of Electronic Commerce, 3,* 7–20.

Milgrom, P., & Weber, R. (1982). A theory of auctions and bidding. *Econometrica, 50,* 1089–1122.

Shapiro, C., & Varian, H. R. (1999). *Information Rules.* Boston, MA: Harvard Business School Press.

Online Auction Site Management

Peter R. Wurman, *North Carolina State University*

INTRODUCTION

Internet auctions appeared on the scene in the mid-1990s and quickly became one of the most successful applications of electronic commerce. eBay, the premier consumer-to-consumer (C2C) Internet auction site, is generally held up as an exemplar for the industry. It is widely predicted, however, that the potential transaction volume in business-to-business (B2B) auctions will be much greater than in the C2C channel (Keenan, 2000; Rosenthal, 2002).

In the B2B marketplace, auctions were initially pressed into service as tools to dispense with excess inventory. One-time market leaders such as Onsale helped companies, primarily consumer electronics manufacturers, sell products near the end of their lifecycle. Onsale was founded in 1996 and made a name for itself early on as an auction-based outlet store. Eventually, they added a C2C component in an attempt to compete with eBay. In what now seems like the prototypical dot-com arc, Onsale merged with Egghead in 1999, and the combined company filed for bankruptcy in 2001.

The current wave of B2B integration represents a much deeper integration of auction technology into the day-to-day operations of many businesses. In particular, companies are using auctions in many procurement situations in an effort to extract better prices from their suppliers. This move toward more formal and rigorous negotiation with suppliers folds neatly into the need to better manage the supply chain and make operations more efficient. The third wave of integration will involve the use of dynamic pricing through the entire product lifecycle, on both the sales and procurement sides. We are already starting to see evidence that products are being sold by auction earlier in their lifecycles; IBM and Sun regularly sell relatively new products via a distribution channel on eBay.

Although eBay sets the standard for features and performance in C2C markets, the potential applications of auctions to B2B markets requires more varied and complex systems and places different burdens on the auction administrators. On the one hand, B2B auctions may require more elaborate and sophisticated auction processes, such as combinatorial auctions (Rothkopf, Pekec, & Harstad, 1998). On the other hand, B2B markets tend to have well-known and authenticated bidders and lower overall communication loads.

Auction systems have three distinct sets of users: the bidders, the auction initiator, and the auction system administrator. Each class of users requires different core and complementary features of an auction system. Note that we do not treat sellers and buyers separately; in the general case, they are both bidders and have similar needs. On most C2C sites, sellers are the auction initiators and place their one and only bid to establish the reserve price during the creation of the auction, if at all. The auction system administrator is the person (or group of people) who installs, configures, and maintains the auction site.

This chapter discusses many of the issues related to Internet auctions for B2B, B2C, and C2C marketplaces. The discussion is framed by a brief history of auctions on the Internet, followed by an introduction to some central concepts. The rest of the article is divided into two parts: a description of the core features of auction systems followed by a discussion of features that are complementary and commonly affect the design or selection of an auction system.

BRIEF HISTORY OF INTERNET AUCTIONS

It is difficult to pinpoint the earliest auction held on the Internet, but it is clear that auctions were conducted via e-mails and newsgroups as early as 1988. (See the historical record, formerly called deja-news, at http://groups.google.com/.) As the Web developed in the early 1990s, it was only a matter of time before people began using this new technology to enhance their online auctions. At first, sellers simply provided a static source of multimedia information (i.e., text and images) about the products being auctioned and continued to collect bids via e-mail. Later, sellers used Web forms to collect and process bids. eBay, founded in 1995, was among the earliest known auction services available on the Internet. Within a year, it had many competitors, including uBid, Onsale, Z-auctions (now defunct), and more. In the B2B arena, Fastparts.com was an early market maker, having bulletin-board-based systems as early as 1991 and launching a Web-based system around 1996 (information from http://www.fastparts.com/fpwebsite/jsp/about/about/jsp). Internet auction platforms were also being developed in research laboratories in 1996, including the Michigan Internet AuctionBot (Wurman, Wellman, & Walsh, 1998), the FishMarket system (Rodriguez, Noriega, Sierra, & Padget, 1997), and GEM (Reich & Ben-Shaul, 1998).

The need to build hundreds of electronic auction sites rapidly in a wide variety of industries created opportunities for packaged and hosted software systems. To satisfy the demand, several companies built and marketed auction software. Among the best known of the auction service providers were OpenSite, Trading Dynamics, Moai, and FreeMarkets. FreeMarkets was founded in 1995, and Moai formed in 1996. OpenSite began life as Web Ducks around the same time, and Trading Dynamics was founded in 1998. OpenSite and Trading Dynamics were acquired in 2000 by Siebel Systems and Ariba, respectively. Meanwhile, IBM was building its own auction engine (Kumar & Feldman, 1998), and CommerceOne acquired CommerceBid to obtain auction technology. Moai and FreeMarkets both remain independent companies. Although these more established companies remain the leaders, the field has become crowded with new ventures, such as i2, ProcureNet, Frictionless, and ICG Commerce, and procurement solutions introduced by multinational companies like SAP, General Electric, and IBM.

In the early days of e-commerce, it was logical for auction software companies to design flexibility into their products so that they could be employed in many application markets. Thus, these products typically are designed to allow customization, to some degree. The amount and type of customization depends on the particular details of the target market and whether the auction engine was slanted more toward B2B, B2C, or C2C applications. Today most major e-commerce vendors have an auction product targeted at B2B procurement, although these products vary widely in features and customizability.

More recently, there has been growing interest in advanced auction formats. These advanced formats include *combinatorial* auctions—auctions that allow bidders to place offers on sets of items. Several companies have recently begun offering systems that manage combinatorial trading. CombineNet, Emptoris, NetExchange, and TradeExtensions are among the leaders in commercial combinatorial auction systems. Another type of advanced auction that is starting to appear in electronic systems is the multi-attribute auction, in which the object (or contract) being negotiated has several negotiable parameters in addition to price. An oft-cited example is a negotiation between a manufacturer and a supplier in which product quality (or purity) and delivery date are negotiable. In a multi-attribute auction, different suppliers will offer contracts that differ in the promised product quality or delivery schedule, and the buyer needs a well-defined method for selecting among the contracts and (optionally) generating the equivalent of price quotes.

It is also important to note that Internet auctions were not the first auctions facilitated by electronic networks. NASDAQ's Small Order Execution System began operation in 1984, and many of the stock exchanges now incorporate electronic trading or electronic support systems. Historically, financial systems ran on closed networks, with custom trading stations at designated, access-controlled locations. Operating markets on the public networks raises new concerns, such as security, privacy, identify verification, and network availability. In addition, the much wider variety of products and services being negotiated on the Internet requires versatile trading systems.

AUCTION CONCEPTS

Although the term *auctions* tends to bring to mind the classic situation with a seller offering a single item to the highest bidder, like eBay or the stereotypical face-to-face auction, the accepted definition of the term in economics includes a wide variety of negotiation mechanisms. The English auction, procurement auctions, and stock markets are members of a large class of negotiation mechanisms that can rightly be called auctions and can be precisely defined by their rules. It is useful to collect the rules into three related sets: rules that govern the admission of bids, rules that govern the information revealed by the auction, and rules that govern how the auction computes trades. A detailed mathematical treatment of these rules is available elsewhere (Wurman, Wellman, & Walsh, 2001). Of concern here is how the parametrization affects the choice of an auction system.

Flexible auction systems are built around some notion of parametrization, although most auction systems have simplified interfaces that make it easy to implement common configurations. For example, an auction initiator can simply select "English auction" to get a standard ascending auction rather than go through the potentially confusing task of specifying the bidding rules and quote and clear policies that define the English auction.

Product Specification

The first task is to specify the product or service to be traded. In C2C auctions, it is common to sell single objects. C2C systems also permit *multiunit* auctions in which multiple copies of a single object are offered for sale. In B2B scenarios it is common to procure services and longer term contracts for physical products. In procurement auctions groups of objects are often purchased together. When the objects can be bid on independently, the auction is referred to as a *multiobject* auction. When some of the objects must be bid on as a group, they are called a *lot*.

Bidding Rules

Bidding rules define the types of bids that are allowed and which participants are allowed to place them. In a single seller auction, the designated seller is the only participant who can place a sell offer and is typically required to do so at the beginning of the auction. If that offer is nonzero, then the seller's offer is called the *reserve price*. In a procurement situation, only the designated procurer can place a buy offer, and typically, the members of a prescreened set of suppliers are the only participants who can place sell offers. In an open, continuous double auction (CDA), such as the stock market, any participant can place either a buy or sell offer. In fact, in a CDA, a bidder can place offers to buy and sell simultaneously, as long as the buy offers are less than the sell offers. Much of the flexibility of auction systems comes from treating buyers and sellers symmetrically.

Different auction mechanisms may have different languages for expressing the bids. The simplest type of bid is an offer to buy or sell one unit at a specified price. When multiple units are being traded, the language may allow the bidder to express a *bid schedule* in which the buyer (seller) expresses how many units she would like to buy (sell) at various prices. This allows a bidder to, for instance, say, "I'll buy one unit if the price is less than $50 and two units only if the price is $35 or less." A language that allows multiunit bids must have clear semantics. In particular, it must be clear whether the auction has the right to satisfy the bids partially. For example, the statement, "I'll buy two units if the price is less than $25," can be interpreted as, "I'll buy exactly zero or two units if the price is less than $25" or "I'll buy zero, one, or two units if the price is less than $25." In the latter case, the bid can be partially satisfied, whereas in the former case it cannot. When bids cannot be partially satisfied, the auction's clearing and quote generation tasks may become computationally intractable.

In combinatorial or multiattribute settings, bidding languages can become even more complex, and the ability to express bids concisely becomes an issue (Nisan, 2000). Moreover, the auction designer can choose among several rules that restrict bids as a function of the current price quote or of the bidder's previous offers. The *beat-the-quote* rule requires that a new bid be better than the price information revealed by the auction. In the English auction, the interpretation of the beat-the-quote rule is obvious and simple to implement. In advanced auctions, however, the rule can be subtle to interpret and apply and does not always achieve the desired outcome of progressing the auction. The *improve-your-bid* rule is an alternative that requires that a bidder's new bid be a strict improvement over his or her previous bid. Although not obviously useful in an English auction, in advanced auctions the auctioneer will often need to use the improve-your-bid rule in conjunction with the beat-the-quote rule to achieve the desired results. Ausubel and Milgrom (2002) proposed a combinatorial auction in which the improve-your-bid rule is the only bid improvement rule.

Recently, market designers have started using *activity rules* to determine which types of bids are permitted. Like the improve-your-bid rule, activity rules define allowable new bids based on the bidder's previous bid state. Activity rules condition the allowable bids on a variety of measures of previous activity, such as the number of objects the bidder was winning or the number of improved offers that were made in the last bid. The discussions that surround the government auctions for spectrum licenses focus, to a large extent, on the selection of activity rules that encourage progress without overconstraining the bidders (McAfee & McMillan, 1996).

In addition to controlling the *who* and *how* of bidding, the rules control the *when*. In simple scenarios, bids are accepted at any time. Some auctions are organized into timed rounds, however, whereas others require an offer from each designated participant before they can progress.

Intermediate Information

Auctions that generate no intermediate price information are called *sealed-bid* auctions. The vast majority of online auctions generate intermediate information to help guide the bidders. Typically, the information generated is a current price, often presented in the form of a *bid-ask* spread. Although the bid-ask concept is inspired from the stock market—a continuous double auction in which the buyers and sellers offers do not overlap—these concepts can be generalized to a larger class of auctions (Wurman, Walsh, & Wellman, 1998). In addition to price information, an auction may reveal a list of the current bids—called an *order-book*—or the identities of the bidders.

Price information becomes more complex to compute and express with advanced auctions. For instance, in multiunit auctions that allow bidders to make all-or-none offers, a simple price-per-unit announcement may not accurately inform the bidders whether they are currently winning the auction. In fact, an all-or-none offer may be turned down in favor of an offer that has a lower per-unit value because the overall value of the trades is greater. For example, an auction in which the seller is offering three items may choose a bid to buy exactly two items at $4 each, and a bid to buy one item for $2, and not select an offer to buy exactly two units at $3 each. The former two bids generate $10 of revenue for the seller, and no combination using the third bid can do as well. Most C2C auctions finesse the problem by listing the current undominated bids and whether they are winning, instead of announcing prices. It is left up to the bidder to compute what bid they would have to place to become a winner, or to simply improve their bid and observe the effect.

Clearing

The final collection of rules determines when and how an auction computes trades. The act of computing trades is called *clearing* and is handled by a well-defined policy instantiated as an algorithm in the auction software. An auction may clear whenever a bid is received (i.e., continuous clearing), when no new bids are received for a specified time (e.g., the typical English auction), at a prescribed fixed time (e.g., eBay), or on a prescribed schedule. The policy that is used to compute trades is generically called the *matching function*. There are many types of matching functions, some of which are described in the companion article on online auctions (see Anders, this volume). The various matching functions trade off desirable properties that cannot be achieved simultaneously even when a single item is being sold by a seller to a single buyer. In combinatorial scenarios, the tradeoffs are more complex and are further influenced by computational factors.

THE CORE OF AN AUCTION SYSTEM

The rules discussed here suggest a generic architecture that delegates the responsibilities of the three main rule sets to corresponding components of the architecture. Although specific auction systems will vary in their actual software architecture, most can be abstracted into the generic software architecture shown in Figure 1.

The diagram illustrates some of the core interactions that bidders and initiators have with an auction system, and some of the primary internal procedures. In actual commercial systems, the internal components may be more or less clearly delineated than in the figure, but the responsibilities remain essentially the same. Industrial strength databases are used to store information about the auctions, bids, users, and transactions. The use of a database ensures data reliability and facilitates the communication of data between the disparate components of the system. Although clearly beneficial, adding a database also adds significant computational overhead, and communicating with it is often a performance bottleneck.

An auction is created when a user (the initiator) interacts with the *auction manager* to specify and launch a new auction. Auction creation can be a time-consuming process. The initiator must clearly describe the product, including as much detail as necessary for bidders to know exactly what they are buying or selling. Often the product description will also contain information about the shipping and handling, payment process, and other details pertinent to the transaction. In B2B procurement scenarios, the process is even more onerous, and many procurers soon realize that they have to generate precise definitions of the products and the contractual obligations before the suppliers will have the information they need to bid properly. Occasionally, the initiator will need to clarify information during the course of the auction. Generally, updates should augment, rather than replace, the original text, and bidders who placed bids before the update should have the option of retracting their bids if the description has changed dramatically.

The initiator must also select the rules of the auction and choose key parameters such as the duration, bid increments, and format of the bids. In many B2B scenarios, the initiator will specify the set of bidders that are permitted to participate in the auction.

The *bid manager* component is responsible for enforcing the bidding rules and admitting bids that satisfy them into the system. To enforce some types of bidding rules, the bid manager may have to extract the auction description and current state from the database. For instance, to enforce the beat-the-quote rule, the bid manager needs to extract the auction's current quote information, whereas to enforce the beat-your-bid rule the auction must extract the bidder's previous bid. Once the bid manager has determined that a bid satisfies the conditions, it *admits* it into the set of current bids and stores it in the database. Bids that are admitted are termed *valid,* and those that are not admitted are *rejected*.

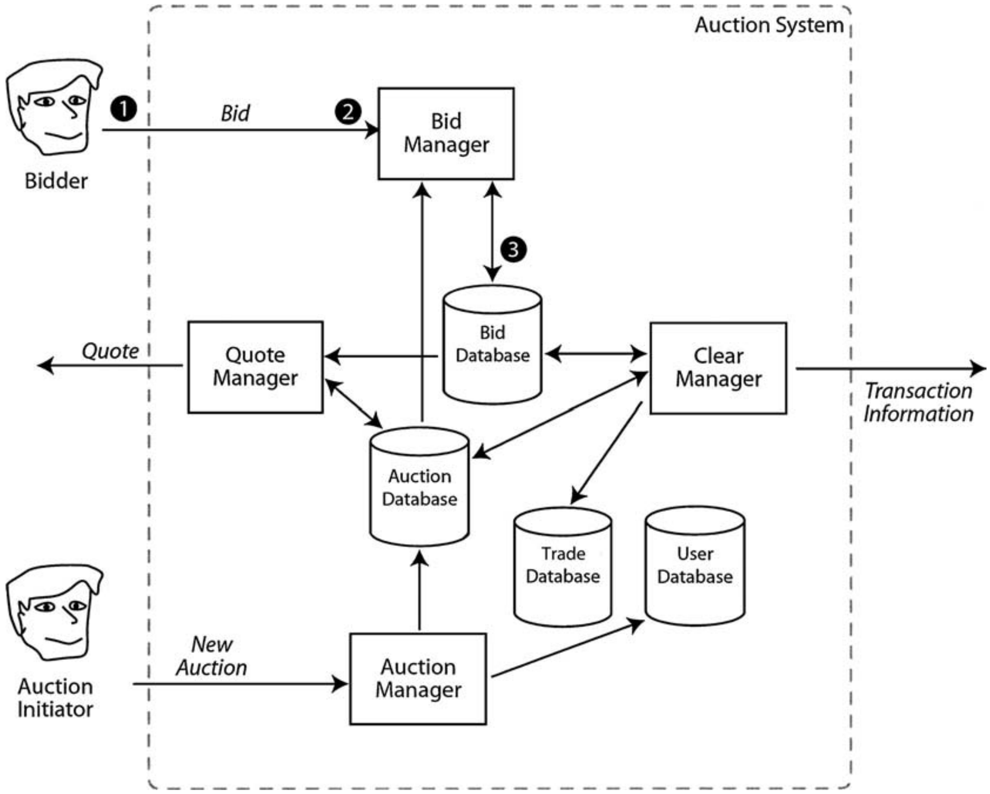


Figure 1: A model auction architecture.

The *clear manager* is responsible for executing clear actions. It must read the bids and the auction's configuration from the database and compute *trades* according to the auction's policies. The new trades are stored in the trade database. In addition, information about the bids and the auction may have to be updated in their respective databases. For example, a bid that becomes part of a trade needs to be marked as transacted in the bid database.

The *quote manager* is responsible for generating price or other information about the auction in accordance with the auction's policies. The quote manager may update the bid database in order to track which bids are currently winning, and may store new quote information as part of the auction database. Both the quote manager and the clear manager must communicate the results of their actions with the interested parties, as indicated by the arrows in Figure 1.

Scheduling Actions

Not shown in Figure 1 is a system for triggering the actions and coordinating the activities of the various components. A critical feature of an auction system is its ability to guarantee the proper sequencing of actions. For example, if a bid is received 1 second before a clear event, the bid should be considered when the clear is computed. If two bids are received nearly simultaneously, the first bid should be handled before the second or an incorrect trade may result. An auction system that provides a guarantee that events are handled in correct chronological order is *consistent* and *auditable*—if an external auditor looked at the bids after the auction was over, she would be able to show that the auction did the right thing. Often, these demands conflict with the desire to make the system perform in real time. Various auction systems handle this complicated trade-off differently.

One method of guaranteeing consistency is to have a *sequential* system. In a sequential system, there is only one thread of control for a given auction, thus guaranteeing that events happen in the correct order. A sequential system cannot handle a new event until it has finished processing the previous one, however; it cannot accept a new bid while it is generating a price quote or admitting a previous bid. If any of the events take a noticeable amount of time to compute, the system will appear to be unresponsive, or unavailable for other uses. A sequential system also does not benefit from the advantages of multithreaded processing, such as improved use of processor and I/O (input/output) resources.

A *parallel* auction system can handle many auction events simultaneously, but greater care must be taken to ensure that the events are handled in the correct order. It is relatively straightforward for the auction program to sequence the quote and clear activities correctly when they are internally triggered, but when they are triggered by the asynchronous submission of bids, sequencing is more difficult. The *bid submission time*—the point at which the bid is officially considered admitted—is a key design element of the system. Essentially, there are three choices, corresponding to points at which the bid is *spoken, heard,* and *understood*. These points are labeled (1), (2), and (3) in Figure 1.

Users would like the bid submission time to be the moment when they click the submit button on a Web form (1); this is problematic for a number of reasons, including the fact that the submission time is measured by the local machine, whose clock cannot be verified by the server. The second choice is to stamp a submission time on the bid when it reaches the server (2). In this case, all the bids can be tagged with a central clock, and the server's response to the user can state that the bid is officially submitted but has not been validated by the bid manager. In addition, the system must be designed so that clears and quotes wait until the bid manager has actually validated all of the submitted bids, introducing a level of asynchrony into the system (Wurman et al., 1999). The third choice is to consider the bid submitted only after it has been validated by the bid manager (3). The advantage is that a definitive response regarding the admission of the bid can be returned to the user. In many cases, however, validation requires a call to the database and therefore takes time. In advanced auctions, the evaluation of a bid's validity may be computationally complex, and the response back to the bidder must be delayed until this computation is complete.

Some auction systems relax the guarantee regarding the admission of bids for some types of events. For example, at the time of this writing, eBay's system does not guarantee that its beat-the-quote rule is enforced for bids received within a very small window of time. While examining the data, we have found cases in which a bid was admitted with a time stamp that is one second after a bid that would have made the former bid rejected (Shah, Joshi, & Wurman, 2002). Presumably due to the asynchrony, the earlier bid had not been fully processed and the new minimum bid computed when the later bid was admitted. Although eBay relaxes the temporal ordering while processing two competing bids, it does not do so when processing the clear event.

These issues become more complicated when the auction system is distributed over several hardware devices. Like other three-tier architectures, auction systems can rely on a database that resides on another machine. Many auction systems also benefit from the fact that the auctions themselves are independent of one another. This enables, for instance, all auctions related to Product 1 to be stored in a database on Machine A, whereas all auctions related to Product 2 are stored in the database on Machine B. The ability of the auction software to be distributed in this way goes a long way toward making the system *scalable*. As in other distributed architectures, load balancing becomes an issue. The issues of effective load balancing in auction systems have not been carefully addressed in the literature on auctions. (Interestingly, markets have themselves been promoted as solutions to load balancing problems; Clearwater, 1996. Essentially, the machines in the cluster auction their time; the busier a machine is, the more expensive its time becomes.)

Notification

Figure 1 suggests that information flows out of the auction system to the users in the form of price quotes or information about trades. In the early days of auction systems, participants received up-to-date information either through text-based e-mail or by checking the Web site. These two methods represent two approaches to the communication process. E-mail based notification is a *push* technology, that is, when information changes, the server pushes the new information to the clients. Web-based dissemination is a *pull* technology, and clients must explicitly request the latest information from the server. Even as auction systems are increasingly connected to more mobile information devices (eBay can send notifications to e-mail enabled mobile phones, pagers, and personal digital assistants), the push-pull distinction still divides the multiplicity of communication channels.

Whether an auction system uses push or pull technology, or both, depends on the application, but the two formats put different stresses on the architecture. In a push system, users will typically subscribe to various types of notifications. Each time an event occurs for which some people have requested notification, the server must compose and send a message. In some settings, this process may generate more messages than are needed to keep the users informed. For instance, if I am away from my computer for a while, when I log on again, I may not need to know all of the intervening price changes, but only the most recent.

Pull-based communication does away with the excess of information but may introduce and excess of information-less messages. In a pull system, users request information only when they are ready to look at it. Because they don't know whether the information has changed on the server, however, they will be forced to make frequent requests to keep their information up-to-date, and the majority of those requests have no new content. Thus, when users need the most up-to-date information (e.g., at the end of an eBay auction), they will make frequent requests to the server and can easily overwhelm the system. This may be true even when push-based information is provided because push channels—like e-mail—are generally asynchronous themselves and don't provide the most up-to-date information.

There are several approaches to controlling the load placed on the server by information requests. One method is to cache the information in files (or memory) so that it can be served quickly. This method is successful for content-oriented sites but can be problematic for dynamic sites like auctions. eBay's category listings are emblematic of the difficulty of keeping all views of an auction up-to-date. The category listings are not refreshed in real time, and it is common to follow a link from the catalog to the auction description and find that the current price is much higher than the price listed on the catalog, especially near the end of the auction. A system architect's alternative to caching the catalogs is to generate them on demand, an approach that would be untenable given the number of user requests and the number of database queries involved on a site like eBay, but which may be practical for smaller sites.

Scaling

Like any Web system, the hardware and software architecture has to be scaled to handle the expected load. An auction site manager should be concerned with at least three types of loads: *page views, bidding actions,* and *internal actions* like clears and quotes. Page views measure the number of requests for HTML documents from the clients. Static pages can be served faster than dynamic pages (which require calls to a database). Unfortunately, auctions are highly data-driven sites, and many of the pages, like the auction status, must be constructed at a moment's notice. Bidding actions put an extra strain on the system because writing to a database is slower than reading from one, and temporarily locks others out. Thus, the load due to bidding should be considered separately from that due to page views. Finally, the events triggered internally compete for computational resources, and steps must be taken to minimize negative effects from concurrent internal events.

To scale the system properly, it is necessary to estimate the loads. This can be done easily from an estimate of the number of auctions hosted and educated guesses about the behavior of the users. For example, if a C2C auction site hosts x auctions per time period t, and we expect an average of b bidders each with r bids per auction, we immediately have the expected number of bids, $brx\ /\ t$. To get an estimate of the number of page views, suppose that the seller and bidders track the auction by viewing the auction description v times each. The number of page views due to this activity will be $vx(b+1)/t$. In addition, bidders will look at n auctions before selecting one in which to bid. Thus, we would expect an additional nb/t page views due to searching. It is relatively easy to estimate some of these parameters by examining bidding patterns on existing sites such as eBay. These calculations return an average load. A system administrator planning the hardware to run an auction site will also be concerned with the peak loads, which can also be estimated with reasonable accuracy.

The choice of auction rules also affects the load. Consider the effect of the rules used by eBay compared with those used by Amazon's auctions. On eBay, auctions close at a fixed time, regardless of activity. This encourages a bidding technique called *sniping* in which a bidder attempts to place a bid high enough to win at the very end of the auction. To execute this strategy, the bidder must have the latest information about the current price, and thus the bidder makes many requests to eBay's server as the auction nears its end. On Amazon, auctions are extended for 10 minutes following the last bid, and close when no new bid is received in that period. This avoids the end-of-auction frenzy that is common on eBay and distributes the bidding more evenly over time (Ockenfels & Roth, 2002).

eBay employs two techniques to help smooth out the load. First, it does not give the seller direct control over the closing time of the auction. Instead, the auction closes exactly (to the minute) 3, 5, 7, or 10 days after it is started, where the duration is chosen by the seller. This bounds the number of auctions that can close at a given time by the number that were created at the front end of the time window. Although tools are available that help sellers automate the listing of items, these tools must still operate by interacting with eBay's Web servers by submitting HTML form data or, more recently, Extensible Markup Language (XML) documents, via hypertext transfer protocol

(HTTP). Thus, requests to initiate auctions are funneled through a gateway. The net effect of eBay's rigid control of auction creation and duration is that auction closings are distributed over time, diminishing the peak loads.

The second approach eBay uses to smooth the load is to allow *proxy bidding*. EBay's proxy bidding system allows users to submit a maximum willingness to pay, and then automatically increases the user's bid on his or her behalf until the maximum is reached or the auction closes. When users take advantage of eBay's proxy bidding, they reduce the direct interaction with the system. In addition, when two proxy bids are submitted, eBay can directly compute the result of the subsequent bidding war without actually computing all of the intermediate bids. The theory (Roth & Ockenfels, 2002) predicts, and empirical studies (Shah et al., 2002) support, that straightforward proxy bidding is not the only strategy employed by bidders.

Despite the fact that eBay provides a proxy bidding feature, third-party bidding software is becoming popular with eBay's users. Companies like eSnipe and Auction Blitz offer sniping services, and there are many "free" programs available for download. These sniping tools are simple, but the potential benefits of bidding assistants—*agents*—are tremendous, in both C2C and B2B environments. Although auction sites have discouraged these tools, the potential benefits to bidders are so great that it seems unlikely that bidding agents can be suppressed for long. This has serious implications for online auction managers. An architecture in which agents make requests for Web pages and parse them to extract the relevant information is inefficient for both the agent and the server. As the number of automated bidding agents increases, auction sites are going to feel the need to provide specialized communication channels. It is clear that XML will be the meta-language of choice for constructing formal languages for querying and bidding in auctions. Indeed, most market-based research platforms now support XML based messaging (O'Malley & Kelly, 1998).

Although we have yet to see bidding agents explicitly designed for B2B auctions, there is a growing recognition of the value of optimization approaches to generating bids. Decision support systems for generating bids in procurement auctions have been demonstrated by IBM (Goodwin, Akkiraju, & Wu, 2002) and others.

COMPLEMENTARY FEATURES OF AUCTION SYSTEMS

Auction systems rarely exist as stand alone Web sites. Generally, an auction Web site provides complementary features that enhance a user's experience with the site. In fact, many of these features are essential to the success of an auction business.

The importance of the features will depend on the particular type of marketplace. Figure 2 shows how several types of marketplaces differ with respect to the number of bidders and the number of differentiated products available. Large C2C markets such as eBay have many users and many products simultaneously auctioned. Commodity and security markets, like the New York Stock Exchange, have many bidders but relatively few and static products. B2B exchanges, like FastParts.com or the automotive markets run by Covisint, have a moderate number of products and screened traders. Procurement auctions typically cover fewer products (e.g., contracts) and have a small set of preselected suppliers.

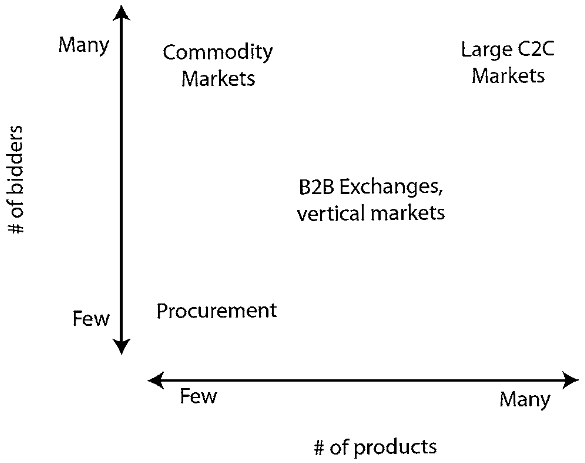


Figure 2: Space of marketplace structures.

Personalization

Bidders need an interface that allows them to create accounts, find product listings, and place bids. All sites require that bidders log on to place bids or initiate auctions. Authentication also allows the server to customize the user's view of the auction site. Typically, the custom view lets users keep track of their bids, watch auctions of interest, and manage other aspects of their accounts.

Catalogs and Search

Catalog and *search* features are two methods that enable bidders to find products, and both are typically provided on sites with large numbers of products. Catalogs are based on a natural, hierarchal organization of the products. Generally, catalogs are constructed by the auction managers, and thus the manager needs tools to construct and manage the catalog over time. The placement of a product listing in a catalog is usually left to the auction initiator. Thus, catalogs are effective only if the initiators place the products in appropriate locations. In C2C auctions, inevitably, some sellers either misunderstand the hierarchy or abuse the freedom, and list auctions in the wrong location. Fortunately, enough sellers correctly position auctions in the catalog that it is an effective way to locate categories of products.

Catalogs suffer from two other problems: they don't work for users who have hierarchal view of the product space different from that of the catalog organizer, and they often require many "clicks" to navigate. Search interfaces alleviate these problems. At a minimum, the search interface must provide a case-insensitive search for text in both auction titles and product descriptions. Standard techniques from information retrieval can be used to improve the quality of the results. Better implementations give users greater control over the search by allowing them to constrain the time, price, or other aspects of the auction. Many search implementations allow users to scour both open and recently closed auctions. Historical auction data provides valuable information about the recent market prices, and, if past bids are revealed, the bidding behavior of the market participants.

Payment and Escrow

C2C auction sites increasingly offer extra services to improve the security and efficiency of the marketplace. Two of the most common offerings are *payment* and *escrow* services. eBay affiliates with Escrow.com to offer escrow services and recently purchased PayPal, a popular payment service that had become widely accepted among eBay users. Amazon has integrated payments for auctions into its system-wide 1-Click ordering feature and guarantees that buyers will be reimbursed, up to specified limits, if they pay and fail to receive the product or the product is materially different than the seller described. (Information can be found on Amazon.com's Web site.) To support these efforts, the auction system must be integrated with the payment systems and provide tools to investigate disputes.

Some aspects of electronic payment are more straightforward on B2C and B2B sites, in part because there is less anonymity in these marketplaces, which enables standard payment methods to translate to the electronic domain more successfully. B2C sites generally accept credit cards as a form of payment, whereas B2B auction sites are more likely to invoke purchase orders. B2B sites are more likely to be involved in international exchanges, however, and monetary exchange rates become a factor.

Reputation Management

Reputation mechanisms are another common method of combating the fraud that often comes with the freedom afforded by anonymity. Reputation mechanisms collect feedback about users and allow them to accumulate a personal history. A positive history is an emblem of trustworthiness that can be used to assure potential trading partners that a transaction will go smoothly. A positive reputation is a valuable asset, and serves to make the site "sticky;" sellers with positive reputations are reluctant to take their business to a new site where they would need to build a new reputation from scratch. Typically, reputation mechanisms collect both a numerical evaluation and text comments. To avoid fraudulent inflation of reputations in a C2C site, users should only be allowed to comment on other users once per mutual transaction. In a B2B environment, this type of fraud is less common, and the designers may give the users more freedom to comment on one another.

Although reputation systems have become a de facto standard feature, their social effects are not well understood. Negative feedback is exceedingly uncommon (only 0.3% of all transactions were rated negatively; Resnick & Zeckhauser, 2001), and when a seller does receive negative feedback, the user can simply start a new account with a clean reputation. In June 2001, eBay took steps to curtail this behavior by tying seller accounts to credit card numbers. This provides a means to (somewhat) track identities, although new credit cards are easy to obtain. Nonetheless, negative reputations remain rare. Resnick and Zeckhauser suggested that there is an atmosphere of reciprocity and retaliation in the feedback system; a buyer is reluctant to submit negative feedback on the seller because the seller will retaliate. Despite these challenges, eBay's reputation system, and others like it, create a sense of community and a perception of trustworthiness and are considered crucial to the success of a C2C marketplace.

Censorship

The global reach, relative openness, and anonymity afforded by C2C auctions permits the posting of politically incorrect material (i.e., Nazi memorabilia, which is illegal to trade in some countries) or illegal products (i.e., illicit drugs or terrorist weapons). C2C sites have established policies and procedures to filter the postings to eliminated unwanted auctions. In addition, many specialty markets focus on collectable products that are targeted at children. Auction sites that expect to have children visitors must censor the product listings, feedback, newsgroups, and other content submitted to the site to be sure it is appropriate for minors, or create a gateway to prevent minors from accessing the adult-oriented content.

Fraud

Fraud, unfortunately, is common on C2C sites. In fact, the Internet Fraud Complaint Center (IFCC) reports that auction fraud accounts for 64% of all complaints, making it the single most common type of fraud on the Internet (IFCC Report, 2001). The most common type of auction fraud is nondelivery; the buyer sends the money, and the seller never sends the product. Many other problems exist as well, including *shill bidding, multiple bidding, feedback extortion,* and false reports of fraud (Freedman, 2000). Shill bids are bids placed by the seller or a collaborator to force legitimate buyers to pay more. On eBay, shills appear to be used to ferret out the maximum bid, information that wouldn't necessarily be revealed by the proxy bidding system. Shill detection software would be a complementary feature to auction software targeted at C2C sites. eBay has a large fraud division but does not publicize its techniques, and fraud detection is only recently being addressed in the academic literature (Shah et al., 2002).

Multiple bidding is a problem similar to shill bidding, except that it is perpetrated by a buyer to defraud a seller. The buyer uses one identity to bid just enough to win, and a second identity to bid very high. This high bid drives away competition, and eventually wins the auction. Then, however, the buyer defaults on the high bid, and the seller may exercise her option to offer the product to the second highest bidder at that bidder's (significantly lower) price. The seller may report the delinquent high bidder but may never know that the two identities can be traced to a single individual.

Feedback extortion occurs when a buyer demands a better deal after the auction under the threat of slamming the seller with negative feedback. Negative feedback can be so damaging to a seller's reputation that the seller will often capitulate to the blackmailer's demands. eBay is loath to get too involved in the resolution of these disputes because it opens the door for false reports of fraud for the sole purpose of discrediting competitors.

In B2B settings, the auction initiator will often specify the certified bidders and may actually create accounts on behalf of the bidders. This greatly reduces the opportunities for fraud. In B2B scenarios, bidders may be granted special privileges, as when they are preferred suppliers in a procurement situation. The initiator may maintain more control over the auction, and may have power to intercede if unforeseen circumstances occur, much the way the stock market will suspend trading on a stock with excessive volatility. Collusion among sellers is still possible, however, and extremely hard to detect.

Integration

It is common, especially in B2B or B2C settings, for the auction system to be connected to a variety of other software systems. Consider a company selling its products via an auction. When an auction closes, the company needs to update its inventory, compose a shipping request that is routed to the fulfillment department, transfer information about the sale to a customer relationship management system, and initiate a billing process. Although XML provides a framework for structured communication between components, integrating these systems still requires a significant amount of time and effort. Thus, administrators of a B2B or B2C site need access to the underlying software components, including the auction engine, the database, the Web server, and the e-mail server, so that they can integrate them with payment systems and other backend applications that may support billing, inventory management, credit authorization, and logistics.

Often, the auction system will be a component of a larger Web presence, and as such must be integrated into the entire Web site and "branded," raising issues of content management. For example, consider Amazon's auction offering. All of the auction pages share Amazon's common graphic and navigation elements (often referred to as "assets"), which are probably managed by a content management system. Thus, when integrating an auction engine into the larger site, a significant amount of work may go into marrying Web content generated by the auction engine with assets served up by the content management system. The customization tools provided with commercial auction systems vary from simply allowing custom graphics to providing scripting tools to create an entirely custom interface.

In addition to working out the details of data flow between the components being integrated, the developers must consider the effect of the integration on the system's responsiveness. In general, the system is only as fast as its slowest component. Thus, if the auction data is stored in a database shared by other mission critical operations, the combined load may overwhelm the database server. Similarly, if authenticating a bidder involves a query through a bottleneck unified authentication system, users may be unable to get through the front door of the auction site.

The integration issues on a C2C auction site are somewhat different. Although the inventory is managed by third parties, these sellers are becoming more technically sophisticated. In fact, many eBay sellers run quite sophisticated operations, and established businesses, both large and small, are using eBay as a direct sales channel. These types of sellers need the ability to connect the outcomes of the auction to their internal order fulfillment processes. Currently, the glue that holds the channel together is provided by small companies that specialize in managing the "auction" distribution channel. (The list of companies that provide this service includes Andale, AuctionHelper, Auctionworks, ChannelAdvisor, CommerceFlow, FairMarket, Zoovy, all of which can be found at their respective dot-com addresses.)

In both C2C and B2B environments, auction managers will need tools to handle the unexpected operational missteps that are bound to occur. For example, most auction sites have a policy to extend the duration of auctions that are adversely affected by system downtime. Great care must be taken when the system is brought back online. A naïve implementation of the system architecture will come back on line and notice that a large number of auctions are past due and will immediately clear them, informing the buyers and sellers of the trades. If the system administrators intend to extend the affected auctions, they would be forced to correct the auction and bid data manually and inform all of the affected buyers and sellers that the transactions have been nullified. In short, a naïve implementation will create an embarrassing mess. A better implementation will give the system administrators control over how auctions are handled when the system is brought back online, including allowing affected auctions to be extended without clearing, and notifying the auction initiators, and perhaps the bidders, of the changed schedule.

CONCLUSION

All auctions share the same core functionality: admit bids, generate information, and clear. The sequence and frequency with which the auction performs these three actions is specified by the auction rules. The rules also determine what types of bids are acceptable, what format the intermediate information takes, and how trade prices are computed. A configurable auction server will be able to implement a greater variety of auction types by mixing and matching the auction rules. This flexibility demands more complex software architecture, however.

In addition, auction systems vary widely in their ability to integrate with (or provide) a variety of complementary features that are common on Internet auctions. Whether selecting from existing auction engines or building one from scratch, the fit between the software's functionality and the needs of all three classes of users (bidders, auction initiators, and administrators) must be considered. The importance of different features will depend on the particular type of auction site being developed—a public C2C market has different requirements than a one-time procurement auction with certified participants.

Although there have been dramatic successes and failures in the short history of Internet auctions, they will continue to be a central fixture in the automation of commerce activities. The potential gains from the pervasion of dynamic pricing in the economy far outweigh the short-term costs associated with developing the technologies.

GLOSSARY

Activity rules Rules that constrain bidders' options based on their actions in previous rounds of the auction.

Auction A system for accepting bids from participants and computing a set of trades based on the offers according to a well-defined policy.

Auditable An auction system that can be inspected by an outside auditor to confirm that the auction's actions were in accordance with its policies and the bids received.

Beat-the-quote rule A rule that requires bidders' new offers to improve on their current offer with respect to the current quote information.

Bid An offer to buy or sell a product or service submitted to an auction.

Bid admission The act of certifying that a bid satisfies the bidding rules and has become an active bid in the system.

Bid schedule An offer that specifies a desire to buy or sell at more than one price point.

Bid submission time: The time at which the auction officially considers the bid submitted.

Clear The act of computing trades in accordance with the auction's policy and removing the associated bids from the set of active bids.

Combinatorial auction An auction that allows bidders to make offers on combinations of items.

Consistent An auction system that handles all bid, quote, and clear events in the order they ought to occur.

Content management system A Web publishing tool that facilitates the management and presentation of large amounts of content on the Web.

Double-sided auction An auction with multiple buyers and sellers.

Improve-your-bid A rule that requires bidders' new offers to improve on previous bids in a precise manner.

Matching function The policy the auction uses to compute trades from bids.

Multiple bidding A strategy involving a fake bid placed by a buyer with a secondary account to scare off other buyers; the high bidder then defaults on the bid and the real buyer wins with less competition.

Order book The list of current bids and the bidders who placed them.

Parallel architecture An auction architecture that, to some degree, allows multiple actions to be handled simultaneously.

Proxy bidding The use of a software program to place bids on behalf of the user; A proxy bidder, unlike an agent, is generally operated by the auction.

Quote The activity of computing and announcing information about the current state of the auction; the information computed and announced.

Reputation system A system that allows a community of traders to comment on each other's performance, thus enabling users to establish reputations within the community.

Reserve price In a single-seller auction, the minimum price at which the seller is willing to part with the item; in a single-buyer auction, the maximum price the buyer is willing to pay.

Sealed bid An auction in which no information is revealed until the auction closes and the winners are announced.

Sequential architecture An auction architecture that performs one action completely before beginning another.

Shill bid A fake bid placed by a seller, or a collaborator, to raise the price paid by real buyers.

Single-sided auction An auction with either a single buyer (and multiple sellers), or a single seller (and multiple buyers).

Trade A declaration by the auction that one bidder should exchange a particular quantity of the item with another bidder at a particular price; generally, the execution of the trade is outside the scope of control of the auction system.

Valid bid A bid that satisfies the bidding rules.

CROSS REFERENCES

See *Cybercrime and Cyberfraud; Internet Censorship; Legal, Social and Ethical Issues; Online Auctions; Web Content Management.*

REFERENCES

Ausubel, L. M., & Milgrom, P. R. (2002). Ascending auctions with package bidding. *Frontiers of Theoretical Economics, 1,* 1–42.

Clearwater, S. (1996). *Market-based control: A paradigm for distributed resource allocation.* River Edge, NJ: World Scientific.

Freedman, D. H. (2000, November 27). Sleaze bay. *Forbes ASAP.*

Goodwin, R., Akkiraju, R., & Wu, F. (2002). A decision-support system for quote generation. *Proceedings of Fourteenth Innovative Applications of Artificial Intelligence Conference* (pp. 830–837). Menlo Park, CA: AAAI Press and Cambridge, MA: MIT Press.

Internet Fraud Complaint Center. (2001, May). IFCC 2001 Internet fraud report [Technical report]. Retrieved September 2002, from http://www1.ifccfbi.gov/index.asp

Keenan, V. (2000, April). *Internet exchange 2000: B2X emerges as new industry to service exchange transactions* [Technical report]. San Francisco: Keenan Vision.

Kumar, M., & Feldman, S. I. (1998). Internet auctions. In *Proceedings of the Third USENIX Workshop on Electronic Commerce* (pp. 49–60). Boston, MA: USENIX.

McAfee, R. P., & McMillan, J. (1996). Analyzing the airwaves auction. *Journal of Economic Perspectives, 10,* 159–175.

Nisan, N. (2000). Bidding and allocation in combinatorial auctions. In *Proceedings of the Second ACM Conference on Electronic Commerce* (pp. 1–12). New York: ACM Press.

Ockenfels, A., & Roth, A. E. (2002). The timing of bids in internet auctions: Market design, bidder behavior, and artificial agents. *AI Magazine, 23,* 79–87.

O'Malley, K., & Kelly, T. (1998, September). An API for Internet auctions. *Dr. Dobb's Journal,* pp. 70–74.

Reich, B., & Shaul, I. B. (1998). A componentized architecture for dynamic electronic markets. *SIGMOD Record, 27,* 40–47.

Resnick, P., & Zeckhauser, R. (2001, February 5). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system [Technical report]. University of Michigan, Ann Arbor.

Rodriguez, J., Noriega, P., Sierra, C., & Padget, J. (1997, April). *FM96.5 a java-based electronic auction house.* Paper presented at the Second International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (PAAM '97), London, UK.

Rosenthal, R. (2002, August). 2002 competitive assessment of eRFx solution providers. Retrieved September 2002 from http://www.idc.com

Roth, A. E., & Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the Internet. *American Economic Review 92*(4), 1093–1103.

Rothkopf, M. H., Pekec, A., & Harstad, R. M. (1998). Computationally manageable combinatorial auctions. *Management Science, 44,* 1131–1147.

Shah, H. S., Joshi, N. R., & Wurman, P. R. (2002, July). *Mining for bidding strategies on eBay.* Presented at the Fourth WEBKDD Workshop: Web Mining for Usage Patterns and User Profiles. Edmonton, Canada.

Wurman, P. R., Walsh, W. E., & Wellman, M. P. (1998). Flexible double auctions for electronic commerce: Theory and implementation. *Decision Support Systems, 24,* 17–27.

Wurman, P. R., Wellman, P., & Walsh, W. E. (1998). The Michigan Internet AuctionBot: A configurable auction server for human and software agents. In *Proceedings of the Second International Conference on Autonomous Agents* (pp. 301–308). Minneapolis/St. Paul, MN: ACM Press.

Wurman, P. R., Wellman, M. P., Walsh, W. E., & O'Malley, K. A. (1999). Control architecture for a flexible Internet auction server. Paper presented at the *First IAC Workshop on Internet Based Negotiation Technologies,* Yorktown Heights, NY.

Wurman, P. R., Wellman, M. P., & Walsh, W. E. (2001). A parametrization of the auction design space. *Games and Economic Behavior, 35*, 304–338.

Online Banking and Beyond: Internet-Related Offerings from U.S. Banks

Siaw-Peng Wan, *Elmhurst College*

INTRODUCTION

The Internet has been in existence since the late 1960s; however, access then was only available to government agencies and academic institutions. It was used primarily for research purposes and no commercial traffic was allowed. All this changed in the mid-1990s, when the Internet was privatized, allowing the general public and commercial enterprises to freely access and use it.

The privatization of the Internet has led, in a short few years, to a dramatic transformation of the U.S. business landscape. New entities such as Internet pure plays and dot-com companies began to appear in massive numbers. As brick-and-mortar companies felt the pressure from these new online entities, many of them started establishing their presence on the Internet, thus transforming themselves into click-and-mortar companies. In other words, companies have both offline and online channels to reach and serve their customers.

Despite the well-publicized flameouts of a large number of Internet pure plays and dot-com companies in the late 1990s, the availability of the Internet to commercial enterprises has fundamentally transformed the way many traditional companies do business. There is no turning back for these companies. The Internet is now a well-integrated channel for many companies to reach and serve their customers, to trade with other companies, and to connect with their trading partners.

Banks are no exception when it comes to the influence of the Internet. Most banks view the Internet as an important channel in reaching their customers. Virtually all banks have established Web sites to provide information regarding their products and services to their customers. An increasing number of banks are also using the Internet to provide consumer, small business, and corporate banking services to their customers. This allows bank customers to conduct many of the banking transactions and activities on the Internet simply with browsers.

Today, banks are offering more than just banking services and products to their customers via their Web sites. They are attempting to serve all aspects of their customers' needs over the Internet. Most banks are turning their Web sites into online portals to compete with other online entities. These portals serve as gateways to information resource centers and financial supermarkets for bank customers, ranging from individuals to corporations. Customers can now use a bank's Web site as a central access point to a variety of information (both financial and nonfinancial) and products and services (both banking and nonbanking) (Fonseca, Hickman, & Marenzi, 2001). The Internet is also providing an opportunity for banks to branch out into offering other products and services. A number of large banks have recently offered small businesses and large corporations a variety of tools and instruments to facilitate their activities in the business-to-consumer and business-to-business e-commerce markets.

ORIGIN OF ONLINE BANKING AND OTHER INTERNET-RELATED OFFERINGS

The roots of online banking can be traced back to home banking, which was first offered by major banks in the mid-1980s when they attempted to bring the convenience of banking services to the homes of their customers (Gray, 1994; Serwre, 1995). Home banking was initially available in two forms, phone-based banking and PC-based banking. With phone-based home banking, bank customers conducted all their banking needs with touch-tone telephones. On the other hand, with PC-based home banking, banks were developing proprietary software packages that allowed their customers to dial into their systems directly to carry out their banking needs (Nixon & Dixon, 2000). In other words, bank customers were simply using their PCs as dumb terminals.

Both forms of home banking services offered bank customers very limited banking functions. They could only check their account balances, make payments to selected billers, and transfer funds among different accounts within the bank. However, home banking was not widely available among banks. Only 340 U.S. banks, out of approximately 12,000, were offering home banking services by the late 1980s (Gart, 1992). One of the reasons for this limited availability was that home banking was accessible only through proprietary systems that required a huge technology investment, which only bigger banks could afford. Another reason was that the demand for PC-based home banking services had been very low because the number of households with PCs was very low and therefore most banks could not justify the investments needed to offer such services (Nixon & Dixon, 2000).

In the early 1990s, PC-based home banking found a new life as commercial services such as Prodigy began to offer a whole host of services, which included electronic banking services, to their subscribers. About 10 banks began offering such banking services on Prodigy in 1991. However, the growth of electronic banking services was slow. There were only 16 banks offering such services on Prodigy by 1996 (Radigan, 1996). However, this did not discourage other commercial services from offering similar electronic banking services. America Online (AOL) began offering electronic banking services to its subscribers in late 1995. Wells Fargo and Bank of America were among the 16 banks that had joined AOL's Banking Center by the end of 1986 (Yamada, 1995).

While commercial services were bringing electronic banking to their subscribers in the early 1990s, personal financial management (PFM) software packages such as Microsoft Money and Intuit Quicken were also incorporating online banking features into their latest versions. Both companies were successful in contracting with a number of banks to offer electronic banking services to their customers who used either MS Money or Quicken to manage their finances. As in the case with earlier PC-based home banking services offered directly by the banks, electronic banking services available with MS Money and Quicken were limited to checking account balances, transferring funds, and paying bills. By the end of 1993, customers of Michigan National Bank, U.S. National Bank of Oregon, and First National Bank of Chicago could conduct electronic banking using MS Money, and customers of Bank One, First National Bank of Omaha, Wells Fargo, and Meridian Bank could do so using Quicken (Arend, 1994).

PC-based home banking with commercial services and PFM software packages allow bank customers to conduct their banking transactions any time but not anywhere. Anytime-and-anywhere banking did not materialize until banking services began to appear on the Internet, allowing bank customers simply to access them with browsers. However, it did take a few years before banking services became commonly available over the Internet. When the Internet first became available to commercial enterprises in the mid-1990s, most of them, including the banks, viewed it as just a new channel to distribute information to their customers. First Union Bank was one of the first banks to establish a presence on the Internet. In 1995, it established the First Access Network, a customer-service-oriented Web site providing information on the bank, its product and services, and other areas (Online for cyberbanking, 1995). In other words, banks were simply putting their existing brochures on their Web sites.

Some banks, however, did see the Internet as a new channel to offer banking services to their customers. In 1995, the Office of Thrift Supervision (OTS) approved Cardinal Bancshares Inc.'s application to provide limited banking services over the Internet. As a result, Security First Network Bank, the first Internet-based bank, was established (Messmer, 1995; Hoffman & Kim, 1995). However, many banks were still skeptical of the Internet's potential as a channel to offer banking products and services to their customers. According to Online Banking Report, there were only 24 banks offering account balances and transaction processing to their customers over the Internet in early 1997 (Wilson, 1997). However, as online competition for customers from other nonbank entities such as brokerage firms and financial Web sites intensified, many banks began to step up their efforts in offering online banking services over the Internet. The number of banks offering Internet banking services increased to 200 by 1998 and grew to 819 by 2000 (Orr, 1998; Online Banking Report true Web banks & credit unions database, 2001).

Not only has the number of banks offering Internet-based (or Web-based) banking services increased over the past several years, but the types of products and services offered by banks on their Web sites have also expanded. Banks are now offering products and services that go beyond the ones traditionally offered by banks, often through partnerships with other Web-based companies and providers. In addition, as security measures and Internet technology advance, banks are also expanding their Web-based offerings to individual customers to include small businesses and corporate clients.

SOURCES OF INFORMATION

The Internet is well suited for delivering information because (1) it can be accessed anywhere and anytime, (2) it can be updated easily and instantaneously, (3) the cost of delivering the information is much lower than in other media, (4) much more information can be delivered compared to other channels, and (5) the information can be multimedia and interactive.

Banks, just like other commercial enterprises, are taking advantage of the Internet's qualities to deliver a variety of information to various target audiences and to provide interactive and unique experiences to those accessing these information sources. The majority of the information available on a bank's Web site is designed for its current customers and its potential customers. The information can be grouped into four general categories: (a) marketing information, (b) educational materials, (c) financial information, and (d) nonfinancial information. All of these sources of information are freely available to anyone, but some banks do require visitors to register before gaining access to certain information.

Marketing Information

Most marketing information focuses on introducing the various products and services a bank has to offer to its customers and on helping its customers locate its physical branches and ATM machines (if there are any). In general, this information is grouped according to the various customer groups the bank serves, such as individuals, small businesses, and corporate clients. Other than providing descriptions of their products and services, some banks have also chosen to help their customers gain a better understanding of their products and services by offering Frequently Asked Questions (FAQ) sections. For example, if a LaSalle Bank customer is interested in applying for a home equity loan, he or she can find out more about the procedure by clicking on the *How do I apply for a home equity loan?* link in the FAQ section.

Many banks are offering this information in a static manner. In other words, customers have to plow through pages and pages of information in order to find the information they need. Some banks are now taking advantage of the inherently interactive nature of the Internet environment to provide their customers with a more dynamic and pleasant experience when accessing such information. For example, instead of simply providing written information on its online banking services, Firstar Bank offers its customers an online demonstration of the services' various features. Its demonstration is made up of a series of Web pages with illustrated screen captures to help its customers understand what types of tasks they can perform online and how to go about performing them. Customers simply need to click on the different links to access the appropriate pages to learn about the various features. Net Bank, on the other hand, chose to provide a Macromedia Flash-based demonstration to help its customers learn more about its online banking features. Therefore, Net Bank's customers do not need to click on any links to advance to the next features. All they need to do is sit back and enjoy the presentation, which incorporates animation, sound, video, and other multimedia elements. These two approaches are the two most common ones adopted by banks when demonstrating their products and services. Although they do incorporate various multimedia features into their demonstrations, they are not hands-on. In other words, bank customers do not have the opportunity to test drive the various features. Some banks are striving to offer their customers such opportunities. When accessing the online banking demonstration, customers of Westport National Bank can try out the various features by modifying the given information and see the changes reflected in the appropriate screens immediately. For example, if a customer transfers $500 from the checking account to the savings account, the demonstration will present him with a screen to confirm the transaction. Once it has been confirmed, the presentation will show how the transaction affects the various bank account balances.

Often though, it is still difficult for bank customers to find answers to their questions despite the banks' efforts in offering an extensive amount of information on their Web sites. The bank customers can send e-mail messages to the bank's customer service department to have their questions answered. However, it often takes some time before they receive responses from the bank. Some banks have been pushing to offer their customers instantaneous responses to their questions by offering them an opportunity to interact with customer service representatives via a chat room environment. For example, Harris Bank's customers can click on the *Push to Talk* button to connect with an online customer service representative. The chat room environment allows customers to carry out real-time "conversations" with customer service representatives.

Educational Materials

With the understanding that an educated customer is a good customer, it is no longer sufficient for banks simply to provide information on their products and services. They need to help their customers understand why they need these products and services. Most banks are now offering a variety of educational materials to their customers, focusing on individuals and small businesses, to help educate them on a wide range of topics.

These educational materials are often categorized into different sections to help the intended target audiences easily locate the information they need. Different banks have chosen to categorize the materials in different manners. For example, when it comes to educational materials for individuals, some banks are grouping the materials into general topics such as education, home, and retirement planning, investing; some banks are grouping them based on activities such as planning for emergencies, getting married, and supporting a growing family; and some banks are grouping them based on demographic groups such as teenagers and women.

Within these various categories, customers can find a number of articles and tutorials related to a specific topic and a glossary that provides an explanation for the many common terms used in the materials. These articles and tutorials cover not only financial-related materials but also non-financial-related materials. For example, customers accessing Citibank's educational materials on college planning can find not only articles on how to fund a college education, but also articles on how to plan summer college visits and how to prepare for college admissions.

In addition to written materials, many banks are also offering a number of interactive calculators and planning tools to help their customers make certain financial decisions, such as decisions related to financing a vehicle by determining the monthly payments for their auto loans. The calculators prompt the users to provide certain key information and then provide a detailed explanation of the outcomes. To help the customers pick out the appropriate calculators, banks generally word them in an easy-to-understand manner, often in the form of a question. For example, when Bank of America's customers are trying to choose a *Home Purchase* calculator, they can pick from ones with titles such as *Which is better: fixed or adjustable?* and *How much should I put down for a new home?*

Many calculators are created using JavaScript and are best suited for simple calculations. Some banks are choosing to offer Java applet calculators that can handle more complicated calculations. The *Visual Planners* from Bank One are Java applets that provide detailed visual summaries of the outcomes. In addition, customers can also save the inputs and results and return to them at a later time.

The JavaScript or Java applet calculators generally help the users deal with specific issues such as choosing between buying and leasing a car. Some banks, such as Fifth Third Bank, are beginning to offer more comprehensive planning tools that help their customers deal with more complex issues such as retirement planning. Through its *Retirement Center*, the bank is providing a planning tool that offers a step-by-step approach that helps the customers determine their retirement needs and plans from beginning to end.

Financial Information

Banks, just like any other entities with an Internet presence, want to keep visitors at their sites as long as possible. To prevent their customers from visiting other financial Web sites for basic financial information, many banks are now offering similar information sources at the banks' Web sites. Customers can now access financial information such as delayed stock quotes, charts, current market activity, and economic and company news. These information sources are often located on a bank's home page so that its customers can easily access the information.

Nonfinancial Information

In addition to offering basic financial information on their Web sites, some banks are also offering nonfinancial information such as general news feeds and local weather forecasts. This is especially the case with some smaller banks that are attempting to create closer ties to the local communities by acting as gateways to a variety of local information. For example, City National Bank of Florida is providing links to a number of third-party Web sites that offer information on various local activities such as South Florida arts, home and garden events, and local sports. Citizens Financial Bank's Web site also provide links to local organizations such as schools, YMCAs, and hospitals.

Some banks are even offering the local community a venue on the banks' Web sites to post information regarding certain activities and events. Anchor Savings Bank has an online *Community Event Form*, which allows local organizations to submit information on charitable or not-for-profit events to be posted on the bank's *Community Calendar*.

ONLINE BANKING PRODUCTS AND SERVICES FOR INDIVIDUALS

Because e-commerce (more specifically Internet commerce) first gained momentum in retail markets that catered to individual consumers in the mid 1990s, it is logical for banks to target their online banking products and services to individual bank customers. There were more than 800 banks in the U.S. offering some form of banking products and services over the Internet by 2000. Many of these banks present very similar features for their offerings in terms of the types of activities and transactions individual bank customers can perform via the banks' Web sites.

Opening an Account

Most banks with an Internet presence allow their customers to initiate the account application process over the Internet. Some banks provide account application forms on their Web sites, requiring their customers to download and print the forms, fill them out, and then either drop them off at a local branch or mail or fax them to the bank. On the other hand, some banks will allow their customers to complete and submit their applications online.

Even though it is convenient for bank customers to be able to submit their account applications online, it is not without its share of frustrations. On many occasions, bank customers find that they did not have all the relevant information they needed after they are halfway through with the online application process. This usually means that they will have to start the application process again if they cannot easily locate the missing information. A number of banks have taken measures to ease the process associated with collecting information for the applications. These banks allow their customers to save their incomplete applications for a certain period of time, allowing them to return during that period to enter any missing information and to modify existing information. Customers at Chase can save their incomplete applications for up to 30 days.

Even with all the information provided on the various bank products, many bank customers do run into problems picking the right bank products to best fit their needs. Some banks, such as LaSalle Bank, are offering interactive tools to help customers find the appropriate products. These tools generally ask the customer a series of questions to form the foundation of their recommendations.

Accessing Account Information

Once bank customers activate online banking services with their banks, they can access a variety of information related to their accounts via the banks' Web sites. They can view the account balances on their checking, savings, CDs, mortgage, IRA, line of credit, credit card, and other accounts. For example, TCF Bank's customers can access online a variety of information on their mortgage accounts, such as their mortgage payment status, their mortgage and escrow balances, and the interest and tax paid information for the current and previous year.

Customers can also view their account transaction histories online. How far back in time bank customers can go depends on the bank. Bank of America's customers can view their account transactions for the past 70 days, whereas Fleet Bank's customers can do so for the past 180 days. For example, they can easily determine whether a particular check has cleared simply by clicking on the link associated with that particular check. However, most banks do not provide information such as the payee of the check as part of the account transaction information unless the customers request the information. To simplify the lives of their customers, some banks are now letting their customers view the front and back images of cleared checks. This makes it a lot easier for customers to determine who the payee is on each cleared check without having to look it up in their checkbooks.

In order to provide their customers with timely information on various accounts, some banks are beginning to offer their customers the opportunity to view their bank statements online. Net Bank provides electronic bank statements in both HTML and Adobe Acrobat formats that its customers can view online or save to their hard drives. Similarly, customers of First Tennessee Bank can also review and download monthly electronic bank statements for their accounts for the past 16 months.

In addition to offering access to information related to the various accounts customers have with the banks, many banks have recently begun to provide them with centralized locations for information on other online accounts not related to the banks. In other words, banks are offering their customers a single access point for information from various online accounts using account aggregation. Bank customers can set up account aggregation to access information on their online banking, brokerage, and credit card accounts. In addition, they can also access their e-mail, review their frequent flier miles, and receive news feeds from various third-party sources. Currently, banks that are offering account aggregation are doing so by partnering with one of two major account aggregators, Yodlee.com and My Accounts from Quicken.

Some banks are also sending out e-mail messages to alert their customers when certain customer-predesignated events have taken place. For example, Fleet Bank will send out an e-mail message if a customer's account balance is either above or below a certain specified amount or if a bill payment has not been cleared after a predetermined amount of time. In addition to account-related activities, Zions Bank's *My Alerts* feature will also send out e-mail notification when the CD or money market rates reach a certain pre-specified level and as personal reminders such as for anniversaries, birthdays, and meetings.

Initiating Transactions

When bank customers sign up for online banking services with their banks, they can choose which accounts they would like to be linked online. Once the linkages have been set, customers can easily transfer payments among these accounts if permitted. For example, they can easily transfer funds between their checking and savings accounts. And if the customers have credit cards issued by their banks, they can easily pay the credit card balances simply by transferring money from their checking or savings accounts to their credit card accounts.

In addition, customers can also choose to sign up for an online bill payment feature if their banks offer it. Once this feature has been activated, bank customers can schedule future payments for up to a year in advance or set up recurring payments such as mortgage payments. Online bill payment can be used to pay for almost anything regardless of whether the payees accept electronic payment or not. If they do not accept electronic payment, the banks will simply cut the checks and mail them out. The only exceptions to online bill payments are court-ordered payments, payments to purchase securities, tax payments, and payments outside the U.S. With online bill payment, customers can easily track their expenses by payee, date, account, or category.

Some banks are also bundling online bill presentation service with their online bill payment service. If bank customers choose to activate the online bill presentation feature, they will be able to receive electronic bills rather than paper bills from selected billers. The bank customers can then pay these bills using the banks' online bill payment features, which will transfer funds from the customers' accounts to the billers' accounts.

Online Self-Serve Customer Service

The Internet offers many commercial enterprises, including banks, the opportunity to reduce many of the costs associated with customer service by allowing them to provide many traditional customer services via their Web sites. Many banks have now set up online automated services that enable their customers to perform certain functions so that they no longer have to go to the banks or call up a customer service representative. They can now perform online a number of account-related activities such as reordering checks, ordering paper copies of checks and bank statements, or requesting a stop payment on a check. In addition, customers can also change certain personal information such as their addresses, PIN numbers, and other information easily online without having to fill out a number of forms at a local branch.

Other Features

Many banks are also branching out online to provide many products and services to their customers that are not account-related. Some of these services have traditionally been performed at local branches. For example, a number of banks are now offering their customers the opportunity to order traveler's checks and foreign currency for their travels directly through the banks' Web sites. Wells Fargo, Bank of America, and Net Bank are a few of the banks that are currently providing such services. Once the banks transfer the appropriate funds from the customers' accounts, the traveler's checks and foreign currency are then delivered to the customers' doorsteps.

Some banks are also offering services that are not traditionally offered by banks. For example, some banks are beginning to provide certain tax assistance to their customers. LaSalle Bank is providing advice on tax issues to its customers over the Internet through its *Tax Hotline* service. In addition, individuals are now able to prepare and file their tax returns online at a number of banks using the provided Quicken Turbo Tax for the Web. Bank One, Busey Bank, People's Bank, and Webster Bank are a few of the banks that are currently offering such service. Another new service that is currently provided by a handful of banks is the ability for customers to send and receive money over the Internet. Citibank and Bank One have introduced such services, named *c2it* and *eMoneyMail*, respectively. These services enable the banks' customers to send money over the Internet to pay for online purchases and to donate to charities without having to mail out paper checks or to provide credit card information online.

ONLINE BANKING PRODUCTS AND SERVICES FOR SMALL BUSINESSES

The commercialization of the Internet created a new channel for many small businesses to access a variety of banking and business solutions provided by banks. Traditionally, it had been the practice of many banks to tailor their offerings to big corporate clients and to individual customers. To meet their business needs, small businesses were forced to choose between the expensive services for corporate clients and the oversimplified offerings for individual customers. Banks could repackage many of these services to meet the needs of small businesses but many banks could not dedicate the time and the attention to service this particular customer group. The Internet provided banks the opportunity to offer many of the repackaged services via their Web sites without expending a great deal of resources (Bers, 1996).

Most online banking services available to small businesses are very similar to the ones available to individual customers. Through the banks' Web sites, small businesses can perform a number of activities such as accessing account balance information, reviewing bank statements, viewing images of cleared checks, transferring funds between accounts, paying bills online, filing federal taxes, and ordering foreign currency and traveler's checks.

The major difference between online account-related services for individual customers and online account-related services for small businesses is that small businesses have the ability to grant and limit access to the various features for different employees. For example, First Tennessee Bank allows small businesses using its online account services to designate individuals as super-user. The super-user will then determine the types of features other employees can use and grant them the appropriate levels of access.

Other than the banking services mentioned above, many banks are also providing a number of web-based products and services to help small businesses better manager their operations. In addition, many banks are also partnering with third-party providers to offer other products and services. As a result, banks' Web sites have become online hubs for small businesses to access a variety of products and services from different sources. Payroll services, accounting system, human resource tools, and educational workshops are four of such products and services commonly available via the banks' Web sites.

Online Payroll Services

A number of banks are bringing the convenience of anywhere and anytime payroll services to small businesses by offering them through the banks' Web sites. Small businesses can now easily make any changes to the payroll information anytime they choose over the Internet. In addition, they can also review the payroll amounts for each employee online before approving and releasing the payroll for processing. Once everything is processed, these online payroll services will also process payroll taxes for the small businesses as long as their accounts are sufficiently funded.

Banks that offer online payroll services to their small business customers tend to do so through partnerships with third-party providers. For example, JP Morgan Chase Bank offers *Powerpay* through its partnership with Ceridian and First Tennessee Bank offers *POWERPayroll* through its partnership with PayMaxx.

Internet-Based Accounting

Most small businesses rely on software packages such as Intuit Quickbook to manage their accounting ledgers and processes. These packages offer the convenience of *anytime* accounting for the small businesses but not *anywhere* accounting. Some banks are now offering small businesses an alternative to such accounting packages by providing a web-based accounting program on their Web sites. For example, small business customers of Well Fargo can access the Oracle Small Business Accounting program on the bank's Web site.

These Web-based accounting programs are offering small businesses anytime-and-anywhere accounting. They can perform a number of activities with just a browser and an Internet connection. In addition to managing all their accounting ledgers and processes online, small businesses can also print checks, create their invoices, track their project expenses, and create their budgets online with the program.

Human Resources Tool

Understanding the complexities many small businesses encounter when dealing with human resource issues compliance requirements, some banks are striving to help their small business customers handle such issues by offering a variety of human resource tools on their Web sites.

First Tennessee Bank is one bank that is offering a number of online human resource tools to its customers, through its partnership with Firstdoor.com. Small business customers of First Tennessee Bank have access to three different modules offered by Firstdoor.com. One of the modules (RESEARCH FIRSTdoor) offers small businesses the tool to search an extensive library of HR materials, and another module (TRAIN with FIRSTdoor) provides various online training courses and tools to employees of the small businesses. If the small businesses have any questions regarding HR issues, instead of contacting their attorneys or accountants, they can receive customized answers by submitting the questions using the ASK FIRSTdoor module.

Business Workshops

In order for small businesses to succeed, they need to be familiar with the marketing, legal, financial, and other tools that are at their disposal. Some banks are stepping in to help their small business customers learn more about the various tools and to incorporate the tools into their businesses.

Wells Fargo and U.S. Bank are two banks that are offering, through their partnership with DigitalWork, a number of workshops tailored for small businesses. The objectives of these workshops are to help familiarize small businesses with various business tools and to help them put the tools to work. Small businesses can choose from areas such as Web site development, public relations, online advertising, and market research. For example, small businesses that are interested in online advertising can learn how to send promotions and newsletters online, how to submit their Web sites to search engines, and how to send direct e-mail. If the small businesses are interested in putting the tools to work, the banks do offer simple-to-follow guidelines for some of the tools and also links to third-party providers that will perform the tasks for the small businesses.

ONLINE BANKING PRODUCTS AND SERVICES FOR CORPORATIONS

Many banking services for corporate clients have long been available online via dedicated network connections such as financial EDI (electronic data interchange) and ACH (automated clearinghouse) networks. It was not until the late 1990s when security measures and Internet technology made dramatic advancements that banks began offering Web-based interfaces for many of these services. Corporate clients can now access such online banking services anywhere and any time with a browser and an Internet connection. In addition, as business-to-consumer (B-to-C) and business-to-business (B-to-B) e-commerce flourished in the late 1990s, many banks began to offer a number of tools and services to facilitate the online activities.

Treasury Management

Online treasury management is the most commonly available online banking service for corporations among banks. The basic features of the service allow a corporation to exchange information and documents with a bank and to initiate a variety of payments over the Internet. These features generally include ACH origination and direct deposit, controlled disbursement of funds, depository services, wholesale lockbox services, and information reporting (Wilson, 2001).

Just as in the case of small businesses, corporations can check their account activities, move funds between accounts, set up recurring payments, and initiate payments and transfers. The only difference from small businesses is that corporations are dealing with more complex transactions. For example, when it comes to initiating payments and transfers, corporate clients can conduct both domestic and international wire transfers, using the appropriate online treasury management feature. In addition, they can also initiate ACH transactions such as direct deposit of payroll and federal and state tax payments.

Most online treasury management services also provide image archive services to corporate clients. For example, U.S. Bank's *Image Check* service enables its clients to retrieve and review sharp, detailed images of any paid check on the Internet the day after it clears. LaSalle Bank's image archive service through its *Cash Pro Web* provides its clients a 7-year archive of check images. In addition to check images, some banks are also offering online archive services for other documents. The *Wholesale Lockbox Imaging* service from Fifth Third Bank provides companies online access to payment files such as remittance and correspondence through a secure Web site.

International Trade and Foreign Exchange

Many banks are enabling corporate clients that have international business dealings to conduct many of their foreign exchange activities online, such as the initiation of payments and the acceptance of payments to settle international trades. LaSalle Bank currently offers *Max Trad* (formerly *Trade Station Online*), an online letter of credit and international receivable service for corporations. Importers can create and send applications for letters of credit online and exporters can receive letters of credit over the Internet.

Another activity that corporations can conduct online is the management of their international payments. For example, corporate customers of Comerica Bank can conduct, via the bank's eFX service, a number of activities such as initiating and executing payments in different foreign currencies and receiving real-time foreign currency quotes. In addition, some banks are also offering foreign exchange trading services where corporations can hedge their foreign currency risks online using a variety of foreign exchange instruments such as spot, forward, swap, and option contracts. Two of the banks offering foreign exchange services are U.S. Bank with its *Foreign Exchange Web* and Fifth Third Bank with its *FX Internet Trading*.

Electronic Bill Presentation and Payment

The latest development that has caught the attention of a lot of corporations is the ability to present bills and invoices to their customers and trading partners and to accept payments over the Internet. Electronic bill presentation and payment (EBPP) can be divided into two different categories depending on the parties the corporations are dealing with: business-to-business EBPP and business-to-consumer EBPP.

B2B EBPP enables a corporation to invoice its corporate customers and process the payments over the Internet. Citibank is one of the banks that is offering its corporate clients B2B EBPP service through its *B2B e-Billing* feature. For corporations that have signed up for this feature, Citibank will extract the necessary information directly from the corporations' accounts receivables files, create the invoices, and present them formatted on a secured Web site. Citibank will then send out e-mail notifications to the corporations' trading partners, alerting them that the invoices are ready for them. The trading partners will then access the invoices by simply using a browser. They can then review the invoices, process them, and initiate payment if there are no disputes regarding the invoices. If there is any dispute on a particular invoice, the trading partners will have the ability to adjust it. The corporations and their trading partners can export the billing and payment information into their accounts receivables and accounts payable systems, respectively.

On the other hand, B2C EBPP is designed for corporations to bill and process payments from regular customers. For example, a corporation that has signed up for Citibank's *B2C e-Billing* feature will transmit its billing files to Citibank when it is ready to bill its customers. Once Citibank receives the information, it will translate the information, format the bills, and send them to each customer's designated site, which is prearranged with the corporation by the customer. The designated site could be the corporation's own Web site, the customer's bank's Web site, or a consumer bill consolidator's Web site. The customers will then access the designated sites to access the electronic bills, review them, and schedule their payments. The customers' payment authorizations will then result in instructions to electronically debit their accounts and credit the corporation's account. Citibank will also transmit the file to the corporation's accounts receivable system for updating.

CROSS-SELLING OTHER BANKING PRODUCTS

A bank's Web site is a good channel to cross-sell its other products such as loans and mortgages, insurance, and online brokerage services. In general, the offerings of these products are available either through the bank's affiliates or through a partnership with a third-party provider.

Loans and Mortgages

Online lending is defined as the ability to completely process a loan, from application to closing, using Internet banking channels (Treadwell, 2001). It generally involves four steps: (1) taking the application over the Internet, (2) underwriting the loan and providing an instant credit decision, (3) delivering relevant documents (if necessary), and (4) funding the loan (Courter, 2001).

Due to the complexity involved in some loan applications, especially with mortgage loans, many banks have not even taken the initial step in offering online lending through their Web sites, namely accepting loan applications online. These banks are merely providing loan application forms on their Web sites for their customers to download and then print out. Their customers still need to fill them out and then either drop them off at a branch or mail them in. Some banks will accept online loan applications for certain loans but require hard copy applications for other loans. For example, People's Bank will accept online applications for personal loans but not for home equity, credit line, or auto loans or personal credit lines.

As for the banks that are accepting online loan applications, some of them are allowing their customers to save their incomplete applications for a certain period of time. For example, Charter One and Washington Mutual allow their customers to save their online applications and calculations for up to 30 days and 60 days, respectively. Customers can continue to enter new information and modify existing information on their saved application during that period.

Once the applications have been submitted online, some banks will provide their customers an instant online credit decision regarding their loan status. On the other hand, some banks will take longer to process the loan applications, requiring the customers to return later to check on their application status. Even though some banks are able to provide instant approval for their loan applications online, most of them are not able to fund the loans immediately because of the complex paperwork and signatures required for many types of loans.

To ease the frustration associated with loan applications, some banks are taking measures to help their customers determine if they are qualified for the loans they want to apply for before they begin the application process. Banks such as Washington Mutual are offering online prequalification tools on their Web sites to help customers determine if they might qualify for the loans before filling out the applications. If the customers do qualify for the loans and they decide to apply for the loans, the prequalification tool will transfer all the information provided and prefill it on the online application forms, thus saving some time for customers.

A few banks are also offering other services associated with loan products. For example, First Union has a Relocation Center that provides articles and resources related to areas such as buying and selling a house, looking for a new home, and moving across the country. Bank One, on the other hand, is providing its customers a link to MovingStation.com.

Insurance Products

Another bank-related product that is increasingly available through banks' Web sites is insurance. The majority of banks that are taking insurance applications online are partnering with third-party insurance providers. Some of these partnerships offer a very extensive list of insurance choices, whereas others offer only very limited choices. Net Bank, through its partnership with Insurance.com, is offering a wide variety of insurance products to its customers. Customers can apply online for auto, life, home, long-term care, health, boat, motorcycle and RV/ATV, travel, and pet insurance. On the other hand, Comerica Bank is offering only health insurance through its partnership with eHealthInsurance.com.

Many other banks have chosen not to accept any online insurance applications, but instead have chosen to offer only online quotes. Standard Federal Bank, HomeStreet Bank, Hibernia National Bank, and United California Bank are a few of the banks that currently only offer online insurance quotes.

Some banks are also offering other insurance-related services via their Web sites. For example, to expedite the insurance claim process, BNC National Bank and Washington Mutual are two of the banks that are allowing their customers to file their claim submission through their Web sites.

Online Brokerage Services

As in the case with their online banking services for individuals, banks' online brokerage services offer many of the same features for brokerage accounts. Investors can either apply for a brokerage account online or download and print out a paper copy of the application forms. Once they have set up their online brokerage accounts, customers can begin placing a variety of orders (e.g., market, limit, and stop orders) for different types of financial assets (e.g. stocks, bonds, and mutual funds). They can also view their account balances and investment activities, such as their order status. In addition, they can view their trading histories for the entire account or for a particular financial asset. If the investors have banking and brokerage accounts with the same bank, they can transfer money between the accounts.

Investors with online brokerage accounts also have access to a variety of online resources and tools. They can receive real-time quotes, charts, market data, and different news feeds on the economy, a specific industry, and specific companies. In addition, they also have access to a wealth of research materials from third-party providers such as Market Guide, Valueline.com, Standard & Poor's, Thomson Investors Network, Briefing.com, and others.

Some banks are also providing a number of interactive tools to help investors pick the most appropriate assets for their portfolios. Many online brokerage services offer screeners, which are interactive tools that help investors find the stocks and mutual funds that meet the customers' criteria. Some brokerage services also provide portfolio trackers that help investors monitor the performances of a number of mock portfolios. For example, Citibank's portfolio tracker service offers investors the ability to create up to 30 portfolios, each containing up to 40 financial assets.

Electronic Safety Deposit Boxes

Several banks are beginning to offer their customers the ability to store important documents electronically on their Web sites. Fleet Bank, Zions Bank, and JP Morgan Chase Bank have introduced file TRUST, z-vault, and I-vault, respectively. These are electronic safety deposit boxes for their customers to load, archive, and retrieve various documents and records such as contracts, deeds, titles, wills, and other documents via the Internet. Customers can choose to selectively provide online access to these documents to their partners, lawyers, financial planners, accountants, and others.

E-COMMERCE SOLUTIONS AND TOOLS

Many of the larger banks have recently jumped on the ever-growing e-commerce bandwagon and have begun offering a number of front-end and back-end e-commerce solutions and tools to small businesses and corporations. Some of these tools focus on facilitating e-commerce activities in B2C online retail markets and some focus on facilitating e-commerce activities in B2B markets and electronic exchanges.

B2C E-commerce Solutions and Tools

A number of banks are assisting small businesses to establish online storefronts by helping them build their Web sites from scratch, by providing the necessary tools to conduct online selling, and by hosting online storefronts. On the front end, banks can help online merchants come up with designs for their online storefronts and create a number of components to help shoppers find the products they need and to help Web merchants take orders on their Web sites. For example, Chase's *eWEBuilder* service helps Web merchants establish electronic catalogs so the shoppers can easily search for products. In addition, it also creates shopping carts, tailored to the Web merchants' needs, which allow shoppers to select a number of items to be purchased during their visits. It also programs the checkout screens to tally the shoppers' purchases and collect payment and shipping information from the shoppers. On the back end, banks can help online merchants process payments and transfer them to the appropriate accounts. Chase's *eWEBuilder* service offers Web merchants a payment gateway, which is a secure server connection that encrypts the payment information collected at the checkout screens and sends it out for authorization and payment processing.

Chase, together with Wells Fargo and Zions Bank, is one of the banks that provide all the front-end and back-end services associated with establishing and running online storefronts completely in-house. However, some banks are only offering some of the services in-house and are offering the other services through a third-party provider. For example, PNC Bank is providing the Web site creation and hosting services through its partner Gateway and the bank focuses on providing credit card processing support.

Instead of providing a complete package to help Web merchants establish online storefronts, some banks have chosen to focus solely on offering a secure way for Web merchants to accept credit card and other payments from their shoppers and on processing the payments. In addition, some banks will also help Web merchants provide some follow-up customer service to their customers. For example, LaSalle Bank will send out e-mail messages to the Web merchants' customers, notifying them of the their order and shipping statuses.

B2B E-commerce Solutions and Tools

In recent years, electronic exchanges have flourished, which enables companies to broaden the pool of companies that they trade with. However, when companies are trading online through an electronic exchange, they are often dealing with companies that they have never traded with before. As a result, companies are assuming a great deal of counterparty risk. In other words, buyers are uncertain if the sellers will ship the products and sellers are uncertain if buyers will pay after they have received the shipments.

Many banks have been stepping in to reduce the uncertainties associated with trading on an electronic exchange. For example, Citibank has offered companies trading on electronic exchanges the *CitiConnect Escrow Service*, which serves as a trusted third party to the transaction. The sellers will only ship their products after the bank has received the payments from the sellers and the bank will only release the buyers' payment to the sellers after the buyers have received and accepted the shipments.

In addition, to ensure that companies are confident in the true identities of the companies they are trading with on an electronic exchange, some of the banks are issuing digital certificates, basically electronic tools, that help verify the identities of the companies. One such online identity service is Identrus from LaSalle Bank.

ONLINE SHOPPING AND E-PROCUREMENT

The Internet has provided a new channel for both individuals and companies to shop for products and services. However, it is sometimes difficult for them to find the products and services they need. Many banks have stepped in to play the matchmaker role, helping individuals and companies with their shopping and procurement efforts. A number of banks are providing gateways via their Web sites to different online merchants to help individuals and companies search for products and services. In addition, a number of banks are beginning to offer online forums in the form of electronic marketplaces for companies to trade with one another.

Online Shopping for Individuals

There are a number of ways for banks to offer an online shopping experience to individuals: (1) partner with a few Web merchants to bring very specific products and services to their customers, (2) offer their own shopping malls that house a number of Web merchants, and (3) partner with existing shopping malls.

Partner with a Few Web Merchants

A common way for some of the smaller banks to bring online shopping opportunities to their customers is by partnering with a selected few Web merchants. Bank of Smithtown, Carrollton Bank, and Anchor Savings Bank are three such banks that have partnered with Trip.com to bring travel-related services and resources to their customers. Through their Web sites, the banks' customers are able to purchase plane tickets, to book hotel rooms, and to reserve rental cars. In addition, customers can also access destination guides, street and airport maps, driving directions, weather forecasts for their destinations, and other travel related information.

Offer Their Own Shopping Malls

A number of banks have chosen to offer their own shopping malls or marketplaces that offer products and services from a relatively large number of Web merchants. Shoppers visiting these online shopping malls can often find web merchants or products available grouped in various categories such as apparel and accessories, books, music, and movies, and toys and video games. Citibank and Bank of America are two banks that offer online shopping malls on their Web sites.

Partner with Existing Shopping Malls

Two of the biggest credit card companies, Visa and MasterCard, offer online shopping malls to their card members: Visa with its *Visa Reward Online* and MasterCard with its *MasterCard Exclusives Online*. These online shopping malls offer special deals from various Web merchants to credit card holders. A number of banks are partnering with one or both of these credit-card-sponsored shopping malls to bring online shopping to their customers. LaSalle Bank, Fifth Third Bank, Wells Fargo, Charter One Bank, and People's Bank are some of the banks that are offering *MasterCard Exclusives Online* on their Web sites, and First Union Bank and People's Bank are two of the banks that are offering *Visa Reward Online*.

In addition to bringing online shopping to their customers, many banks are also offering a number of other shopping-related resources to their customers. For example, customers of Chase can access the bank's *Buyer's Guide* section to learn more about the different features they need to look for and pay attention to when purchasing certain products. Through its partnership with Epinions.com, Citibank is offering its customers the opportunity to search for reviews of the products that are of interest to them.

Some banks are also providing a number of shopping services to their customers. For example, Bank of America offers its customers a search engine to help them easily compare prices from different Web merchants. Customers visiting its online shopping mall, the Shopping Network, can search for products by specifying the type of product within the category, the brand, the minimum price, and the maximum price they are willing to pay. The Web site will then display the products and pricing information from Web merchants that meet the criteria specified by the customers.

Online Shopping and E-procurement for Small Businesses and Corporations

Compared to online shopping malls that cater to individual shoppers, there are much more limited choices for such online shopping malls for small businesses. A few banks do provide links to a small number of online stores for their small business customers. For example, Wells Fargo provides a link to OfficeDepot.com and Fleet Bank provides a link to Stamps.com. On the other hand, a few banks do offer bigger collections of links to Web merchants that provide products and services to small businesses. For example, Hibernia National Bank's Market Place groups the online merchants by categories such as business resources, taxes & accounting, financial management, and sales & marketing. Customers accessing the business resources category can find links to Web merchants such as Bplans.com, which offers business planning software packages, and Credit FYI, which offers small businesses credit checks on other small businesses.

The largest source of products and services available to small businesses online is the numerous electronic marketplaces available through a number of large banks' Web sites. Most of these electronic marketplaces are designed for the banks' corporate clients, but many of them are also accessible to small businesses. These electronic marketplaces are virtual marketplaces where buyers and sellers can get together to trade on a variety of products and services. Some of these electronic marketplaces are offered directly by the banks such as Web Procure from Fleet Bank and Key Procure from Key Bank. A number of other banks are offering them through their partnerships with eScout. Harris Bank, Citizens Bank, Charter One Bank, and First Tennessee Bank are several of the banks partnering with eScout.

These electronic marketplaces allow firms to search through a large quantity of goods and services, such as office supplies and computer products, from a variety of suppliers. They provide a streamlined procurement process with automated tracking capabilities and versatile control. With the aid of privilege-based security, firms can control their procurement expenses by predetermining the appropriate categories of products that can be purchased by their employees.

WHAT COULD BE NEXT?

In a relatively short period of time, Internet offerings from U.S. banks have blossomed from simple marketing information related to their products and services to a variety of online banking products and services for consumers, small businesses, and large corporations. Banks' customers now have access to numerous online tools to conduct their daily financial and business activities from anywhere and at anytime simply with their web browsers. In addition, the commercialization of the Internet has also presented banks an opportunity to establish themselves as a key player in the e-commerce arena by developing tools for both online B2C and B2B transactions and by establishing the infrastructure for conducting such transactions.

What could be next on the horizon in terms of U.S. banks' Internet offerings? The following are some potential near future developments.

Consumer Banking

Over the past few years, U.S. banks have focused on developing a number of online consumer banking tools to serve their customers. The purpose of these tools is to offer bank customers the opportunity to carry out many of their banking activities on their own via the banks' Web sites. In other words, these are convenient self-service tools with minimum human contact with the bank. However, as bank customers become more comfortable conducting their banking activities online, they also become more demanding in terms of the customer services they receive online (Pastore, 2001). For example, waiting for an e-mail response to a question is no longer acceptable for many bank customers. There will be a trend among many banks to adopt chat technology to offer instant responses to their customers' inquiries. In addition, to build a closer relationship with their customers, many banks will also increase their offers of customized e-mail reminders where the customers can preselect the events that will trigger an e-mail notification from the banks.

Since the passage of the Electronic Signatures in Global and National Commerce Act in 2000, the U.S. banking industry is still in the process of developing a universal standard for accepting electronic signatures over the Internet. Once such a standard is adopted and the proper security measures are put in place by the banking industry, which is expected to take place in the very near future, bank customers will be able to apply for a number of bank products completely online. For example, bank customers will no longer need to submit hard copies of their signatures when applying for an account. Banks will also be able to process a loan, from application to closing, completely over the Internet.

Banks will also strive to provide their customers with real-time information on their accounts. Currently, account information accessible online is generally not updated until the end of the business day. In other words, if bank customers withdraw money from an ATM machine or use their debit cards at a grocery store, that information will not be reflected on their account if they logged on to their accounts immediately.

Many banks will also be extending their reach to offer many of the currently available online tools to their customers through means other than computers. As Web-enabled wireless devices such as cellular phones and PDAs become more commonly available, many banks will be modifying many of their online offerings to allow delivery via such devices (Molwig, 2001). In addition, there will also be a trend among banks to Web-enable their ATM machines so that they can offer their customers another channel to access their online tools. Fleet Bank announced in late December of 2001 that it will be Web-enabling its ATM machines to allow its customers to carry out activities such as accessing and printing their 30-day account statements and paying their credit card and utility bills (Mearian, 2001).

Commercial Banking

Online commercial banking tools, thus far, have come primarily from the larger U.S. banks because they are the ones that traditionally offer commercial banking services such as treasury management and foreign exchange services to larger companies. They will continue to take the lead in offering many of their more complex commercial banking services over the Internet as new Internet tools and standards are developed and adopted. For example, as Internet security measures such as encryption technology continue to advance, there will be a trend among larger U.S. banks to provide Web interfaces for many of their existing electronic commercial banking services. In other words, banks will be offering their customers the ability to conduct many more complex banking activities with a browser rather than with specialized software and dedicated system. In addition, the adoption of new standards such as the extensible markup language (XML) will pave a new and more efficient way for banks to exchange data-heavy financial documents with their customers.

As the technology for online commercial banking tools matures, the cost of providing such tools will also drop dramatically. This will enable smaller banks to enter the arena of commercial banking that was previously not easily accessible to them due to the high costs associated with them. Even the smallest community banks will be able to offer many of these services (Wilson, 2001). The access to such online tools means that smaller banks will now be able to offer such tools to small businesses, which have traditionally been underserved by larger banks. There will be an increasing number of smaller banks offering online commercial banking services via their Web sites in the very near future. Unlike their larger counterparts, many of these smaller banks will focus their online offerings on the most basic commercial banking tools, namely treasury management services, because most small businesses tend to have less sophisticated needs.

E-commerce

Banks will continue to be involved in facilitating e-commerce activities in the retail market but the larger U.S. banks will be focusing more on the B2B e-commerce arena. Currently, the majority of B2B e-commerce activities are concentrated in the procurement area, where buyers and sellers trade with one another online. Many of the larger U.S. banks have already been heavily involved in developing many of the standards for such B2B online transactions. In addition, a number of these banks have also partnered with other entities to develop the infrastructure for conducting such transaction, such as B2B exchanges. These banks will continue to dedicate their resources in further developing transaction-oriented initiatives and infrastructure for B2B e-commerce activities. They realize that it is important for them to play a key role in shaping the financial transaction settlement infrastructure for this particular market and they are positioning themselves to be the key players as B2B e-commerce activities evolve beyond procurement and into the entire supply chain of an industry.

GLOSSARY

Automated clearinghouse (ACH) network An electronic system for settling transactions among financial institutions.

Electronic data interchange (EDI) A system, among big companies, to exchange information over a private network, via dial-up lines or dedicated leased lines.

Java A cross-platform programming language that is often used to create Internet-based applications, i.e., applets.

JavaScript A scripting language used to control the behaviors of certain elements in a Web page.

Pure play A company that exists solely online and does not have a physical establishment.

CROSS REFERENCES

See *Electronic Commerce and Electronic Business; Electronic Data Interchange (EDI); Electronic Procurement; Internet Literacy.*

REFERENCES

Arend, M. (1994, February). Son of ATM is growing up fast. *ABA Banking Journal, 86,* 74–76.

Bers, J. S. (1996, October). Capturing the burgeoning small business market via the Internet. *Bank Systems & Technology,* 40–43.

Courter, E. (2001, August). Quicker shopping. *Credit Union Management,* 26–30.

Fonseca, I., Hickman, M., & Marenzi, O. (2001, Summer). The future of wholesale banking: The portal. *Commercial Lending Review,* 23–35.

Gart, A. (1992, Spring). How technology is changing banking. *Journal of Retail Banking, 14,* 35–43.

Gray, D. (1994, May). Here comes home banking—Again. *Credit Card Management,* 54–58.

Hoffman, T., & Kim, N. (1995, October 23). In net we trust. *Computerworld,* 14.

Mearian, L. (2001, December 17). FleetBoston to Web-enable ATMs, revamp technology. *Computerworld,* 14.

Messmer, E. (1995, May 15). Banks surge ahead in electronic commerce. *Network World,* 12–14.

Molwig, D. (2001, August). Pragmatism reigns. *Credit Union Management,* 55–56.

Nixon, B., & Dixon, M. (2000). *Sams teach yourself e-banking today*. Indianapolis: Sams Publishing.

Online Banking Report true Web banks & credit unions database [data file]. Retrieved December 1, 2001, from http://www.onlinebankingreport.com/resources/

Online for cyberbanking. (1995, February). *Financial Technology International Bulletin,* 9.

Orr, B. (1998, June). Community bank guide to Internet banking. *ABA Banking Journal, 90,* 47–53.

Pastore, M. (2001, October 25). *Connecting channels key to online banking.* Retrieved June 30, 2002, from http://cyberatlas.internet.com/markets/finance/print/0,,5961_910511,00.html

Radigan, J. (1996, April). Have on-line services peaked? *US Banker,* 54–58.

Serwre, A. (1995, June 26). The competition heats up in online banking. *Fortune,* 18–19.

Treadwell, T. (2001, November). The state of online lending. *Credit Union Management,* 36.

Wilson, C. (1997, June). The Internet comes to home banking. *Community Banker,* 10–15.

Wilson, C. (2001, May). Cash management services. *Community Banker,* 22–25.

Yamada, K. (1995, November 27). Intuit, America Online team up for electronic banking. *Computer Reseller News,* 65.

FURTHER READING

Anchor Savings Bank (http://www.anchornetbank.com)
Bank of America (http://www.bankofamerica.com)
Bank of Smithtown (http://www.bankofsmithtown.com)
Bank One (http://www.bankone.com)
BNC National Bank (http://www.bncbank.com)
Busey Bank (http://www.busey.com)
Carrollton Bank (http://www.carrolltonbank.com)
Charter One Bank (http://www.charterone.com)

Chase (http://www.chase.com)
Citibank (http://www.citibank.com)
Citizens Bank (http://www.citizensbank.com)
Citizens Financial Bank (http://www.citizensfinancialbank.com)
City National Bank of Florida (http://www.citynational.com)
Comerica Bank (http://www.comerica.com)
Fifth Third Bank (http://www.53.com)
First Tennessee Bank (http://www.firsttennessee.com)
First Union Bank (http://www.firstunion.com)
Firstar Bank (http://www.firstar.com)
Fleet Bank (http://www.fleet.com)
Harris Bank (http://www.harris.com)
Hibernia National Bank (http://www.hibernia.com)
HomeStreet Bank (http://www.homestreetbank.com)
Key Bank (http://www.key.com)
LaSalle Bank (http://www.lasallebank.com)
Net Bank (http://www.netbank.com)
People's Bank (http://www.peoples.com)
PNC Bank (http://www.pncbank.com)
Standard Federal Bank (http://www.standardfederalbank.com)
TCF Bank (http://www.tcfbank.com)
U.S. Bank (http://www.usbank.com)
United California Bank (http://www.unitedcalbank.com)
Washington Mutual (http://www.wamu.com)
Webster Bank (http://www.websterbank.com)
Wells Fargo (http://www.wellsfargo.com)
Zions Bank (http://www.zionsbank.com)

Online Communities

Lee Sproull, *New York University*

INTRODUCTION

The Internet was not invented as a social technology, but it has turned out to be one. From the earliest days of the ARPANET (network of communicating computers established in the late 1960s with U.S. government funding), people have shaped and used the technology for social purposes. Today millions of people use the Net as a means of making and maintaining connections with other people who share a common experience, interest, or concern. The Net-based social contexts range from family e-mail to fantasy games with hundreds of thousands of players. This chapter focuses on a subset of net-based social contexts, which in recent years have come to be called online communities. These are large voluntary online groups, composed primarily of people who have no preexisting ties with one another and who may never meet face-to-face. Online communities range in technical sophistication from Usenet discussion groups to complex multiplayer fantasy games supported by proprietary software. They range in purpose from entertainment to political dissent. They range in access restrictions from totally open to closed behind corporate firewalls. (This chapter does not discuss private online communities.)

In addition to describing how technical and social factors mutually interact to produce and sustain online communities, the purpose of this chapter is to begin to produce a differentiated view of online communities. Online communities share some underlying attributes and processes, but they differ in member interests, goals, processes, benefits, and negative consequences. A more differentiated view will make possible more productive theorizing, research, and design.

DEFINITION AND ATTRIBUTES

Definition

Offline communities are defined as collectivities based on members' shared experience, interest, or conviction, and their voluntary interaction in the service of member welfare and social welfare (Etzioni & Etzioni, 1999; Knoke, 1986). Examples include neighborhood communities, religious communities, civic and social communities such as youth scouting or service clubs, and collections of like-minded enthusiasts such as sports fans. Communities in the physical world can be described by structural attributes that exist independent of any member such as rules, roles, and resources. Thus one can talk about the size of a community, membership requirements and obligations, resources and amenities, member characteristics, interaction patterns. Communities can also be described by psychological attributes internal to their members such as feelings of trust, identification, and commitment. Both structural and psychological attributes exist along a continuum, so one can find more or less well-structured communities with more or less committed members. In all cases, however, the term "community" usually suggests positive feelings, prosocial behavior, and choice. (People rarely talk about a prison community even though its inmates interact and have experiences in common.)

The definition of online community is also based on shared experience, interest, or conviction, and voluntary interaction among members in the service of member welfare and social welfare. An online community is defined as a large, voluntary collectivity whose primary goal is member or social welfare, whose members share a common interest, experience, or conviction, and who interact with one another primarily over the Net. Online communities can have more or less structure and more or less committed members. This definition excludes electronic work groups and virtual teams, whose primary goal is economic, whose members are paid, and whose size is relatively small. It excludes ad hoc friendship groups and buddy lists, which are relatively small and unstructured and whose members may interact primarily in the real world. It also excludes nominal groups such as all-the-people-who-use-Google (who neither share a common interest nor interact with one another) or all-the-people-who-donate-cycles-to-Seti@home (who may share a common interest but do not interact with one another). Others who have offered definitions of electronic communities include Figallo (1998), Kim (2000), Powazek (2002), Preece (2000), and Rheingold (2000).

Until the mid-1990s, almost no one used the term "online community." Instead these groups were named after the technology that supported them and were called "newsgroups," "listservs," "mailing lists," "BBSs," or "Free-nets." In some ways, online communities bear little resemblance to communities in the physical world. They own few tangible resources and amenities; they require no visible or tangible commitment from members (such as taxes, dues, attendance at meetings); members may never see or meet one another face-to-face. Yet some online communities have resources that may be economically valuable: their domain name, the wisdom accumulated in their FAQs (frequently asked questions), the intellectual property created by their members. Fantasy game characters and properties have yielded nontrivial sums for their creators on eBay auctions. The members of one voluntary online community collected enough money in member donations to buy a new server to host the community (Boczkowski, 1999). Nevertheless, calling any electronic site where people may gather an "online community," doesn't make it one. As in the physical world, the term carries positive connotations, and some who have used it are merely guilty of wishful thinking. This may be particularly true for those who, in the late 1990s, aspired to create online communities for profit (e.g., Bressler & Grantham, 2000; Hagel & Armstrong, 1997).

Supporting Technologies

In addition to the packet-switching technology of networked communication, many online communities rely on message-based group communication applications to support member interaction. Asynchronous discussion is often supported via mailing lists or bulletin board applications. Mailing lists, a push technology, send group messages to a person's e-mail inbox where they intermingle with the person's other mail. With bulletin boards, a pull technology, a person reads group messages organized by topic in a file exclusively devoted to that group. (Some people establish filters to move all mailing list messages into separate folders, thereby making distribution lists function somewhat more like bulletin boards.) Synchronous discussion may be supported via talk programs such as IRC (Internet relay chat) or text-based virtual reality (VR) environments. MUDs (multiuser dungeon, domain, or dimension) and MOOs (MUD object oriented) began as text-based virtual reality games, a computer-based version of fantasy games such as Dungeons and Dragons. Today some are still organized as fantasy games; others are organized for professional or social purposes. Their spatial metaphors and programmable objects and characters are the precursor of today's graphically based fantasy games.

With the spread of the Web and graphical browsers in the late 1990s, online communities could support more varied forms of interaction on their Web sites. Some use real-time chat for discussions that are scheduled and announced in advance. Some use special file formats to share image, sound, or video files. Online game communities are supported by more or less elaborate software that supports play in board games like chess or supports avatar creation and interaction in fantasy worlds. Most discussion among members on online community Web sites still occurs in message-based discussions, however. Even the fantasy games have discussion boards.

General Attributes and Processes

The Social Psychology of Message-Based Communication

Reduced social context cues characterize all discussion-based communities, relatively to face-to-face discussions. Weak social context cues mean that messages carry few explicit reminders of the size of the group or members' physical appearance, social status, or nonverbal reactions compared with face-to-face or telephone communication. (In virtual reality communities, people can create entirely new personas.) Therefore communication tends to be relatively frank and open (Kendall, 2002; Reid, 1999; Sproull & Kiesler, 1991). Weak social context cues can lead to different effects, even for people within the same community. It can increase affiliation and commitment among members because objective differences among them are obscured and subjective similarities, based on their common interest, are magnified (Galegher, Sproull, & Kiesler, 1998; Mackenna & Bargh, 1998; Sproull & Faraj, 1995). Alternatively, it can increase disaffection and dropout because it may be more difficult to establish common ground or consensus and manage conflict (Carnevale & Probst, 1997; Cramton, 2001; Dibbell, 1998; Herring, 1994; Kollack & Smith, 1996). Asynchrony and weak social context cues condition communication in many online communities. They allow for a potentially greater geographic and social diversity of participants than many physical communities do. At the same time they offer few cues to that diversity in interaction, except those revealed through linguistic cues. There is a body of research on gender differences in online language (e.g., Thomson & Murachever, 2001).

In synchronous communities, people can participate from any place that has technology available. In asynchronous communities, people can also participate at any time as well as from any place. Common interest and technology become the only two requirements for community participation; geographic location is irrelevant.

Microcontributions

Many of the tools for electronic group discussion are based on a relatively fine granularity of time and attention—the text message. Although messages can be any length, ones sent to asynchronous discussion communities typically range between 10 to 30 lines or one to two screens of text. For example, Winzelberg (1997) reported a mean of 131 words; Galegher et al. (1998) reported a mean of 8 to 20 lines of new text; Wasko and Faraj (1999) reported a mean of 25 to 30 lines of text; Sproull and Faraj (1995) reported a mean of 22 to 42 lines of new text. Single synchronous contributions in a chat room or VR environment are usually even shorter. In asynchronous communities, people can read one or more messages and post or send one or more messages at their convenience. The message can be thought of as a microcontribution to the community. When people are online much of the day as a part of their work, voluntary microcontributions can be interspersed throughout the work day. Even in synchronous communities, some members report that they keep a community window open on their screen while they are working. Every once and a while they "check in" on the community (Kendall, 2002). Or participation can be interspersed with other activities at home in the evenings. Some people may devote hours a week to an online community, but they can do so in small units of time at their own convenience.

Communities based on microcontributions have relatively low barriers to entry. It is fairly easy to read enough microcontributions to know if a particular community is relevant or appropriate. If so, that reading readily demonstrates the appropriate form that a newcomer's own microcontributions should take. The production and posting of initial microcontributions is relatively low effort. Then, if all goes well, the newcomer receives positive reinforcement in the form of (easy to produce) responding microcontributions from other community members. Whereas microcontributions create low barriers to entry, they may also create high barriers to commitment. It can be difficult to develop complex arguments or achieve nuanced understanding through microcontributions. Communities that are easy to join are often just as easy to leave.

Aggregation Mechanisms

Although people can make microcontributions at random, they must be organized into larger units for efficiency and social effectiveness. Both technical and social mechanisms are necessary to organize the smallest unit of contribution into larger units that are useful to participants. Software for asynchronous discussion lets people indicate that their contribution is a response to a previous one. All contributions so designated could be aggregated by the software and displayed as "threads"—a seed message and all reactions to it. Forms of threads common from the earliest days of the net include a question with replies and a proposal or statement with comments. Threads organize microcontributions so that everyone can see their constituent parts, making it easy for potential contributors and beneficiaries to see what has already been said. (Software also allows readers to mark threads they have already read or to display only unread messages.)

Asynchronous discussion communities may have tens or hundreds of threads active at the same time, necessitating a level of organization beyond the self-organizing thread. Here a human designer may suggest or impose a topic map or architecture to group threads into more general topic categories. (In some older bulletin board systems, people would vote to create a new top-level topic, which would get its own separate bulletin board.) Web-based software can display these maps graphically so that users can click on a topic that interests them and see all threads related to that topic. The shared interest of a group usually suggests the type of topical map that may be created. For example, medical concern communities may have topics for symptoms, medications and side effects, negotiating the health care system, and managing relationships with family and friends. Movie or television fan communities may have topics for major and minor characters, actors who play those characters, and past and future episodes. Creative communities that build software will have categories for different types of code, bug reports, patches, and documentation.

Threads and topic maps may not be sufficient to structure extremely large numbers of messages. Thus was created another form of micro-contribution, the rating message. Some Web sites now give members the opportunity to rate the contribution of others' messages. Software then aggregates and displays these ratings in an overall quality index for contributions (e.g., Slashdot) or contributors (e.g., MotleyFool). Overt ratings of contribution quality surely increase economic trust within an electronic market like eBay buyer and seller ratings (Kollock, 1999). It is an open question whether they increase or inhibit emotional trust and cohesiveness within an electronic community setting.

Software for synchronous interaction may organize contributions in channels or use a spatial metaphor to organize contributions in "rooms." Residence-based communities may organize contributions around civic functions like the garden club or public library. As these communities increase in membership, the organizers or members themselves construct new rooms, buildings, and territories to organize interaction. Some also offer rating and review functions to rate characters or contestants and properties.

Norms and Motivations

Some norms of community behavior, which were visible from the early days of the ARPANET, prevail across many types of online communities. One is peer review of content (Benkler, 2002). In most electronic discussion communities and creative communities, there are no "authorities" to certify the accuracy of message content. Instead, it is expected that members themselves will comment on the quality, accuracy, completeness, and so on of one another's contributions. Similarly, peer review of behavior is also expected: it is normative for members to chastise or complain about inappropriate behavior and praise helpful behavior. Most important is the norm of altruism. Community members freely offer information, advice, and emotional support to one another with no expectation of direct reciprocity or financial reward.

The fundamental dynamic supporting discussion communities is that someone asks a question or makes a proposal or statement and other people provide answers or comments. Utilitarian self-interest may be all that motivates the askers—a personal need for information. But pure self-interest does not explain the behavior of people who reply. By definition, they are not paid for their replies. Because they are unlikely to have a personal relationship with the person they help, neither friendship obligation nor the expectation of direct reciprocity is likely to impel their behavior. Indeed, an early influential paper (Thorn & Connolly, 1987) predicted that computer-based systems relying on volunteers would be doomed to failure. They argued that people who could give the best replies would have no incentive to participate because they would receive few benefits for doing so. It was unlikely that anyone could answer their questions; their time would be unrewarded. (In social dilemmas, helping in these situations is known as the "sucker's choice.") Over time, therefore, the quality of help would decline until people no longer even bothered to ask questions. The fallacy in this argument is the assumption that replier rewards must come in the same form as the help they give. Yet motivations for helping behavior can be quite complex. In studies of volunteers in the physical world, motivations include commitment to the cause or interest associated with the community, the desire to help others, benefits from displaying expertise, and the personal satisfaction and self-esteem derived from helping others (e.g., Clary et al., 1998; Omoto & Snyder, 1995). Studies of electronic discussion communities document a similar combination of motivations for people who answer questions and otherwise support their online community (e.g., Butler, Kiesler, Kraut, & Sproull, 2002; Lakhani & von Hippel, 2002; Wasko & Faraj, 2000).

In virtual reality communities, the environment indexes participants' motives (Reid, 1999). In VR game communities, the motives are tied to the rules of the game: amass property, kill enemies, design an award-winning room, and so on. In VR professional communities, the motives are tied to the profession: contribute to shared databases, review articles, and participate in policy discussions. In VR social communities, the motives are tied to exploring social worlds.

HISTORY OF ELECTRONIC COMMUNITIES

Both the technical and social trajectories of electronic communities began with the design and early deployment of networked computing in the late 1960s and early 1970s. (See Table 1 for time line.) Whereas computer networking was initially conceived as a way to share scarce computing resources located in one place with researchers at other places via remote access, it soon became a convenient way for people to gain remote access to other people as well as to computers (Licklider & Veza, 1978; Sproull & Kiesler, 1991). During the 1970s and 1980s, people created additional networks and wrote e-mail and bulletin board software, technical innovations and improvements that made networked computing more useful for supporting human communication. During the 1990s, the technical innovations of the Web and the graphical browser supported the broad diffusion of electronic communication to millions of U.S. households and hundreds of millions of people worldwide.

The social trajectory of electronic discussion communities had its beginnings in the same technical community that invented the ARPANET. By the mid-1970s, ARPA program officers and researchers around the country had begun using group e-mail to share results, discuss plans, and organize meetings. Some researchers also began using group e-mail for purposes unrelated to work: for example, to share opinions on cheap Chinese restaurants in Boston and Palo Alto, favorite science fiction books, inexpensive wine, and new movies. This research community invented both the technology (group communication tools) and the new form of social organization (the voluntary electronic group). They appropriated technologies that were created for utilitarian purposes to create self-organizing forums for the voluntary discussion of common interests. The social trajectory of VR communities began in 1979–80 with an effort to program a game that would be like the fantasy game Dungeons and Dragons. The first multiuser VR game was accessible on the ARPANET in 1980.

By the 25th anniversary of the ARPANET in 1994, electronic group communication had become a taken-for-granted process and voluntary discussion communities had become a taken-for-granted organizational form in universities, technical communities and scientific disciplines, and some corporations (e.g., Finholt & Sproull, 1990; Kiesler & Sproull, 1987; Orlikowski & Yates, 1994; Walsh & Bayma, 1996). Multiplayer games were also popular on university campuses. Despite their growth, at this point both discussion groups and games were still in large measure the province of young, technically adept men. The final years of the 20th century saw Net-based communication enter the mainstream of U.S. life because of a combination of technical and economic developments. The technical developments were the Web and the graphical browser, which made it much easier for ordinary people to find and access information and groups on the Net. The economic developments were the commercialization of the Net and AOL's business model. Once the Net was commercialized, corporations began to see potential economic value in electronic communities and so endeavored to support them and, indeed, to "monitize" them (e.g., Hagel & Armstrong, 1997). The online game industry began to grow. AOL emphasized member discussion groups in contrast with other commercial services that were still emphasizing access to databases. By 2000, AOL had 34 million members—more than all the other commercial services combined—many of them participating in electronic forums and communities of interest. That year, 44 million U.S. households were on the Net. Despite the enormous influx of people very different from the ARPANET pioneers, four themes evident from the earliest days of the ARPANET continued to characterize electronic communication at the beginning of the 21st century: access to people as much as to databases, group communication as well as dyadic communication, personal interest topics as well as utilitarian ones, and self-organizing voluntary electronic communities.

TYPES OF ONLINE COMMUNITIES

Until the early-1990s, most electronic communities used similar technology and their members had similar attributes. Highly educated, young, technically adept people congregated electronically with similar others who shared similar interests. These congregations were relatively homogeneous in structure, process, and membership. Now, however, the diversity of Internet users and group goals and processes is so great that it is helpful to differentiate types of communities to understand their costs and benefits in more detail. This section categorizes and describes types of online communities based on the interest that members or sponsors have in common. (See Table 2 for examples.) It is neither an exhaustive nor a mutually exclusive characterization, but it does represent some prevalent types. Despite the differences across types of shared interest, all of these communities are characterized by any-time, any-place communication with weak social context cues, aggregated microcontributions, and norms of interaction. The boundaries across types can be

Table 1 Timeline for Community-Oriented Group Communication on the Net

DATE	NAME	TECHNICAL	SOCIAL
1965–68	ARPA projects	Research on networking to share scarce computing resources	Networking research community forms across small number of labs
1969	ARPANET	First four ARPA sites connected	
1972	RD	First ARPANET e-mail management program	
1973			75% of ARPANET traffic is e-mail
1975	MsgGroup	First ARPANET mailing list	
1978	BBS	First bulletin board system	
1979	Usenet	Free software to share bulletin board discussions	
1979	MUD	First multiuser VR game	
1979	CompuServe	Began offering e-mail to customers	
1980	MUD	MUD first played over ARPANET	
1981	Bitnet	Computer network for non-ARPANET universities	University computer center directors band together to support this
1981	Sendmail	Free software to send mail across networks	
1984–85	Delphi, Prodigy, AOL	Commercial information services founded that offered e-mail	
1985	Listserv	Free software to manage e-mail lists	
1986	IETF		Volunteers focused on technical operation and evolution of Internet
1988			>1,000 public listservs
1990			>1,0000 Usenet groups; >4,000 posts per day
1990	World Wide Web	Invented by Tim Berners-Lee	
1990	LambdaMOO	VR environment created at Xerox PARC	
1991	Linux	Free computer operating system	First message about Linux posted to Usenet group
1992	AOL	Connects to Internet	
1994	Netscape	Introduced graphical Web browser	
1994			>10,000 Usenet groups; >78,000 posts per day; Estimated 3.8 million subscribers to commercial online services
2000			34 million AOL subscribers; 44 million U.S. households online
2001			90,000 Usenet groups; 54,000 public Listserv groups

IETF = Internet Engineering Task Force; MUD = multiuser dungeon, domain, or dimension; VR = virtual reality.

fuzzy; the descriptions indicate central tendencies within types.

Types by Member Interest

Customer Communities: Brands and Fans

Customer communities are composed of people who share a common interest in and are loyal to a particular brand, team, entertainer, or media property. Although people organized fan clubs prior to the net, it was difficult for them to arrange activities on a large scale and frequent basis. Online customer communities have a much broader reach. People voluntarily share their information and passion with thousands or hundreds of thousands of others who share their interests in a particular product,

Table 2 Examples of Online Communities

Customer Communities	
http://www.audifans.com	Fans of the Audi marque
http://www.lugnet.com	Adult fans of Lego
http://www.Britney-Spears-portal.com	Fans of Britney Spears
http://Rec.arts.tv.soaps	Soap opera fans
(A)vocation Communities	
http://www.mastersrowing.org	For masters rowers and scullers
http://www.bikeforums.net	For the avid cyclist
http://www.Everquest.com	For Everquest players
http://www.chessclub.com	For chess fans
http://www.tappedin.sri.com	For K–12 teachers
http://LambdaMoo	Virtual reality environment for social interaction
Place-Based Communities	
BEV	Blacksburg Electronic Village
Netville	Wired suburb of Toronto, Canada
Condition Communities	
http://www.seniornet.org	For people over age 50
http://www.systers.org	For female computer scientists
Alt.support.depression	For sufferers of depression
BHL (Beyond Hearing mailing list)	For people with hearing loss
AML (Argentine mailing list)	For Argentinian expatriates
Concern Communities	
http://www.clearwisdom.net	For practitioners of Falun Gong
http://www.moveon.org	For online political activists
Talk.guns	For handgun advocates
Soc.religion.mormon	For believers in Mormonism
Creative Communities	
LKML	For developing the Linux kernel
http://www.wikipedia.com	Collaborative project to produce an encyclopedia
IETF	For maintaining and improving the Internet
http://www.rhizome.org	For creating and discussing new media art

entertainer, or media property. AudiFans, for example, is composed of more than 1,000 Audi enthusiasts who exchange information about parts suppliers and mechanics, post photos of their cars, and share the joys and sorrows of Audi ownership. The Britney Spears portal contains pictures, MP3 files, news, and forums where thousands of people comment (positively or negatively) on all things having to do with Britney. For most members of customer communities, the shared interest may be an intense one, but it typically represents a fairly small and often short-lived portion of members' lives.

(A)vocation Communities

Experts and enthusiasts form and join voluntary (a)vocation communities to increase their pleasure and proficiency in their hobbies or work. A particular product may be a means to advancing a common interest in an a(vocation) community, but it is not the primary focus of attention. From bicycling to computer programming, dog training to quilting, karate to the Civil War, there are communities for people who share these interests. "How-to" information prevails in discussions and people who share their expertise are greatly appreciated. BikeForums, for example, has more than 3,500 members who discuss and debate bicycle commuting, mountain biking, tandem biking, racing, training, and so on. TappedIn is a professional VR community for K–12 educators. The Internet Chess Club has more than 25,000 members who play chess with other members, take lessons, play in tournaments, watch grandmaster competitions, and so on. Everquest and Ultima are large online communities for people who delight in fantasy games. The shared community interest in (a)vocation communities may represent a relatively enduring part of members' lives.

Place-Based Communities

These communities are organized by and for people who live in a particular locale. Their genesis in the early 1980s had a political agenda—to give residents a(n electronic) voice in the local political process (e.g., Schuler, 1996).

More recent versions have had the broader goal of building social capital by increasing the density of electronic social connections among residents of physical communities (Hampton & Wellman, 1999; Kavanaugh, 1999). In principle, the shared interest should last at least as long as people reside in the community.

Common Condition Communities

In these communities, people share the experience and interest of a common condition. It may be based on a demographic characteristic such as race, age, or ethnic background; a medical or psychological condition such as arthritis or depression; or being an alumnus/a of a particular organization such as a college or branch of the military. People join condition communities to learn how others experience or are coping with their condition and to share their own experiences. Along with practical information and advice, "you are not alone" sentiment prevails in many discussions. BeyondHearing, for example, has more than 1,000 members who have a hearing loss or who have a loved one with a hearing loss. Topics range from cochlear implants to the best audiologists to funny stories about lip reading mistakes. Systers' membership is more than 3,500 female computer scientists and engineers who discuss female-friendly graduate schools and employers, how to manage male subordinates, and so on. The shared community interest is often a long-term or lifetime one for members.

Concern Communities

In these communities, members share an interest in a common political, social, or ideological concern. Because members often wish to influence the state of affairs in the physical world, these communities usually have multifaceted ties to that world. They may announce and comment on real-world events and organize letter-writing campaigns, rallies, fundraisers, and so forth. They may use click-and-donate applications to raise money or pledges of volunteer time. MoveOn, for example, began as an online petition drive to censure Bill Clinton and has grown to include many online advocacy groups that organize volunteer campaigns. More than 1,000 members of soc.religion.mormon discuss and debate Mormon doctrine and practices. The shared interest of concern communities is likely to be a deep and abiding one for their members.

Creative Communities

Unlike other community types whose primary output is talk or amusement, members of creative communities work collaboratively on the Net to produce real products, be they software, literary works, or other artistic creations. Most open source software is produced in voluntary communities today despite the growing interest of the corporate sector (e.g., Raymond, 1999). Indeed much of the design and engineering of the Internet itself is accomplished by a voluntary community that conducts most of its business electronically, the Internet Engineering Task Force (n.d.). Poets participate in writers' communities whose members thoroughly critique one another's work. A more pragmatic writing community is creating a new encyclopedia. In the first 18 months of its existence, the community project has produced more than 40,000 entries (http://www.wikipedia.org/wiki/Main_Page). The shared interest of creative communities is likely to be deeply involving for members, although it need not be as enduring as the shared interest in concern communities.

Types by Sponsor Interest

Only within the past 10 years have online communities been sponsored or organized by anyone other than members themselves (with a few exceptions such as the Well and geographically based bulletin board systems called Free-nets). Many recent third-party sponsors or organizers have been motivated by the profit potential of online communities with revenue models based on either sales (of advertising, membership lists, products, etc.) or subscriptions. During the dot-com boom, third-party sponsors of some customer and demographic condition communities used sales-based revenue models. Profit-oriented community sites were created for L'eggs panty hose, Kraft food products, women (iVillage), Asian Americans (Asia Street), and African Americans (NetNoir), for example. Some third-party-sponsored communities based on subscriptions are relatively healthy; arguably a substantial (but unknown) fraction of AOL's subscriptions are a function of online community memberships. Within the game industry, subscription-based revenue models have been successful. Several multiplayer game communities have thousands or hundreds of thousands of members. Members of one game, EverQuest, report spending an average of more than 22 hours a week playing it (Yee, 2001).

Some corporations have avoided revenue-based models and instead have supported online communities to build market share or increase customer satisfaction and loyalty. Sun Microsystems sponsors the Java Developer Connection, an (a)vocation community designed to support and expand the Java software developer community worldwide. The Lego Corporation supports several "adult-fans-of-Lego" sites, in addition to sponsoring its own site, to support loyal customers. Various software companies support voluntary technical discussion and support communities in the interest of increasing high-quality, inexpensive tech support. Harley-Davidson uses H.O.G., its online members-only group, to reinforce the brand loyalty of Harley owners worldwide.

In the not-for-profit sector, foundations and service organizations have sponsored communities for their target populations with the goal of improving their welfare. Thus, for example, The Markle Foundation sponsored the creation of Seniornet, a not-for-profit online community for people over age 50, which currently has 39,000 members. The National Science Foundation sponsored TappedIn, a not-for-profit online community for K–12 school teachers and teacher developers, which currently has more than 14,000 members.

ONLINE COMMUNITY CONSEQUENCES

Positive Effects

Benefits to Members

Not surprisingly, most studies of online community members report that information benefits are important to them (e.g., Baym, 1999; Lakhani & von Hippel, 2002; Wasko & Faraj, 2002). What is noteworthy is the form that the information takes. It is not the disembodied, depersonalized information that can be found in databases or official documents, which are themselves easily accessible on the Web. Instead, it is profoundly *personalized* information. Its form and content are personal—personal experiences and thoughts. Likewise, its audience is personal. Questions or requests for comment do not look like database queries: They are framed for human understanding and response. (A discourse analysis of Usenet groups found that almost all questions included a specific reference to readers; the few that did not were much less likely to receive replies; Galegher et al., 1998.) Replies typically address the person or situation engendering the request and are based on the replier's own situation or experience. In customer communities personalized information can increase members' pleasure in using or experiencing the product or property. Personalized information can increase members' pleasure or competence in practicing their (a)vocation. It can also challenge one's assumptions and beliefs (e.g., Kendall, 2002).

Members derive more than information benefits from online communities, however. Some also derive the social benefits that can come from interacting with other people: getting to know them, building relationships, making friends, having fun (e.g., Baym, 1999; Butler et al., 2002; Cummings, Sproull, & Kiesler, 2002; Kendall, 2002; Quan y Hasse, Wellman, Witte, & Hampton, 2002; Rheingold, 2002). Occasionally these social benefits are strong enough that they lead some members to organize face-to-face group activities, such as parties, rallies, show and tell, reunions, or meeting at conferences or shows.

Members of medical and emotional condition communities may derive actual health benefits from their participation in addition to information and social benefits. The evidentiary base for these benefits is small but comes from carefully designed studies that use either random assignment or statistical procedures to control for other factors that could influence health status. Benefits for active participants include shorter hospital stays (Gray et al. 2000), decrease in pain and disability (Lorig, Laurent, Deyo, Marnell, Minor, & Ritter, 2002), greater support seeking (Mickelson, 1997), decrease in social isolation (Galegher et al., 1998), increase in self-efficacy and psychological well-being (Cummings et al., 2002; Mackenna & Bargh, 1998).

Membership benefits do not accrue equally to all members. Passive members, those who only read messages, as a class may benefit least. This speculation is consonant with research in physical world groups and communities that finds that the most active participants derive the most benefit and satisfaction from their participation (e.g., Callero, Howard, & Piliavin, 1987; Omoto & Snyder, 1995). Most studies of online communities investigate only active participants because they use the e-mail addresses of posters to identify their research population; they have no way of identifying or systematically studying people who never post but only read. Research that has studied passive members systematically finds that they report mostly information benefits; their total level of benefits is lower than that for more active participants; they are more likely to drop out (Butler et al., 2002; Cummings et al., 2002).

Among active participants, people who participate more extensively report having a greater sense of online community (Kavanaugh, 1999; Quan y Haas et al., 2002). More frequent seekers of information report receiving more helpful replies than less frequent seekers (Lakhani & von Hippel, 2002). More frequent providers of information report greater social benefits, pleasure in helping others, and pleasure in advancing the cause of the community (Butler et al., 2002).

Benefits to Third Parties

Many attempts to directly "monitize" online communities through sales revenue from advertising or commerce have been relatively disappointing (Cothrel, 2001; Figallo, 1998; Sacharow, 2000). Although the potential customer base could be quite large for brand or demographic condition communities, attracting and retaining customer/members is difficult. By contrast, subscription revenues—in the online game industry at least—have been relatively robust. Specific figures are hard to get from privately held companies. But estimates are that revenues from online games were more than $200 million in 2000 and will grow to more than $1 billion in 2004 (Zito, quoted in Castronova, 2001).

In customer and (a)vocation communities, substantial nonrevenue benefits may accrue to corporations through reinforcing customer brand loyalty and increasing customer satisfaction. Voluntary personal testimonials about a product or experience in the context of giving help can be quite persuasive, both to the person who asked for help or comment and to others reading the exchange. Motivational theories of attitude formation (e.g., Festinger, Schachter, & Back, 1950) and information processing theories of decision making (e.g., Nisbett & Ross, 1980) both point to the influential nature of voluntary personal testimonials. The process can be so powerful that there have been unsubstantiated reports of paid shills masquerading as community members in customer communities (Mayzlin, 2001).

Much of the help offered on customer and (a)vocational communities is personalized customer support—and potentially quite high-quality support at that. In the software industry, online communities have been recognized as the Best Technical Support Organization of the Year for 1997 and 1999 (Foster, 1999). Information that solves customer problems or enhances their product experience is likely to increase their satisfaction and loyalty. When it is provided by self-organized volunteers, the corporate cost is minimal.

Some online communities offer potential product development benefits. Most remarked are probably open source software communities that have generated product revenues for companies like Red Hat and product enhancements for companies like IBM and Sun Microsystems. Some game and hobby communities offer extensive product testing before widespread product release (e.g., Wallich, 2001). Some actively propose, design, and discuss new product features.

The strategic question for corporations today centers on what type of corporate involvement in online communities is likely to bring the greatest benefit. With few, but important, exceptions, direct corporate ownership and control of online communities is unlikely to be the answer. (This is not to say that corporations won't benefit from Web-based sales and customer support; e-commerce can be profitable even if online community-based ecommerce is not likely to be.) Forging positive and productive relationships with independent online communities can be challenging, however.

Benefits to Society

Rigorous empirical evidence is almost nonexistent for online community benefits to society. If members of condition communities achieve improved health status, the cost of their medical care to themselves and society could decrease. Alternatively, better informed members may seek out additional tests or treatments, thereby increasing the cost of their care. If members of targeted populations such as K–12 schoolteachers or senior citizens derive cognitive, social, and emotional benefits from participating in online communities, then the larger society may benefit as well. Data from Blacksburg Electronic Village suggests that participation in online community activities can increase civic involvement (Kavanaugh, 1999). If members of online concern communities can more effectively mobilize, then the causes served by their advocacy are likely to benefit (e.g., Gurak, 1997; Quan y Haas et al., 2002). Note, however, that online communities can advocate for harmful causes just as easily as they can for helpful ones.

Negative Effects

Although anecdotes are widespread, systematic evidence on the negative consequences of online communities is sparse. Members can be harmed by erroneous, malicious, or destructive information. The norm of peer review of content in discussion communities acts as a damper on this kind of harm but cannot prevent it entirely. More seriously but less likely, members can be harmed by unhealthy, dangerous relationships that can form via online communities. Unscrupulous or criminal intent can be masked by an online persona that inspires trust and friendship within the community context. If a relationship moves out of community scrutiny and into private e-mail, it can lead to emotional harm or even physical danger. Within VR communities, there have been a small number of widely publicized "attacks" that caused emotional harm to their members (Dibbell, 1998; Schwartz, 2001). A group itself can harm its members: cults can exist in cyberspace as well as in the real world. One such, Pro-Ana, extols the joys and personal freedom of anorexia. Its members share tips on how to hide weight loss from family and friends, discuss the importance of personal choice, and praise members' announcements of their weight loss.

Although participating in an online community may not be directly harmful to its members, involved members may withdraw from their physical relationships and responsibilities. People who spend a great deal of time online must be spending less time doing something else. The research thus far has only examined the effects of aggregate number of hours spent online. One study found it was associated with a small decrease in social involvement and psychological well-being for a particular group of new users (Kraut et al., 1998), but that effect was erased with continued use (Kraut et al., in press). Some studies have found it to be associated with an expanded social circle (Katz & Apsden, 1997; Quan y Hasse et al., 2002; Kraut et al., 2002).

Just as members may be harmed by erroneous or malicious information, so too may corporations. Companies may fear liability if advice promulgated in an online community leads to product failure or product-associated damages. Customer complaints can be shared very rapidly with large numbers of people. If accurate, they can snowball rapidly into widespread mobilization. The Intel corporation had to manage a wave of Internet protest, much of it organized through Usenet groups, as it learned about and took steps to correct a flaw in its Pentium processor. Ultimately, of course, fixing the error benefited Intel and its customers (Uzumeri & Snyder, 1996). In a different case, members of a number of Internet groups mobilized to protest the introduction of a new household database product created by the Lotus Development corporation. That protest led to the withdrawal of the planned product (Culnan, 1991). Online communities are not the only means of mobilizing discontent on the Net (e.g., Gurak, 1997), but because they are characterized by member commitment, they can be particularly potent.

Intellectual property infringement is another area of potential harm for corporations. Trademark infringement is easy to spot in many customer communities. Copyright infringement can be particularly troublesome for media property companies. Fan community members create and discuss fan fiction, that is, new story lines or alternate plot endings for their favorite shows, movies, or characters. Corporations routinely issue cease and desist orders against these groups, fearing loss of copyright control (Silberman, 1996). As with mobilizing discontent, online communities are not the only mechanism on the net for intellectual property infringement. Unauthorized media file sharing probably represents a larger area of intellectual property harm for media companies at the moment, but the social reinforcement that is generated when community members praise or critique one another's (arguably intellectual property infringing) creative work can be potent.

RESEARCH METHODS AND ISSUES

Two broadly different research traditions characterize much of the research on online communities. In caricature these can be labeled "insider studies" and "outsider studies." Participant observation studies began with Howard Rheingold's (2000) description of the Well, a Northern California online community begun in 1983. Examples of scholarly ethnographies include those of a soap opera discussion community (Baym, 1999), a social MUD (Kendall, 2002), a lesbian café (Correll, 1995), and an IRC arts community (Danet, 2001). In each case the writer/analyst was a member of the community for an extended period of time. The ethnography evokes the language, personalities, beliefs, interpretations, and daily lives of people inhabiting these worlds.

"Outsider studies" typically extract records of online behavior or attitudes for study outside the community context. Linguists may extract text records for linguistic analysis of online talk (e.g., Herring, 1996). Sociologists may analyze them for norm strength (Sassenberg, 2002). Social psychologists may use questionnaires to survey community members about their social support systems in the online world and the physical world (e.g., Cummings et al., 2002; Mackenna & Bargh, 1998). Sociologists and political scientists may use questionnaires to survey members about their social and civic activities and attitudes (Kavanaugh, 1999; Wellman & Hayathornthwaite, in press).

Online communities are appealing targets for researchers. New or newly visible social phenomena are intrinsically interesting to the social scientist. Moreover online access to one's subjects of study offers beguiling efficiencies. Ethnographers can do ethnographic work without leaving the office. Survey researchers can administer questionnaires to multinational samples without buying a single postage stamp. Linguists have access to entire cultural corpora in digital form. The efficiencies are not problem free, however. Ethnographers have an incomplete picture of their community members' offline lives. Survey researchers often have no access to members who never post (e.g., Nonnecke & Preece, 2000). Linguists do not see private or back channel communication. Still, despite the drawbacks, the past 10 years have seen a substantial growth in social scientists and social science methods oriented toward understanding online communities.

Whereas principles of ethical research are widely shared within the academic social science community (and are governed by federal regulation), procedures for implementing those principles in research on online communities are under debate. Consider just the principles of informed consent and subject anonymity. If a person's words in a community discussion are considered private, consent should be obtained before analyzing them. If they are considered public, consent should be obtained before quoting them (except for fair use). If a researcher plans a participant observation study, she or he should seek permission from community members before beginning the study. Because most communities are open, members who join after the study has begun do not have the same opportunity to give or revoke permission, however. In publishing results, the researcher must honor promises of anonymity by disguising identity. Because full-text search engines are so powerful, verbatim text, even with no identifier, can be used to find the original (identified) source. Bruckman (2002) pointed out that norms in the humanities encourage attribution and reproduction. Several scholarly and professional associations are currently grappling with the ethics of online research (Frankel & Siang, 1996; Thomas, 1996).

CONCLUSION

The Internet has been a home for self-organizing voluntary groups since its inception. As the Net grew, so did the pace of people and technology mutually adapting to form and support new online communities of interest (e.g., Boczkowski, 1999). Despite the large number of online communities and online community members today, the social form has been widespread for less than a decade. The nature of the net means that experimentation and evolution into new variants of the social form can occur rapidly. The next 10 years should see more new community types, new ways of aggregating microcontributions, and new community processes.

The social form is also immature in terms of impact on members and on the offline world, but with a more differentiated view of community types, we should be able to better specify which types of online communities should or could have which kinds of impacts on which types of members. Nevertheless, people live in the physical world. The real online community design payoff may come from supporting online extensions of the places where people live, work, send their kids to school, recreate, vote, and worship. Television has had an enormous impact on family communication patterns, teen culture, political activity, and consumer behavior. Most of the decisions that led to those impacts were made by an extremely small number of wealthy and influential individuals. In online communities, by contrast, everyone has the opportunity to shape the processes that will make a difference.

GLOSSARY

Dot-com Internet sites or businesses designed to make money during the late 1990s.

Internet relay chat (IRC) A program for simultaneous text communication among two or more users.

Listserv A program for managing a distribution list of e-mail addresses.

Multi-user dungeon/domain/dimension (MUD) A text-based virtual reality environment, initially used for fantasy games, now also used for social and professional purposes.

MUD object oriented (MOO) Text-based virtual reality environment in which users can program objects that have persistence.

Seti@home An activity organized over the Net in which people donate idle CPU cycles to process data looking for radio signals.

Usenet A system of electronic bulletin boards.

Virtual reality (VR) A text-based or graphics-based environment that evokes a self-contained world.

CROSS REFERENCES

See *Internet Etiquette (Netiquette); Internet Navigation (Basics, Services, and Portals); Legal, Social and Ethical Issues; Privacy Law; Virtual Reality on the Internet: Collaborative Virtual Reality.*

REFERENCES

Baym, N. (1999). *Tune in, log on: Soaps, fandom, and online community (new media cultures).* Thousand Oaks, CA: Corwin Press.

Benkler, Y. (2002). Coase's penguin, or, Linux and the nature of the firm. *The Yale Law Journal, 112,* 369–446.

Boczkowski, P. J. (1999, Spring). Mutual shaping of users and technologies in a national virtual community. *Journal of Communication,* 86–108.

Bressler, S. E., & Grantham, C. E. (2000). *Communities of commerce.* New York: McGraw-Hill.

Bruckman, A. (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology, 4,* 217–231.

Butler, B. S., Kiesler, S., Kraut, R. E., & Sproull, L. (2002). Community effort in online groups: Who does the work and why? In S. Weisband & L. Atwater (Eds.), *Leadership at a Distance.* Retrieved April 9, 2003, from http://opensource.mit.edu/papers/butler.pdf

Callero, P. L., Howard, J. A., & Piliavin, J. A. (1987). Helping behavior as a role behavior: Disclosing social structure and history on the analysis of prosocial action. *Social Psychology Quarterly, 50,* 247–256.

Carnevale, P. J., & Probst, T. M. (1997). Conflict on the Internet. In S. Kiesler (Ed.), *Culture of the Internet* (pp. 233–255). Mahwah, NJ: Erlbaum.

Castronova, E. (2001). *Virtual worlds: A first-hand account of market and society on the cyberian frontier.* Retrieved January 4, 2003, from http://papers.ssrn.com/abstract=294828

Clary, E. G., Snyder, M., Ridge, R. D., Copeland, J., Stukas, A. A., Haugen, J., & Miene, P. (1998). Understanding and assessing the motivations of volunteers: A functional approach. *Journal of Personality and Social Psychology, 74,* 1516–1530.

Correll, S. (1995). The ethnography of an electronic bar: The lesbian café. *Journal of Contemporary Ethnography, 24,* 270–298.

Cothrel, J. (2001). *Measuring the success of online communities.* Retrieved October 27, 2002, from http://www.participate.com/research/art-measuresuccess.asp

Cramton, C. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science, 12,* 346–371.

Culnan, M. J. (1991). *The lessons of Lotus Marketplace: Implications for consumer privacy in the 1990s.* Paper presented at the First Conference on Computers, Freedom and Privacy. Retrieved October 27, 2002, from http://www.cpsr.org/conferences/cfp91/culnan.html

Cummings, J. Sproull, L., & Kiesler, S. (2002). Beyond hearing: Where real world and online support meet. *Group Dynamics: Theory, Research, and Practice, 6,* 78–88.

Danet, B. (2001). *Cyberpl@y: Communicating online.* Oxford, England: Berg; distributed in the U.S. by NYU Press.

Dibbell, J. (1998). *My tiny life: Crime and passion in a virtual world.* New York: Holt.

Etzioni, A., & Etzioni, O. (1999). Fact-to-face and computer-mediated communities, a comparative analysis. *The Information Society, 15,* 241–248.

Festinger, L., Schachter, S., & Back, K. (1950). *Social pressures in informal groups.* New York: Harper.

Figallo, C. (1998). *Hosting Web communities: Building relationships, increasing customer loyalty, and maintaining a competitive edge.* New York: Wiley.

Finholt, T., & Sproull, L. (1990). Electronic groups at work. *Organization Science, 1,* 41–64.

Foster, E. (1999). Best technical support: It may not be the guy on the telephones any more. *InfoWorld.* Retrieved from http://www.infoworld.com/articles/op/xml/99/11/29/991129opfoster.xml

Frankel, M. S., & Siang, S. (1999). Ethical and legal aspects of human subjects research on the Internet: A report of a workshop. Retrieved from http://www.aaas.org/spp/dspp/sfrl/projects/intres/main.htm

Galegher, J., Sproull, L., & Kiesler, S. (1998). Legitimacy, authority, and community in electronic support groups. *Written Communication, 15,* 493–530.

Gray, J. E., Safran, C., Davis, R. B., Pompilio-Wietzner, G., Stewart, J. E., Zacagnini. L., et al. (2000). Baby Care-Link: Using the Internet and telemedicine to improve care for high-risk infants. *Pediatrics, 106,* 1318–1324.

Gurak, L. J. (1997). *Persuasion and privacy in cyberspace: The online protests over Lotus Marketplace and the Clipper Chip.* New Haven, CT: Yale University Press.

Hagel, J., & Armstrong, A. G. (1997). *Net gain: Expanding markets through virtual communities.* Cambridge, MA: Harvard Business School Press.

Hampton, K. N., & Wellman, B. (1999). Netville online and offline: Observing and surveying a wired suburb. *American Behavioral Scientist, 43,* 475–492.

Herring, S. (1994). Politeness in computer culture: Why women thank and men flame. In M. Bucholtz, A. C. Liang, L. A. Sutton & C. Hines. (Eds.), *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference* (pp. 278–294). Berkeley: Berkeley Women and Language Group, University of California.

Herring, S. (Ed.). (1996). *Computer-mediated communication: Linguistic, social and cross-cultural perspectives.* Amsterdam: John Benjamins.

Internet Engineering Task Force (n.d.). A novice's guide to the Internet Engineering Task Force. Retrieved September 22, 2002, from http://www.ietf.org/tao.html

Katz, J., E., & Aspden, P. (1997). A nation of strangers? *Communications of the ACM, 40,* 81–86.

Kavanaugh, A. (1999, September). *The impact of computer networking on community: A social network analysis approach.* Paper presented at Telecommunications Policy Research Conference, Alexandria, VA.

Kendall, L. (2002). *Hanging out in the virtual pub: Masculinities and relationships online.* Berkeley: University of California Press.

Kiesler, S., & Sproull, L. (1987). *Computing and change on campus.* New York: Cambridge University Press.

Kim, A. J. (2000). *Community building on the Web: Secret strategies for successful online communities.* Berkeley, CA: Peachpit Press.

Knoke, D. (1986). Associations and interest groups. *Annual Review of Sociology, 12,* 1–21.

Kollock, P. (1999). The production of trust in online markets. *Advances in Group Processes, 16,* 99–123.

Kollock, P., & Smith, M. (1996). Managing the virtual commons: Cooperation and conflict in computer communities. In S. C. Herring (Ed.), *Computer mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 226–242). Philadelphia: Benjamins.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist, 53,* 1017–1031.

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues, 58,* 49–74.

Lakhani, K., & von Hippel, E. (2002). *How open source software works: "Free" user-to-user assistance.* Sloan School of Management Working Paper #4117. Cambridge, MA: MIT. Retrieved April 9, 2003, from http://opensource.mit.edu/papers/lakhanivonhippelusersupport.pdf

Licklider, J. C. R., & Veza, A. (1978). Applications of information networks. *IEEE Proceedings, 66,* 1330–1346.

Lorig, K. R., Lorca, K. R., Laurent, D. D., Deyo, R. A., Marnell, M. E., Minor, M. A., et al.. (2002). Can a back pain e-mail discussion group improve health status and lower health care costs? *Archives of Internal Medicine, 162,* 792–796.

Mackenna, K. Y. A., & Bargh, J. A. (1998). Coming out in the age of the Internet: Identity 'de-marginalization' from virtual group participation. *Journal of Personality and Social Psychology, 75,* 681–694.

Mayzlin, D. (2001). *Promotional chat on the Internet.* Unpublished manuscript, Yale University.

Mickelson, K. D. (1997). Seeking social support: Parents in electronic support groups. In S. Kiesler (Ed.), *Culture of the Internet* (pp. 157–178). Mahwah, NJ: Erlbaum.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nonnecke, B., & Preece, J. (2000). *Lurker demographics: Counting the silent. In Computer–Human Interaction 2000* (pp. 73–80). New York: ACM Press.

Omoto, A., & Snyder, M. (1995). Sustained helping without obligation: Motivation, longevity of service, and perceived attitude change among AIDS volunteers. *Journal of Personality and Social Psychology, 68,* 671–687.

Orlikowski, W., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly, 39,* 541–574.

Powazek, D. (2002). *Design for community: The art of connecting real people in virtual places.* Indianapolis, IN: New Riders.

Preece, J. (2000). *Online communities: Designing usability, supporting sociability.* New York: Wiley.

Quan y Hasse, A., Wellman, B., Witte, J. & Hampton, K. (2002). Capitalizing on the Internet: Social contact, civic engagement, and sense of community. In B. Wellman & C. Haythornthwaite (Eds.), *The Internet in Everyday Life,* (pp. 291–394). Oxford: Blackwell.

Raymond, E. (1999). *The cathedral & the bazaar: Musings on Linux and open source by an accidental revolutionary.* Cambridge, MA: O'Reilly.

Reid, E. (1999). Hierarchy and power: Social control in cyberspace. In M. A. Smith & P. Kollock (Eds.), *Communities in Cyberspace* (pp. 107–133). London: Routledge.

Rheingold, H. (2000). *The virtual community* (rev. ed.). Cambridge: MIT Press.

Sacharow, A. (2000). *Consumer-created content: Creating and valuing user-generated programming. Vision Report.* New York: Jupiter Media Metrix.

Sassenberg, K. (2002). Common bond and common identity groups on the Internet; Attachment and normative behavior in on-topic and off-topic chats. *Group Dynamics, 6,* 27–37.

Schuler, D. (1996). *New community networks: Wired for change.* Reading MA: Addison-Wesley.

Schwartz, J. (2001, January 4). Virtual mayhem arouses real anger at hackers' attack. *New York Times.* Retrieved January 4, 2003, from www.nytimes.com/2001/01/04/technology/04HACK.html.

Silberman, S. (1996). Paramount locks phasers on Trek fan sites. *Wired News.* Retrieved February 12, 2001, from http://www.wired.com/news/culture/0,1284,1076,00.html

Sproull, L., & Faraj, S. (1995). Atheism, sex and databases: The net as a social technology. In B. Kahin & J. Keller (Eds.), *Public Access to the Internet* (pp. 62–81). Cambridge, MA: The MIT Press.

Sproull, L., & Kiesler, S. (1991). *Connections: New ways of working in the networked organization.* Cambridge, MA: MIT Press.

Thomas, J. (Ed.). (1996). *The Information Society, 12* [Special issue].

Thomson, R., & Murachver, V. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology, 40,* 193–208.

Thorn, B. K., & Connolly, T. (1987). Discretionary data bases; a theory and some experimental findings. *Communication Research, 14,* 512–528.

Uzumeri, M. V., & Snyder, C. A. (1996). Information technology and accelerated science: The case of the Pentium^{TM} flaw. *California Management Review, 38,* 44–63.

Wallich, P. (2001, September). Mindstorms: Not just a kid's toy. *IEEE Spectrum,* 52–57.

Walsh, J. P., & Bayma, T. (1996). The virtual college: Computer-mediated communication and scientific work. *The Information Society, 12,* 343–363.

Wasko, M. M., & Faraj, S. (2000). "It is what one does": Why people participate and help others in electronic communities of practice. *Journal of Strategic Information Systems, 9,* 155–173.

Wellman, B., & Haythornthwaite, C. (in press). *The Internet in everyday life.* Oxford, England: Blackwell.

Winzelberg, A. (1997). The analysis of an electronic support group for individuals with eating disorders. *Computers in Human Behavior, 13,* 393–407.

Yee, N. (2001). *The Norrathian scrolls: A study of Ever Quest.* Retrieved from http://www.nickyee.com/eqt/report.html

Online Dispute Resolution

Alan Gaitenby, *University of Massachusetts, Amherst*

INTRODUCTION

The emergence of the Internet and myriad online social and commercial interactions has spawned another realm of human disputing. Information technology entrepreneurs applying alternative dispute resolution (ADR) practices and theories created online dispute resolution (ODR) to manage and ameliorate ever-increasing online conflict.

Need for ODR

Online interaction is the source of conflict to which ODR is designed to respond. Online interactions include commerce, group identification and maintenance, and conducting personal or professional relationships. Commercial transactions produce many online disputes in need of resolution and have been subject to concerted ODR development and marketing efforts (Rule, 2002). Conflict in commerce is costly, thus amelioration through open and fair terms of trade and dispute resolution arising from that trade are critically important and desirable. There are noncommercial interactions as well, such as discussions and e-mail exchanges in chat spaces or message boards. Conflict erupts in all of these spaces.

Most online conflict is resolved via informal means such as direct negotiation. Some conflict necessitates more formal responses however. ODR is part of that response, sharing dispute resolution responsibilities with positive law, as well as filling niches where that law's reach is challenged. ODR has become the default for managing certain forms of online conflict because of its "fit" with those disputes (i.e., both the dispute and the resolution are manifested online). ODR also offers cost advantages over traditional remedies, the delivery of dispute resolution expertise at a distance, and the convenience of settling disputes from the comfort of an individual's computer.

According to Henry H. Perritt, Jr., "three characteristics of the Internet make traditional dispute resolution through administrative agency and judicial procedure unsatisfactory for many controversies that arise in Internet-based commerce and political interaction." First, "the Internet's low barrier to entry invites participation in commerce and politics by small entities and individuals who cannot afford direct participation in many traditional market and political arenas." Second, "the geographic openness of electronic commerce makes stranger-to-stranger transactions more likely." Third, "the Internet is inherently global" (Perritt, 2000, p. 675). ODR suggests a synergistic fit with particular online interactions and contexts, as well as recognizing significant concerns with the practical effectiveness of state sponsored legal remedies (Bordone, 1998). Online transactions with geographically disparate participants, especially transactions of modest value, would be best served by online dispute resolution. ODR offers convenience because participants do not have to meet in person; dispute resolution expertise is provided to their own personal computers (PCs). ODR is also well suited to conflict management processes marked by repetitive and relatively discrete information exchanges.

Despite national and international property, commercial and contract law questions persist about whether that law ought, or feasibly can, reach all online interactions and transactions. This is exacerbated when the participants are at distant locations, and the service providers for Internet connections or applications are also at a distance. An online transaction gone awry between a business and individual located in a single country poses a considerably different legal and practical challenge than between parties in different countries. Cost structures also mitigate favorably to ODR over traditional law when a transaction crosses considerable distance and numerous political or geographic boundaries, or importantly when the transaction is not of great value but significant enough to one or both parties to pursue resolution.

The practical effectiveness of positive law is challenged by both subject matter and party jurisdiction over cyberspace. Online interaction is relatively new and thus may not yet be covered or described effectively in law's various forms (e.g., cases, codes, and commentary). Additionally, uncertainty as to "location" of online interaction and the reach of political sovereigns into cyberspace pose serious challenges. Legal institutions and actors may be slow to act, and even as political sovereigns seek to exert and define jurisdiction over online activity, much will be unattended unless other mechanisms are developed and employed.

ODR's potential for managing online conflict remains strong because of its fit and particular advantages over traditional legal mechanisms. Where ODR processes are seen to alleviate systemic strain and deliver improved social justice courts and judicial administration will likely adopt ODR much as court sponsored mediation has been incorporated to great success by family and divorce courts in the United States.

Nature of ODR

ODR refers to alternative dispute resolution processes implemented largely using Web-based information management and communication and in some cases database applications. ODR relies on the Web and users' abilities to access sites and submit or download information from the Internet. Because of the relatively rapid diffusion of networked information technology, ODR has the potential to increase access to conflict management dramatically among people who would otherwise forego third-party facilitation. Whether litigation or ADR there are physical variables to be addressed which ODR transcends (Kassedjian & Cahn, 1998). ODR lowers barriers to entry by reaching into homes, offices, schools, and halls of government all over the globe. ODR however presents other obstacles, most significantly the need for networked information technology and bridging the digital divide.

ODR is described as a "Fourth Party," supplementing traditional ADR third party facilitators or neutrals critical to resolving disputes outside of traditional litigation (Katsh & Rifkin, 2001). ODR systems tend to be dedicated to particular dispute resolution models such as mediation, arbitration, and negotiation. Two prominent examples found niches providing dispute resolution to immense online marketplaces like eBay, as well as relatively small but valuable exchanges like ICANN domain names registration. ODR is part of the developing law in these contexts, law that is directly associated with the marketplace or auction, operating in the shadow of ADR and positive law itself (e.g., "eBay law"; Katsh, Rifkin, & Gaitenby, 2000).

There are two basic ODR approaches: dispute avoidance (DA) and dispute resolution (DR). Dispute avoidance is informal ODR, where information technology is employed to lubricate online social interaction so that fewer full-blown disputes erupt that necessitate formal resolution. DA takes advantage of the fact that online information exchanges, regardless of social setting or context, are considerably less robust than offline exchanges.

Primary DA technologies include credit card chargebacks and online money transfer, payment, or escrow services facilitating a wide range of online transactions from marketplaces such as LLBean.com to free-wheeling auction sites such as eBay.com. Chargebacks are debt erasures for consumers by credit card companies (e.g., Citibank, American Express). Chargebacks occur when consumers wish to challenge or reverse a transaction they made with a credit card and are unable to otherwise get satisfaction from the merchant. Chargebacks are popular with consumers because financial liability is largely removed from them and transferred to merchants (i.e., the benefit of the doubt rests with the consumer).

Payment systems are relatively simple and help reduce some transaction costs associated with online markets (e.g., seller having to wait until a money order arrives via "snail mail" before sending the purchase). Third-party payment systems (e.g., Paypal.com) maintain accounts, much like bank accounts, in which users make deposits via check, money order, or electronic fund transfer, although these accounts are not regulated or ensured as banks. Users maintain balances, but rather than accessing the money to conduct transactions, the money transfer service facilitates the deal after being notified that a transaction has been requested and authorized. Money transfer systems then either cut a check, credit a credit card or bank account, or adjust an interest-earning balance for users' accounts.

Other DA mechanisms provide consumer and seller information profiles derived from data about transaction-related behavior, these profiles are a track record or reputation that can be incorporated into decision making by potential buyers or sellers. The eBay feedback rating system is a prime example of such an information database readily available to buyers and sellers alike. eBay buyers and sellers are ranked by those with whom they do business, and eBay then computes overall positive and negative ratings upon which decisions to trade may be tied.

Trustmarks and trustseals are other prominent DA applications. Trustmarks and seals are protocols or standards of good practice for online commerce, privacy, and dispute resolution. Online services and enterprises capitulate to the standards and are accredited by a trustmark or trustseal service provider (e.g., TrustE, SquareTrade). Online services or enterprises pay for the mark or seal to appear on their site to verify accreditation and satisfaction of promulgated standards. Trustmarks and seals are important to ODR not only because they help parties avoid disputes in the first place but also because they stipulate standards of dispute resolution.

EVOLUTION OF ODR FROM ADR

ADR Primer

The groundwork for ODR was laid with the establishment of alternative dispute resolution (ADR) in the latter half of the 20th century (Auerbach, 1983). ADR represents a collection of professional practices and theories designed to resolve disputes outside of courts. ADR is largely voluntary, comprising direct or facilitated negotiation, mediation by a neutral party with variable levels of power in structuring the exchange and in solution crafting, binding or nonbinding arbitration with a neutral decision maker, and decision support practices.

The adversarial model of courts and litigation came under scrutiny during the 1960s and 1970s because it was perceived to be less responsive to evolving needs for social justice, especially as that relates to the maintenance or reparation of damaged, and highly valued, relationships (Fuller, 1978). Litigation may consume considerable resources and time, because courts tend to be busy and the lawyers who access them costly. Thus, powerful practical disincentives exist to adversarial dispute resolution in addition to an ideological opposition to a one-size-fits-all, win–lose proposition. Substance and process concerns with litigation led to the development and ultimate acceptance of ADR (Fuller, 1978).

The adversarial model's formalisms of rights, advocates, neutrals, openness, due process, and enforcement of remedies, however, remain powerful incentives for litigation. ADR can adhere to some of these formalisms (e.g., binding arbitration's litigation-like qualities), but enforcement of remedies is the source of much criticism for ADR and now ODR. Because the voluntary nature of mediation and negotiation, ADR must depend on more than force to be effective. ADR appeals to disputants' desire to repair relationships as well as seek recompense for a given act or transgression. ADR's "teeth" reside within the desires of participants themselves and the potential of various ADR processes to be applied to facilitate achieving those desires or next-best alternatives.

ADR offers opportunities for so-called win–win solutions and plays heavily on its cost advantages. Litigation takes time and money, thus a strong incentive to settle quickly plays to ADR's favor. Not all ADR is without force however; arbitration can be binding and recognized as law through statutes (e.g., U.S. Federal Arbitration Act, 2002) or transnational conventions (e.g., The New York Convention on the Recognition and Enforcement of Foreign Arbitral Awards; http://www.uncitral.org/en-index.htm). Binding arbitration is increasingly part of professional service contracts stipulating that as disputes arise they must be settled through binding arbitration.

ADR moves the power in the dispute resolution setting to the participants and away from some external authority or the process itself. Most significant in terms of substantive justice is the flexibility in outcomes and the eschewing of the standard win–lose dichotomy.

> ADR has proven that moving justice away from the courthouse is often desirable and that the arena of dispute resolution, once thought to be the exclusive domain of law and courts, is markedly different from what it was several decades ago. Mediation arbitration, and other forms of "alternative" dispute resolution are now the most common approaches to dealing with conflict. (Katsh & Rifkin, 2001, p. 27)

ODR: An Early History

Although practitioners and scholars of ADR began using personal computers in the 1980s, e-mail in the early 1990s, and the Internet by the middle of the 1990s, ODR systems did not emerge until the Web became ubiquitous. The Web was the innovation that spurred many to recognize the potential for commerce and other forms of social interaction in cyberspace. Entrepreneurs, scholars, lawyers, and policy makers realized that all this interaction included the propensity for conflict, and that had a damping effect on the vitality and value of online interaction. ODR's early history has had three distinct periods between the years 1990 and 2002. This ends with the Internet "bubble" bursting in 2000–01 and the adverse impact on the fledgling ODR field, a challenge that gave way to ODR's push to maturity, marked by product differentiation, niche filling, and innovative entrepreneurial dispute avoidance tools.

Evolution of Cyberspace and Online Disputes: 1990–1995

Before 1995 there was no universal online standard like the Internet. Rather, there were several key activities, supported by software and a protocol to use the Internet. E-mail and finding and moving files through textual interfaces required that users know some operating system command language. Interfaces between users and the Internet were sufficiently complicated to slow Internet use, and access and use was concentrated among academics, scientists, researchers, and select others who did not encompass a broad social cross-section. Dial-up bulletin board and listserv services, moderated e-mail lists to which users subscribed, were the most active areas of cyberspace then. At the far edge of cyberspace in those days were the pioneers of virtual realities and communities, the "gods," "wizards," and everyday users of multiuser object-oriented realities (MOOs) and multiuser dungeons (MUDs; Gaitenby, 1996).

Listservs were common asynchronous communication protocols, sustaining all variety of conversations. Listservs were sometimes moderated, but more likely were left alone; in both instances discursive norms were developed formally and informally to structure textual interaction. In a moderated list, authority was more overt and terms of use more explicit. Unmoderated lists, on the other hand, were laboratories for self-governance and indigenous normative development. In these early days of cyberspace, flaming was the most common form of dispute to develop in a listserv, often occasioned by a textual bombardment against a particular user, viewpoint, or both. Flaming could result in the shutting down of the listserv because levels of acrimonious traffic have a significant damping effect on online discourse.

Listserv disputes were resolved either by moderator interjection or other authoritative action (e.g., active content management, issuance of warnings to cease, evict participants). Without moderators, disputes were handled informally but also with the adoption of specific conduct norms by consensus or other decision-making method. Such methods could, however, be utterly undemocratic and unfair because there were rarely process or first-order rules to enable the creation of substantive or second order rules. Even in the absence of moderation, listserv authority and management ultimately rested with system administration, the final authority in resolving disputes was the power to shut down a board, a user, or a specific conversation.

More ambitious than listserv technology and practices were MOOs and MUDs, those online textual realities whose players immersed themselves in complicated role-playing games as well as thematically oriented social environments. In most of these environments, users adopted new identities and characteristics. System administrators were more fully vested in the rule structure and dispute resolution mechanisms than in listservs where they filled a more mechanical role. In MOOs and MUDs, system administrators were the "gods," and "wizards" of the universes they were creating and inhabiting, and it was to them that dispute resolution, or at least decision enforcement, ultimately fell.

During 1993, in the virtual reality LambdaMOO, a clever user created a program that allowed him to manipulate the online characters of other users. Manipulation appeared as all other actions in these spaces, as a textual expression, which all users could "see." Characters so afflicted were "forced" to perform self-mutilation or sexual assault upon themselves and characters of other users. A number of LambdaMOO users called for the admitted perpetrator's removal. After a public forum called by general consensus to address the event and its ramifications, the perpetrator was banished by a wizard. Many present felt the removal was a violation of fair process; most, however, felt he had clearly violated a fundamental norm of conduct. Conflict and its management had come fully to cyberspace.

Emergence of ODR: 1995–1998

With faster and cleaner Internet connections, more powerful microprocessors, hypertext markup language editors, and browsers, the Internet became a mass-cultural phenomenon after 1995. As more people went online and the complexity of social interaction increased, so, too, did the likelihood for conflict. Not surprisingly, conflict came readily to that place where valued exchange was most important online during these relatively early days, the provision of an interface by Internet service providers (ISPs).

ISPs hold the keys to Web access for many individual users. Their relationships with users are typically long-standing and have potential for conflict. ISPs provide an uplink to the Internet and often supply applications that allow users to move files and communicate. With those keys come expectations about terms of use, and not far behind is the law, constructing things like liability and responsibility for both user and service provider. The primary point of contention that arose between ISPs and users concerned ISP liability for users' actions. Specifically, ISPs were worried that they would suffer contributory liability for such actions as libelous speech or copyright violation by users.

In 1995, the Virtual Magistrate online arbitration system was established through which ISP disputes could be handled (Post, 1996). Virtual Magistrate, and later the Online Ombuds at the University of Massachusetts and the Family Mediation Project at the University of Maryland, were implemented with support from the National Center for Automated Information Research. Although conflict arising online was the impetus for these developments, applying information technology in dispute resolution more generally was the prime objective. ODR's second developmental period was dominated by two trends, the rapid proliferation of online activity and technologies to support it and the emergence of ODR prototypes from academic institutions.

Entrepreneurial Drive and Burst Bubbles: 1998–Present

Internet entrepreneurs and ADR professionals took the lead as the Web was more vigorously used in commerce and other interactions. Two major features of this period were significant growth in number and diversity of ODR enterprises, and an increased recognition by policy makers that online dispute resolution was needed to promote efficiency and vitality in evolving online markets. From 1995 to 1998, the field grew from a handful of academic-sponsored prototypes to roughly four dozen mostly private entrepreneurial efforts.

Cross-border ecommerce was the primary impetus for government attention to ODR. By 1999, the U.S. Federal Trade Commission and the Department of Commerce formally addressed ODR as an important element to enhanced consumer protection and willingness to trade online. U.S. efforts largely focused on providing incentives and support to private ODR initiatives rather than governmental dispute resolution or command and control regulatory mechanisms. In Europe, the concern for privacy protection in ecommerce were paramount and spurred the adoption of the SafeHarbor protocol, which required privacy protection and grievance handling processes for any ecommerce entity wishing to do business in Europe.

During this period ADR practitioners teamed with Internet entrepreneurs to produce online versions of offline ADR practices in the belief that aside from managing conflict arising online, that ODR could be used to manage many forms of offline conflict as well. These pioneers felt that information technology could be leveraged to enhance the provision of nonadversarial means of conflict resolution regardless of context.

ODR systems were created for specific online marketplaces, in the business-to-business (B2B) context much of this was dispute avoidance through closed-network electronic data interchanges which provided structure and consistency to a variety of online transactions between members of a trading consortium or other private marketplace. In more public consumer marketplaces, such as business-to-consumer (B2C) or consumer-to-consumer (C2C) marketplaces, ODR systems were created to provide direct negotiation or facilitated mediation/arbitration services.

The Internet shakeout of 2000–02 had a dramatic impact on ODR. Many efforts failed or went dormant, retrenchment left a few core areas of activity in auctions, domain names, and insurance claim settlement. Offline disputants have not adopted full-fledged ODR in any meaningful way and likely will not until more ADR practitioners, lawyers, and laypeople are made aware of ODR's advantages and utility. ADR practitioners and lawyers to a lesser degree use e-mail, the Web, and other information technology (IT) tools to enhance performance of their traditional practices.

NEED AND NATURE OF ODR

Domain of Online Dispute Resolution

Many disputes are avoided through possessing information about social environments, understanding the range of possible interactions, the rules and norms structuring them, and the potential participants. In online contexts, code is used to ameliorate conflict by shaping environments and potential interactions, as well as supplying desired information about goods, markets, players, and rules (Lessig, 1999). Avoidance is not perfect or always desirable, thus as the complexity and frequency of online interaction increases disputes occur more readily.

Online disputes require alternative dispute resolution as well as traditional legal remedies. Cost, convenience, and efficiency advantages, in addition to challenges facing traditional litigation, mitigates favorably for ODR. Online interaction falls into commercial and noncommercial categories. Some disputes arising from them are subject to formal legal attention, a large portion are settled informally, and a growing pool is served by ODR.

Online Commerce: B2B, B2C, and C2C

Commercial transactions online produce disputes between businesses (B2B), businesses and consumers (B2C), and between consumers (C2C). These disputes vary greatly in terms of value, location, and rule structures. Online commercial disputes erupt for many of the same reasons as offline commercial disputes: a transaction was unsatisfactory, and one or both parties wish to achieve recompense. Online commerce and business generally is critical proving ground for ODR (Rule, 2002).

B2B online transactions are largely managed through private contracts; for example, an online trading consortium may bring together suppliers and buyers for a particular market. Contracts bind the players together, defining roles and responsibilities, and providing for ready-made dispute resolution mechanism. ODR may yet be a primary dispute resolution method in B2B online disputes as stipulated in these contracts; at present, however, ODR is not significantly relied on. "ODR is not used to any meaningful degree in the B2B market segment since the parties have made other arrangements for the settlement of disputes between them and disputes among them are rare in any case" (American Bar Association Task Force on Electronic Commerce, 2002, p. 15).

B2C online transactions (e.g., a purchase at LLBean.com or an online transaction with a National Association of Securities Dealers member) are regulated by terms of sale, contracts, and consumer law more generally. Most disputes from B2C transactions are settled relatively painlessly under terms of sale; those unsuccessful few in which the relative value of the dispute is significant enough may be settled via consumer protection law or other legal mechanism. Cross-border transactions may complicate this somewhat, but contract and commercial law have been globalized especially effectively through transnational agreements such as the General Agreement on Tariffs and Trade.

C2C online transactions, on the other hand, have provided especially fertile ground for online disputes and their management with ODR. Online auctions (e.g., eBay.com) are the primary manifestation of C2C transactions causing online disputes. Auction disputes are straightforward: either the seller or buyer was dissatisfied with the transaction and seeks alteration of the conditions of trade. Most disputes are buyer initiated, with the purchase not satisfying expectations, and thus the buyer seeks recompense. If a seller declines to negotiate, or disputes the claim altogether, direct negotiation and facilitated mediation ODR may be employed. Applying ODR does not alter rights of access to formal legal processes and remedy.

Domain Names

At the root of all online interaction lies Internet symbolic addressing through domain names. The Internet Corporation for Assigned Names and Numbers (ICANN) exclusively controls domain name registration due to a U.S. government contract subject to periodic renewal and congressional oversight. Symbolic addressing of Internet numeric addresses has a significant value for many organizations and individuals and thus results in particularly vested disputants.

Domain name registration disputes arise when two or more parties assert property interests in similar names or marks, or one party challenges the registration and use of a domain name as being in "bad faith" as defined by the Uniform Dispute Resolution Policy (UDRP) of ICANN. Traditional mark law takes political and cultural boundaries, physical distance, and market distinctiveness into account when evaluating claims of infringement. Thus similar names or marks for two or more distinct applications (e.g., Apple Computer and Apple Records) would likely not be in conflict. Geographic and market distinctions do not hold online in the same fashion and disputes erupt between parties all over the globe as they scramble to lay claim to valued domain names.

Message Boards, Chat Spaces, Listservs, E-mail

Disputes arise within the myriad online discussion forums. Discussion forums are e-mail or Web facilitated, sustaining both synchronous and asynchronous communication. Communication may be private (e.g., a one-to-one e-mail) or public (e.g., an individual posting on a widely read message board). Disputes occur readily, and private ad hoc means are employed ranging from cessation of relationship to one party "giving in" to repair the damage done. In forums such as message boards, however, the very public nature of the experience makes private and informal dispute resolution difficult.

Flaming and overly vigorous or vitriolic debate continue to be the primary dispute profile in discussion forums. In these cases, one or more forum participants post messages or send e-mails that may threaten, libel, besmirch, or simply annoy other forum participants. Most flaming is can be considered nuisance speech, and sometimes a threat threshold is breached; largely, however, it is just unpleasant. The challenge has been twofold: defining the limits for a given discussion forum and establishing a mechanism for managing events when limits are said to be breached or threatened.

Flamed speech in these contexts tends to drown out other speech and quickly squelches participants' desire to be a part of the discussion. The exit option tends to be the most readily available and practical for discussion forum participants (Hirschman, 1970). In this case, the exit option is a function of the continued discretionary nature of participation in these spaces and the profound challenge of managing discourse without stifling debate.

Exit may not be the most attractive option for long. Increasingly discussion forums are being created and used for a wider range of subjects and applications, becoming valuable to users for important professional and personal information exchanges. A particularly vibrant area of online discourse has been forums associated with social and political interests and movements, forums that enable geographically disparate individuals to come together virtually for a variety of purposes.

NativeWeb.org, the foremost online resource and virtual community for and about indigenous peoples, provided resources and expertise at a distance while establishing an infrastructure for the development of numerous discussion forums. NativeWeb ran many discussion forums and message boards and also provided Web site hosting to support more discussion forums and message boards. Between 1998 and 2001, many forums and boards experienced arguments and flames over issues of native peoples' identity, cultural appropriation, misrepresentation, and fraud. Those targeted or implicated were fearful that online disputes would seriously affect offline life (i.e., livelihood and physical well-being). In the face of ongoing discussion forum disputes many participants simply dropped out, taking the exit option was the ultimate dispute resolution method. NativeWeb followed by shutting down all their discussion forums by 2002.

Claims Settlement

Insurance claims settlement exemplifies offline interactions lending themselves to online dispute resolution that may be cost-effective and desirable. When an automobile accident produces a dispute, the resolution of that dispute (i.e. the filing, negotiating, and agreement on settlement) presents a unique opportunity for ODR. Insurance claims settlement is a discrete and repetitive negotiation process that can be reduced to a set of defined information exchanges that can be moved online relatively easily. By avoiding physical meetings for negotiations the cost structure mitigates favorably for ODR.

ODR Building Blocks

ODR depends on the use of networked information technology, beyond that ODR is predicated on three essential building blocks: trust, expertise, and convenience (Katsh & Rifkin, 2001). ODR is largely voluntary, and therefore it must rely on its contextual fit, attractiveness, ease of use, and effectiveness.

Trust

ODR is about trust, according to ODR pioneer and scholars Katsh and Rifkin (2001). Large-scale online commerce is constrained by a trust gap, some of which is visceral. Individuals and organizations have a disinclination to trade without certain firsthand interactions as well as secondhand accounts. Most of the gap, however, is derived from risk appraisal that evaluates the possibilities for reversing an unsatisfactory transaction in addition to potential trading partners.

Contracts have long been used to reduce risk by filling the trust gap with reliance on third-party adjudication and enforcement. Trust is an interpretation or perception of security, a weighting of the likelihood that object and compensation will be exchanged smoothly. Courts and ADR are important in bridging the trust gap offline so individuals will interact with relative unknowns. In cyberspace, the trust gap between potential buyers and sellers is a significant barrier to interaction. ODR addresses those deficits, providing potential alternatives to litigation as well as assuaging some doubt regarding trading partners.

In ODR, participants, or prospective participants, must also trust that the process is fair and secure and that the likelihood of resolution reasonable. Fairness in dispute resolution requires neutral facilitation, process openness, and opportunity for hearing. Security is a key concern with online commercial exchanges and dispute resolution. Users must trust that the systems and processes are not easily exposed and that private financial or personal information is not rendered public or disclosed inappropriately.

Expertise

Expertise is closely related to trust. Prospective ODR users, and dispute resolution participants generally, are concerned with the quality of the system and process, requiring the presentation of expertise in the fourth-party environment as well as in third-party ADR practices themselves. When an individual or organization takes a claim to court, there are expectations about professional expertise in those who articulate, contest, and adjudicate cases. ADR has also had to build up its stock of experts, practitioners and scholars who constitute the profession and represent an accumulated knowledge base. Expertise is built into ODR software systems, for instance, structuring communication and management of information in direct negotiation systems or settlement algorithms in blind bidding. ODR is relatively new, however, and its stock of experts and knowledge base is just developing. ODR takes much from ADR to be sure, but ODR also necessitates new skills and practices specific to the online contexts.

Convenience

Convenience is ODR's major and most obvious advantage over formal litigation. The ease and reduction of physical costs associated with dispute resolution delivered by IT networks favors ODR. Additionally, online interaction occurs between parties often at considerable distance, when disputes arise from those interactions it is clearly more convenient to pursue resolution online than filing a physical claim in a distant court.

ODR's Manifestations

ODR's current manifestations are direct negotiation and myriad hybrid systems. Direct negotiation ODR is most often formulated as blind bidding claims settlement and is especially well suited to discrete disputes over claims in insurance settlement proceedings. Hybrid ODR covers a fairly wide spectrum of dispute resolution models. Some utilize a mix of online and offline procedures. Hybrids may be stand-alone ODR services such as mediation or arbitration with no offline component, or they may be systems employing online and offline components. At the cutting edge are decision support, collaboration, and multivariable negotiation systems that aggressively leverage the interactive and computational power of information technology to produce greater satisfaction levels for participants.

Most ODR service providers employ similar technical means despite applying different dispute resolution models. All ODR service providers can be found on the Web, some are immediately associated with particular online contexts or marketplaces (e.g., SquareTrade.com and eBay.com). ODR cases commence when individuals or organizations register with an ODR service provider and file claims against individuals, groups, or organizations. Registration and filing may be all online or have mixed online–offline components. Service providers process claims by soliciting participation from the noncomplaining disputant through e-mail, phone, fax, or snail mail. Once respondent agrees to participate, the ODR service provider may use mediation, arbitration, or direct negotiation via e-mail, video conferencing, chat, a variety of Web communication, document handling, and database applications. As with ADR, offline ODR strives to provide a structure for solution development, relationship reparation, and ultimate dispute disengagement.

Although direct negotiation may be completely automated in some instances and employ solution determination algorithms, other ODR models use human facilitators to promote participation, communication, and crafting of solutions. Service providers are also using IT applications (e.g., SmartSettle.com) to help formulate solutions that maximize satisfaction for disputants in complex multivariable negotiations. The population of ODR service providers is fluid; for updated list of service providers visit the Center for Information Technology and Dispute Resolution (http://www.umass.edu/dispute).

Blind Bidding

Blind-bidding direct negotiation is most effectively applied to insurance claims cases. Blind-bidding structures an iterative communication of blind demands and offers with on-the-fly calculations to determine whether the case specific settlement barrier as been breached. In 2002, CyberSettle.com and Clicknsettle.com are examples of operative blind-bidding systems. Dispute resolution in blind bidding is reduced to a simple form: a claim for reimbursement or damages is made against a party, an acceptance or attribution of some responsibility by the party against which the claim is made, an opportunity to present relevant information, and then an iterative process of blind demand and blind offer until settlement or impasse. Blind-bidding systems apply split-the-difference algorithms to determine settlement or impasse; never do participants know exact demands or offers, only that they are requested to submit another or resolution has been achieved. Participants create a settlement or contract to ensure enforcement of remedy; if, however, a party wishes to further the contest, more traditional legal options (i.e., civil suits) are available, and their relative positions have not been exposed.

Direct negotiation is also applied by ODR hybrids. Many online disputes, especially those deriving from auctions, are straightforward, with a small domain of disputes and solutions, and nearly all relationships are one time, discrete events (Katsh et al., 2000). SquareTrade.com, following a facilitated mediation pilot project by the Center for Information Technology and Dispute Resolution (Katsh et al., 2000), established a direct negotiation ODR system for eBay, now handling 80% of auction-related disputes (Katsh, 2002) and overseeing successful resolution of more than two thirds of those. According to SquareTrade, on a given day in August 2002, they received 1,000 complaints. Online direct negotiation systems such as SquareTrade.com are Web and e-mail based, with claims and responses filed via Web interface, and then exchanges are facilitated through an e-mail message scheduling database and routing software. This form of direct negotiation requires no human facilitation by SquareTrade.com; for other online disputes, however, simply enhancing communication is not going to be enough to settle conflict.

ODR Hybrids

Although many disputes can be reduced to a mathematical negotiation over a single variable or can be settled simply by enhancing communication between disputing parties, some online disputes will benefit from third-party facilitation in an ODR fourth-party environment. Hybrid ODR systems may be entirely online or employ a mix of online or offline services to cover a broad range of disputes. Hybrids are largely modeled on offline experiences with mediators, arbitrators, or other dispute resolution facilitators including judges, jurors, and lawyers. ODR hybrids are generally designed to provide these experiences while reducing the difficulties associated with face to face interaction such as travel, heated emotions, bias, and power inequities. Hybrids also leverage information technology for complex negotiations or collaborations across physical and temporal space, incorporating multiple participants and interests.

Facilitated Mediation

Direct negotiation in online auction disputes has been utilized to a significant degree; when that fails, however, the dispute resolution service provider—SquareTrade, in the case of eBay—is ready with facilitated mediation options. Using a Web interface and online mediators, most of whom are also offline mediators, ODR service providers have developed relatively simple fourth-party environments for voluntary auction-related dispute mediation.

Claims and responses to online auction claims are filed via the Web; e-mail is used initially to connect mediator with disputants. Mediation consists of a series of communication exchanges at the ODR mediation case Web site. Online mediators employ techniques similar to their offline counterparts, except without much of the informal information exchange taken for granted in physical proximity (e.g., body language, tone of voice, eye contact). After filing a claim and original narrative, online mediators may solicit each participant to expand on elements of their stories; mediators then can recraft somewhat in attempts to find common ground. Disputes arising from online auctions present limited dispute profiles, and thus mediators depend on understanding what is really of value to the disputants, as well as the most likely place for movement. Repetition has allowed SquareTrade's online mediators to develop a knowledge base and a set of practices to manage auction-related disputes.

For all its parsimony and apparent success, facilitated mediation is not binding in the larger legal context, although auction sites may deny participation rights to disputants who do not live up to terms of settlements. Facilitated mediation has had success, but it has been limited to a narrow class of online disputes; that class, however, has many members. Although not binding, facilitated mediation works because of cost and fit advantages; legal remedies are generally considered too costly and cumbersome for the bulk of online auction disputes.

Arbitration

ICANN's UDRP "arbitrations" are designed to settle disputes over Internet domain name registration quickly and cheaply. Domain name registration services, approved and licensed by ICANN, require domain name registrants to sign a service agreement contract which stipulates that any claims arising against a domain name registrant will be processed through UDRP arbitration. Several ODR service providers make available hybrid arbitration processes for these disputes: World Intellectual Property Organization, National Arbitration Forum, and Center for Public Resources, and the Asian Domain Name Dispute Resolution Center.

Arbitration service providers use a mix of online and offline procedures for filing claims and associated documents (i.e., e-mail, file transfer protocols, fax, phone, snail mail, courier). Claimants choose a service provider, submit a claim, and wait for the service provider to approach the other party about participation in the arbitration. Arbitrators view the provided claims and documentation, may request supplementary information, and make decisions based on UDRP "bad faith" rules (http://www.icann.org/udrp/). Arbitrators and disputants use an array of communication types but never have a physical meeting or session. Arbitrators make decisions and submit opinions via their respective service provider to ICANN, which makes it publicly available.

Service providers adhere to standards as understood in the U.S. Federal Arbitration Act and the New York Convention, despite the fact that domain name dispute resolution proceedings are not considered arbitrations under those existing laws. UDRP arbitrations are nonbinding according to both the UDRP and developing case law. In *Sallen v. Corinthians Licenciamentos* (2001), a federal appellate court determined that U.S. courts do have jurisdiction over domain name. In *Parisi v. Netlearning, Inc.* (2001), the court determined that UDRP arbitration decisions are subject to federal appellate jurisdiction. Litigation will play a significant role in how online arbitration ultimately takes shape. Policy and academic actors have roles to play as well, helping to define the terms of debate and interpreting or implementing what courts find and order (Hill, 1999).

Criticisms of ICANN dispute resolution have focused on the propensity of trademark holders to prevail in domain name arbitrations. Additionally, some service providers seemed to lean more heavily toward trademark holders. Such findings have led to perceptions of forum shopping by disputants and catering to a class of disputants by service providers (Geist, 2002; Mueller, 2000).

Decision Support and Multivariable Negotiation

Another class of ODR hybrids is decision support and multivariable negotiation systems (e.g., SmartSettle.com), which aggressively leverage computational information technology as much as information management and communication (Thiessen, 2000). Decision support and multi-variable negotiation is the most robust of the ODR models; third-party facilitators take active roles in assisting participants to identify, quantify, and adjust preferences in robust fourth-party environments. Users break disputes or negotiations into constituent operative variables that are given discrete value ranges and represent a mix of interests that can be adjusted during negotiations. These systems allow parties to adjust variables as they wish viewing the resulting satisfaction index and then propose a new mix.

Virtual Courts and Collaborative Platforms

Beyond mediation, arbitration, and negotiation via ODR, there are a variety of approaches to conflict management. Virtual courts emulate elements of real courts and their processes; I-courthouse.com is an example of the adversarial model in online dispute resolution. Cases may originate offline or online, and disputants are essentially provided an opportunity to present a narrative claim and supporting evidence in a fourth-party environment. Virtual courts are Web accessed and typically sustain rudimentary document management. Sharing that environment are jurors who act as third parties who may query disputants individually and then render nonbinding decisions. Findings are not binding unless a contractual agreement results.

At the opposite end of the spectrum are service providers that bring individuals together in fourth-party environments to seek consensus when diverse positions exist on particular issues. Consensus-building collaborative platforms (e.g., GroupMindExpress.com) present users with a variety of tools to determine group positions and individual deviances from them, as well as ascertaining commonalities among a sea of difference. These tools hold potential for ODR preprocesses, especially in large, complicated multivariable negotiations involving numerous interests.

CONCLUSION

Assessments of ODR from participants and observers are similar to offline dispute resolution; victorious disputants tend to have a more favorable overall perception of the experience. The loudest critical assessment has come from the ADR community and paradoxically ODR entrepreneurs themselves. ADR professionals are concerned that ODR's lack of face-to-face interaction, a pillar of ADR theory and practice, is necessary for both facilitators and disputants. For facilitators, having disputants in front of them physically allows employment of a range of interpersonal skills and cues. For disputants, especially those in longstanding relationships, face-to-face interaction may be useful for rebuilding trust and helping to move from positions of intransigence toward substantive reconciliation. ODR entrepreneurs, although confident in their products and services, are critical of the market itself, of the apparent difficulties finding a fit for many of them.

Other major concerns include trust gaps, openness or transparency of process balanced with security and privacy of information, lack of standards in training and practice, cultural differences among disputants and facilitators, and language. ODR strives to provide trust to online commercial contexts, but there still is a considerable lack of trust in ODR itself. Some of this comes from concerns with enforceability of remedies, but ADR was able to transcend this, and ODR is also likely to do so. Instead it is the fact that ODR service providers may be relatively unknown, their facilitators hidden behind a digital veil. To bridge that gap, service providers make assurances as to the confidentiality and privacy of participants' ODR experiences in conjunction with open and transparent processes and third-party facilitator profiles. Until ODR is better established, the primary manner to assuage trust gaps resulting from concerns with third-party training and practice standards is to draw online facilitators from the ADR field. English is the dominant ODR language, and alternative dispute resolution theories and practices from North America and Europe also dominate. To counter hegemonic claims and criticisms, ODR service providers will first have to adopt language translation technology or services as they become more cost-effective and robust. Bridging the language divide will provide for creating a discourse on the substance of dispute resolution theories and practices as well as ODR models themselves.

ODR's future appears tied to online commerce in both the short and long term; it is also likely that formal dispute resolution institutions and actors will use ODR as part of a bundle of conflict management tools. Entrepreneurial applications that facilitate robust forms of information exchange and case management and that may be flexibly connected will provide more effective dispute avoidance through informed decision making. To more effectively create ODR tools, designers must focus on dispute profiles rather than mimicking ADR models. Dispute profile variables such as context, cause, missing information, outcome ranges, and patterns of behavior provide far more indication of potential resolution or avoidance mechanisms than existing ADR models. Online experiences are fractured; there is a seemingly endless stream of discrete interactions. ODR's future is in diversification and niche filling, finding where market inefficiencies are intolerable and capitalizing on novel means of amelioration.

GLOSSARY

Alternative dispute resolution (ADR) A set of practices to manage conflict outside of litigation and courts. Mediation, arbitration, and negotiation are the primary manifestations often facilitated by a neutral third party employing a variety of interpersonal and managerial skills.

Arbitration A dispute resolution method in which a neutral third party makes a definitive finding that may be binding or nonbinding.

Blind bidding An online dispute resolution method in which disputants make a series of blind bids and offers in a claims settlement negotiation. Settlement occurs when the difference between the bid and offer reach a certain predetermined threshold.

Chargebacks A credit card company practice of crediting customer accounts when transactions are in dispute.

Direct negotiation A dispute resolution method with no neutral third party. Disputants communicate and negotiate with each other directly.

Dispute resolution All methods and models for managing conflict. This includes alternative dispute resolution and traditional adversarial contests in courts.

Dispute avoidance Informal and formal practices to make disputing less likely.

Forum shopping: The practice of disputants selecting apparently favorable litigation or other dispute resolution contexts or facilitators.

Fourth party Software-created dispute resolution environments and constituent practices.

Hybrid online dispute resolution Dispute resolution processes that use both online technologies and offline practices to manage conflict.

Multivariable negotiation and decision support Online dispute resolution processes that handle complex variable-laden negotiations and decision making.

Mediation A dispute resolution method in which a neutral third party facilitates communication and case settlement between disputants. The neutral party does not make final decisions, and solutions are the product of the disputants.

Internet Corporation for Assigned Names and Numbers (ICANN) An Internet domain name registration regulatory and oversight entity.

MOOs/MUDs Early self-contained virtual social environments in which users conducted interaction via textual interface; users took on characters and immersed themselves in various social contexts.

Negotiation A dispute resolution or avoidance technique in which parties bargain over exchanges of value with the goal of reaching a final settlement.

Online dispute resolution The use of online technologies to support or sustain conflict management for disputes arising online and off.

Positive law State or other sovereign legal institutions, rules, and actors.

Third party An alternative dispute resolution facilitator (e.g., mediator or arbitrator).

Trustmark/Trustseal An online seal or signification indicating that an organization, corporation, or individual meets proffered standards of commercial and other online practices.

Uniform Dispute Resolution Policy (UDRP) Procedures for managing domain name registration conflicts under ICANN.

CROSS REFERENCES

See *Business-to-Business (B2B) Electronic Commerce; Consumer-Oriented Electronic Commerce; Online Communities.*

REFERENCES

American Bar Association Task Force on Electronic Commerce and ADR (2002, March). Addressing disputes in electronic commerce: Recommendations and report. Chicago, IL: American Bar Association.

Auerbach, J. S. (1983). *Justice without law?* Oxford: Oxford University Press.

Bordone, R. C. (1998). Electronic online dispute resolution: A systems approach—potential. *Harvard Negotiation Law Review, 147.*

Fuller, L. (1978). The forms and limits of adjudication. *Harvard Law Review, 92.*

Gaitenby, A. (1996). Law's mapping of cyberspace: The shape of new social space. *Technological Forecasting and Social Change, 52,* 135.

Geist, M. (2002, March). Fundamentally fair.com? An update on bias allegations and the ICANN UDRP. Retrieved from http://aixl.uottawa.ca/~geist/frameset.html

Hill, R. (1999, April). Online arbitration: Issues and solutions. *Arbitration International.* Retrieved April 4, 2002, from http://www.umass.edu/dispute/hill.htm

Hirschman, A. O. (1970). *Exit, voice, and loyalty; responses to decline in firms, organizations, and states,* Cambridge, MA: Harvard University Press.

Kassedjian, C., & Cahn, S. (1998). Dispute resolution online. *International Lawyer, 32,* 977

Katsh, E. (2002, Spring). Online dispute resolution: The next phase. *Lex Electronica, 7.* Retrieved April 4, 2002, from http://www.lexelectronica.org/articles/v7–2/katsh.htm

Katsh, E, & Rifkin, J. (2001). *Online dispute resolution.* New York: Jossey-Bass.

Katsh, E., Rifkin, J., & Gaitenby, A. (2000). E-commerce, e-disputes, and e-dispute resolution: In the shadow of "eBay Law." *Ohio State Journal of Dispute Resolution, vol 15.*

Lessig, L. (1999). *Code and other laws of cyberspace.* New York: Basic Books.

Mueller, M. (2000, November). Rough justice: An analysis of ICANN's dispute resolution policy. Retrieved April 4, 2002, from http://dcc.syr.edu/roughjustice.htm

Parisi v. Netlearning, Inc., 139 F. Supp. 2d 745 (2001).

Post, D. (1996, May). *Engineering a virtual magistrate system.* Paper prepared for NCAIR Conference on Online Dispute Resolution, Washington, DC.

Rule, C., (2002) *ODR for business: B2B ecommerce, consumer, employment, insurance, and other commercial conflicts.* New York: Jossey-Bass.

Sallen v. Corinthians Licenciamentos, 273 F.3d 14 (2001).

Thiessen, E. (2000). Beyond win-win in cyberspace. *Ohio State Journal of Dispute Resolution, 15.*

U.S. Federal Arbitration Act, 9 USCS 1 (2002).

FURTHER READING

Perritt, H. H. (2000). Dispute resolution in cyberspace: Demand for new forms of ADR. *Ohio State Journal of Dispute Resolution, 15.*

Perritt, H. H. (1996, May). *Electronic dispute resolution: An NCAIR Conference.* Paper prepared for NCAIR Conference on Online Dispute Resolution, Washington, DC.

Online News Services (Online Journalism)

Bruce Garrison, *University of Miami*

INTRODUCTION

Online journalism has become an important part of information available on the World Wide Web and the Internet. In the past decade, online magazines, wire services, newspapers, television stations, broadcast-cable networks, and subscription newsletters have matured to take advantage of their unique orientation and audiences. New technologies that allow consumers new ways to obtain Web-based information are introduced almost each year. Wireless technology that allows even greater access to online news and information is growing rapidly in the first decade of the new century. Airports, college campuses, business centers, downtown areas, and even entire communities are becoming wireless centers for laptop and notebook computer users, and for personal digital assistant-handheld (PDA) device users. Cell telephone users have even greater wireless access for online news and information. New technologies in development, such as flat-screen portable tablets, are likely to become popular in the near future as well.

Even with new wireless technologies for a variety of receiving devices, wired online services continue to grow in their availability and in use worldwide. The use numbers for online news sites and services, such as their e-mail alerts, have become staggering. News consumers are embracing these new communication tools. The new communication technologies have been viewed by news organizations as opportunities to extend their reach and contact with news consumers at international levels. The result is new forms and formats of news and information, such as news headlines and breaking story alerts delivery to subscribers, that are at the cutting edge of communication at the beginning of the new century.

One example is the recently reorganized and redesigned *Sacramento Bee* Web site (http://sacbee.com). The site, housed in a building across the street from the newspaper offices in the California state capitol, emphasizes local and regional content built on the local and state coverage of the newspaper (Stone, 2002). The local print content has been enhanced with multimedia, entertainment content, and related sites devoted to local entertainment (http://sacramento.com and http://www.sacticket.com/movies). All of this is done with about two dozen staff members, but only two devoted to content production and two others who focus on content management. Other staff members are involved in advertising sales, technical support, site management, and other site business activities (Stone, 2002). Online newspapers such as *The Bee* also understand the Web is a means to compete more directly with the instantaneous nature of the news services, radio, and television (Stone, 2002; Weise, 1997).

This chapter focuses on the nature and status of online news and journalism, including that practiced by newspapers, magazines, and broadcast and network television. It will provide an overview of the evolution of online news over the past decade. This chapter outlines the most significant economic models that have failed, have succeeded, or currently are being tested. Competition between online newspapers, radio stations, television stations and networks, magazines, and other media is also discussed. The chapter examines the changing nature of online news and the impact of multimedia and convergence on their content and staffing. The chapter provides brief descriptions of the major American online news organizations. The primary service models for online news and a summary of online news market sizes and models conclude the discussion.

ORIGINS OF ONLINE NEWS

The news industry is diversifying. Today's news companies own local and national newspapers, local television stations, cable television and radio networks, AM and FM radio stations, general and specialized market magazines, subscription newsletters, professional sports teams, outdoor advertising companies, newsprint plants, Internet service providers, and a growing number of high-profile World Wide Web sites. These popular Web sites cover a lot of informational ground. They include news and information content, but also focus on activities such as entertainment and nightlife, tourism and travel, sports, and reference/research. News companies, particularly those that began with newspapers several generations ago, have emerged as new enterprises in the multimedia world through expansion and contraction, redefinition, and restructuring. Most news companies have embraced this form of convergence and have moved fully into the digital age.

The number of online news sites grew geometrically in the mid-1990s, paralleling the growth in popularity of the Web beginning in 1993–95. However, the origins of online news can be traced to the early 1980s. In 1983, the Knight-Ridder newspaper group and AT&T launched a revolutionary experiment to bring people news on demand through their computers or television sets (AT&T dropped out of the project in 1984). Development work on the videotext service, called Viewtron, began in 1978 but took more than five years to debut. It remains noteworthy because it was a forerunner of today's online news media. Knight-Ridder suspended the Viewtron operations on April 1, 1986, having fewer than 20,000 subscribers and having lost $50 million. A similar venture at the *Los Angeles Times* called Gateway also closed operations in 1986, having only 3,000 users. The failed ventures left journalists wondering whether there would ever be a market for news on demand via computers (Burgess, 1993; Keefe, 1996). Viewtron and its partners even adopted rapidly developing PCs to their service by aggressively marketing software for PCs in 1985–86, but it was not enough. With the right timing, Viewtron might have become an America Online (AOL) or CompuServe, distinguishing itself with a strong component of news and information. Had the service existed only five years later, it would have been at the edge of the then-new Web.

A handful of daily newspapers were available online through proprietary online services in the 1980s. These included eleven newspapers that did an experiment with CompuServe in 1982. A few other newspapers produced online sites in 1990, including *The Albuquerque Tribune,* the *Atlanta Journal and Constitution,* and the *Rocky Mountain News* in Denver. University of Florida Interactive Media Lab Director David Carlson has outlined this development with an online media timeline on the World Wide Web (http://iml.jou.ufl.edu/carlson/professional/newmedia/default.htm).

The *Chicago Tribune* (http://www.chicagotribune.com) was the first newspaper to provide same-day editorial content to America Online in May 1992. Soon afterward, the *San Jose Mercury News* (http://www.bayarea.com/mld/mercurynews), also a pioneer in online news, was the first newspaper to put its entire contents online on AOL in 1993. At the time, when only three million Americans were online, the *Mercury News'* decision to put the newspaper's Mercury Center online was simply an experiment. As a result, most observers regard it as the first online newspaper. It was not surprising that a San Jose newspaper would lead the way in online journalism. Given its location, it considers itself the paper of record for the famed Silicon Valley.

By the mid- 1990s, as the Web expanded rapidly, broadcast news outlets were developing a significant presence on the Web as well. Broadcast and cable networks were among the earliest to put their news content on the Web, but local affiliates began to build Web sites with news content that also served a promotional purpose by offering program listings, biographies and other information about air personalities, and various forms of community service features.

Competitive amateur and professional sports have become a major drawing card for online news sites in the past decade as well. Local and national news organizations have built a user base by providing information about local teams and major sports events while competition is underway. Local newspapers and television stations often offer running summaries and updated scores of events not always on television or radio. They enhance printed and broadcast coverage by offering statistics, schedules, and other information that, because of space or time limitations, cannot be published or broadcast. National news media often provide similar service for professional teams and major college events.

There was also a growth of Web-based electronic magazines at about the same time. These topical online publications are often called "zines" or "e-zines." Most zines did not originate in printed form, but certainly were built upon the printed magazine tradition. They are often highly specialized. Zines have been described as "independent publications characterized by idiosyncratic themes, low circulation, irregular frequency, ephemeral duration, and noncommercial orientation" (Rauch, 2002). Almost overnight, these energetic periodicals appeared all over the Web and received a large amount of popular media coverage and attention as the trend of the moment. The Web created opportunities for this new genre of specialized online news and information in the mid- and late 1990s, and they have continued into the new century with a steady niche audience appeal. There are thousands of zines on the Web today and they attract an audience in the millions (Rauch, 2002).

THE ONLINE NEWS WORLD

For the past half decade, news organizations have scrambled to keep up with competitors to establish their strong presence on the Internet and Web. In some studies, Internet use for reading and viewing news has exceeded traditional news media such as magazines and radio. One study found it approached the use levels of network television and cable television ("Internet Growing as News Medium," 2002). After having set themselves up for the future, online news media have worked to define themselves. This has been difficult for news organizations, whether they are based in newspapers, magazines, television, or radio, or are independents and only on the Web. For online newspapers, the effort to determine the role and model in which to work has already gone through several iterations.

There will be more to come. Just like radio and television before it, the newest technologies undergo a formation period. Online newspapers are experiencing an era of exploration and self-determination. Although there are no geographic boundaries for online newspapers, practically speaking, most people who access a local newspaper do it for local news. Online newspapers differ from most online news media in their intensely local nature provided by the content they offer from their print editions.

Formation of new media can trigger transformation of established media. New media may also have a destabilizing effect on old media, but old media also tend to embrace the new media (Samoriski, 2002). For example, the development of television brought on change in movie newsreels and, ultimately, their disappearance in theaters. However, it is also common for new media to complement and benefit existing outlets (Samoriski, 2002). For example, print newspapers might advise readers to go online for more information not offered in the printed editions due to space limitations or they can offer links to multimedia content-such as interviews with key newsmakers, photo galleries, video and audio of major events as they occurred-available on the online newspaper's Web site.

Online newspapers require professional abilities and skills different from their print counterparts. While both forms require news judgment, speed, and other fundamental journalistic tools, online newspapers have sought producers with greater versatility to work in multimedia as well as computer network technologies (Stepp, 1996). Old skills-based walls have been taken down in favor of individuals who can do it all—write, edit, compile, tell the story visually, and work with audio and video content. The changing staff needs have forced some online newspapers to recruit beyond their own newspaper newsrooms. News companies will continue to look for specialists in different skills areas such as print and television, but will also increasingly seek "multimedia super-reporters" who cover the world with backpacks loaded with laptops, digital video and still cameras, and audio recorders (Gates, 2002).

New news media have also forced change among the more traditional online news media. The popularity of coverage offered by independent news Web sites such as Salon.com (http://www.salon.com/) and Matt Drudge and the Drudge Report (http://www.drudgereport.com/) have brought mainstream news media on the Web to new horizons. Drudge, on his own in early 1998, broke coverage of the Monica Lewinsky and Bill Clinton affair that scandalized the White House and the Clinton administration for several years. Established news organizations, such as those that are part of newspaper or television organizations, have responded to the wide-open approaches offered by these sites with more liberal definitions of news that include less traditional approaches to breaking stories, features, and other content.

CONVERGENCE

The best online newspapers have readily moved from shovelware—the period when what they offered was little more than the published content of the traditional newspaper and the wire services—toward convergence. They seem to have matured past the point of scooping themselves when coverage of either the online or print edition carries a story that the other does not ("An Evolving Medium," 1999; Fee, 2002). Not only do the best online newspapers today offer much more than just the best content of the daily newspaper, they offer an interactive Internet home for their established print edition readers as well as nonreader residents and visitors. Many online newspapers have developed online information *portals* for their communities (Miller, 2001). They want to offer readers a wide range of interactive tools and technologies to experience news and information like it has never been provided before ("The Impact of the Internet," 2000).

One of the most-publicized examples of national and international corporate media convergence was the America Online–Time Warner merger that was announced in early 2000. Some reports described it as a merger, but AOL stockholders actually owned 55% of the new company when the deal was completed. AOL obtained the immense resources of Time Warner, including Time Inc., and CNN, through the buyout involving $156 billion in AOL stock. This event set the stage for dozens of multimedia and convergence opportunities for the new corporation that now includes America Online's worldwide network access, its Internet services, and its proprietary Web content, Time Warner's CNN broadcast news networks, its widely circulated and respected magazines (e.g., *Time, Fortune, Sports Illustrated, Money, People*) and news services, its cable television and broadband customers, its music library, its film library and archives, and much more. It has been characterized by one media observer as a "new age, techno-cultural behemoth" (George, 2000). The possibilities of such resources combined with the Internet and online media are dizzying.

A leading example of convergence at the regional level is the merger of newsrooms by Media General in Tampa, Florida. *The Tampa Tribune* (http://tampatrib.com), the morning daily newspaper serving Tampa Bay, WFLA-TV (http://www.wfla.com) Channel 8, the NBC broadcast television affiliate, and Tampa Bay Online (http://www.tbo.com), the Web site serving both the newspaper and television station, moved into a single facility in 2000 to share and work together in an early effort at convergence of newsrooms and technologies. The three news organizations share news resources, equipment, and staff. In fact, the large newsrooms, studios, workstations, and other facilities spread over several floors of the news center are anchored by a circular central news desk in the building's atrium. Journalists from each of the three news organizations staff the news desk and coordinate use of images, audio, video, wire content, local photographs, and news stories as well as other available content.

Web portals provide much more than just news content. Many traditional news media have established portals for residents and visitors of their communities and regions. *The San Jose Mercury News* (http://www.bayarea.com/mld/mercurynews/), for example, is the daily newspaper portal serving California's Silicon Valley and is the host newspaper for Knight-Ridder Newspapers (http://www.knightridder.com). It serves the Bay Area with the equally suburban *Contra Costa Times* within a site called BayArea.com (http://www.bayarea.com). *The Boston Globe* has established a similar portal for New England with Boston.com (http://www.boston.com). Online newspaper and television portals provide a gateway to a series of sites and links about a region or metropolitan area. They are also often customizable so readers can select types of information displayed on the home page.

ONLINE NEWS WRITING

Online journalism has the opportunity to take the best features and characteristics of existing mass media and news media, eliminate most of the weaknesses, and roll them into a new medium in which stories can be told in new ways using words, video, audio, and still photos together.

To accomplish this, online news writing is evolving. It is, of course, quite different from newspaper, broadcast, or other news media writing. To be truly valuable to users, online news must take advantage of hypermedia and hypertext and their nonlinear nature. Writing for online news sites is often done in layers (e.g., headlines, summaries, leads, and other text), often called "chunking," and with an awareness of clicking and scrolling capabilities of browser software. Online storytelling should involve the wide variety of media forms available on the Web, not just text and images. This includes audio, video, and forms of reader-user interactivity such as chat rooms and instant opinion polls.

Another important part of writing for online newspapers is style. Each publication develops its own style. Sometimes, it is personal, casual, and informal. Other approaches can be more formal. Often the style used is a function of the audience served (Rich, 1999; Whittaker, 2000). Although the Internet allows storing long articles easily, the compact style has not been abandoned. News writers still strive for brevity for many of their stories in order to display an entire story in one screen view without scrolling.

Online newspapers have a struggle similar to that of their print counterparts. Both are fighting for survival in a digital- and video-oriented world. With countless other online news sources coming from television, radio, and magazines, competition for online newspapers is intense.

Add to the mix a new set of news providers—online services such as America Online and Yahoo!—and the job becomes much more difficult. With an online customer base of 29 million customers and nine million log-ins per day compared to three million for *The New York Times'* online version, AOL cannot be ignored as a news provider. Yet analysts are fast to criticize the news judgment and coverage of AOL as overly commercialized, perhaps even unethical (Koerner, 2001). If, as some AOL executives maintain, AOL is used by many of its customers for news and information instead of local online newspapers, the electronic editions of newspapers have a long road ahead for success and, perhaps, survival. As one critic wrote, online news consumers and journalists should "be afraid" of such ventures (Koerner, 2001, p. 25).

The evolution of online newspapers has paralleled the development of other online news media in recent years ("An Evolving Medium," 1999; Strupp, 1999). They are learning what works and what does not. Newspapers have offered electronic mail alert services for breaking and ongoing stories, severe weather bulletins, or even more routine notifications of new editions of their publications (Langfield, 2001; Outing, 2001). While interactive features, such as electronic mail used for news alert services and bulletins, have been successful, Web-based chat rooms have not been used as effectively (Perlman, 1999; Pryor, 2000). Major national tragedies, such as those surrounding the student shootings at Columbine High School in Colorado on April 20, 1999, or the terrorist attacks in New York, Washington, and Pennsylvania on September 11, 2001, provided opportunities to use interactivity, but many online newspapers were criticized for failing to use them well (Outing, 2001; Perlman, 1999).

Newspaper sites have become quite popular during this learning period, but newspapers continue to learn how to use their Web edition to attract and serve readers and advertisers (Strupp, 1999; Pryor, 2000). For example, use of interactive features to enhance reader feedback seems to be growing in newspaper newsrooms ("The Impact of the Internet," 2000). While there is growth in sophistication in developing news content, the same cannot be said about commercial endeavors that would support the news product. Electronic commerce and advertising opportunities remain a mystery to many in the industry. Online newspaper managers certainly have not learned everything necessary about turning the sites into profit centers for the company, but neither have other online news organizations at this point in their development. The online newspaper industry is so young that it remains uncertain how its costs will be paid (McMillan, 1998; Palser, 2001b; Pryor, 2000).

Online news technology has altered traditional news gathering and distribution in at least one additional way. The low cost of Internet access and storage space combined with the user-friendly nature of its software for content creation has led to a form of personal journalism that has become known as the "weblog" or "blog." These personal and noncommercial sites are usually prepared by date of entry in a log, journal, or diary format. A site is usually specialized and focuses on a single topic or subject area and offers information and observations from the author, contributors, or from other sites. Some weblogs are single author and some are collaborations of groups of interested contributors. The sites are often also called blogs. These sites have attracted quite a following in the past decade and led to the phenomenon of weblogging or blogging. Readers find them appealing since they change every day. This highly personal approach to journalism has been used in independent, nonbranded online news sites and has become a feature of some more conventional online news sites as well. The ease of using the software required to create blogs has led to their worldwide growth in popularity (Higgins, 2000; Walker, 2001).

SEARCHING FOR AN ECONOMIC MODEL

New York Times Company Chairman and Publisher Arthur Sulzberger, Jr., effectively expressed the position of newspapers in the online age. He recently stated that "If we're going to define ourselves by our history, then we deserve to go out of business" (Gates, 2002). "Newspapers cannot be defined by the second word—paper. They've got to be defined by the first—news. All of us have to become agnostic as to the method of distribution. We've got to be as powerful online, as powerful in TV and broadcasting, as we are powerful in newsprint."

Newspapers, more than any other traditional news medium, raced to go online during the mid-1990s. In 1994 and 1995, they were offered for free. However, some owners and managers felt they could charge for the content and offered online subscriptions. This was short-lived, however, as the experiment failed when users found comparable information elsewhere on the Web (Grimes, 2002). Owners and managers had expected to be part of a digital gold rush (Dibean & Garrison, 2001; Quint, 1994), but the anticipated gold rush never came because the medium has not found a profitable economic model. To date, there simply has not been sufficient advertising volume and other revenue to support online news. News on the Web, some have observed, spreads on the Web very quickly, has a short shelf-life, and can be found for free if a user searches hard enough (Grimes, 2002).

There are numerous business models for Web content, including the failed videotex model, the failed paid Internet model, the failed free Web model, the failed Internet/Web advertising "push" model, the present portals and personal portals models, and the evolving digital portals model (Picard, 2000). Online news organizations are experimenting today with portal approaches. Within the various portal models, the subscription model may be the future despite previous failed efforts. Financial conditions often lead to such radical change and the economic environment in late 2001 and 2002 following the September 11 terrorist attacks against the United States, combined with a recession, put pressure on sites to generate income. Online advertising was not producing sufficient content and sites were forced to find a different economic model. Suddenly, the subscription model looked better and, as one publication in 2002 stated, "the free lunch is over" (Smith, 2002, February; see also Smith, 2002, April; Trombly, 2002). The collapse of free alternatives will lead consumers to accept the subscription or other fee-based models for operation of online news sites. Commercial sites with valued content have already found this to work. *The Wall Street Journal* and its online site, for example, had more than 609,000 subscribers in 2002. "We already know that consumers will pay for highly valued, proprietary content that either makes or saves them money," stated *EContent* writer Steve Smith (2002, February, p. 19). "For most other content brands, however, it gets more complicated. The fear of competition from free alternatives, the risks of severe traffic loss, and the overwhelming resistance among users has left the fee-based approach more of a pipe dream."

Despite this surge toward fee-based content, it is apparent that consumers have been conditioned to expect to find news online at no cost (Donatello, 2002). Online users see the Web as an open frontier that is free to all, and they adamantly oppose paying even modest fees (Noack, 1998). Since 1997, some online newspapers ceased publication for lack of profitability.

The most noteworthy newspaper to charge for its online content is *The Wall Street Journal's* Interactive Edition (http://www.wsj.com). Print subscribers to the *Journal* and *Barron's* pay a smaller fee for most of the online edition. The *Journal* differs from the other fee-based newspapers in that it is a nationally prestigious newspaper. The *Journal* also offers much editorial content related to business and finance, making it more akin to a specialized online publication to which an interested niche audience is willing to pay a fee (Rieder, 2001).

Online news consumers are upscale and typically live in urban or metropolitan areas. A study of online newspaper consumers in 2001 concluded there is an undeveloped newspaper market for paid access to local information and news. The study concluded that "substantial marketing efforts" were required to develop the consumer base. Consumers, the study stated, are (a) unwilling to pay for content because they are conditioned to finding it for free, (b) they do not see "incremental value" in online news, (c) newspapers are "strong enough to lure a core audience online, but not strong enough to make them pay," (d) the process of paying for content reinforces attitudes against paying for it, and (e) consumers are willing to register and provide personal information, but not to pay cash for access (Donatello, 2002, p. 36).

Online news media, unlike their forerunners in videotex services, are popular. At the very least, newspapers, radio and television stations and networks, and news magazines sustain them to maintain their brand reputations in cyberspace. However, until the industry can find a successful economic model for online news media to make a profit, or at least sustain its operations, they will always be a promotional service in their organizations rather than distinct editorial voices. For several years, news organizations have been searching to find the solution for a profitable Web site. According to one national analysis, about half of the 171 newspapers participating in a 2001 survey reported slight profits, mostly from classified advertising, but they also earned some revenue from banner advertising, sponsorships, and Web-related services fees (Noack, 1998; Scasny, 2001; "The Impact of the Internet," 2000; Trombly, 2002).

The financial models for online newspapers and other online news organizations have not had enough time to test what works and what does not. In fact, one industry expert described the development of online business models as "short and wild" (Palser, 2001b, p. 82). Still, at least a few of these experts feel it is not a bad idea if marketed properly (Palser, 2001a, 2001b) or, if news sites offer more value, paid access will succeed (Scasny, 2001). It does not appear likely, in the minds of observers, that subscription revenues will overtake and replace advertising revenues in the near future. There are too many variables, such as sector served, markets, brands involved, and prices. Online news sites must explore content and merchandise offered. Sites may have to target highly specific audiences and meet their information needs, for instance, to succeed. Ultimately, the situation seems to involve redesigning content offered to create greater demand instead of simply getting customers used to paying for something they used to get for free (Smith, 2002, February).

THE COMPETITIVE DIMENSION

One of the main characteristics of online news media that make them different from more traditional news media is hypertext. Fredin (1997; see also Fredin and David, 1998) noted that in order for online newspapers and other media to provide a news service online, "a computer-based interactive system requires a wide range of links or choices within stories" (p. 1). Most news organizations are only beginning to understand the uses of hypertext, linking, and integration of other media into their coverage of breaking news, features, and other information that can provide the depth and scope of coverage expected of newspapers and their online counterparts (Marlatt, 1999).

Multimedia news content distributed on the Web still has a lot of developing to do. Online audio and video quality is often a function of connection speed. Individuals connecting to an online news source with a modem or wireless device will often experience slide-show like video and out-of-sync audio or worse (Palser, 2000). Individuals with a digital subscriber line (DSL), cable modem, or faster connections still find difficulties with loading sound and video clips and getting them to play. One critic wrote that online news video was in its "adolescence" in mid-2000 (Palser, 2000, p. 84). Its potential in enhancing online newspapers and their coverage, however, seems great no matter how it is seen today, especially for newspapers that have sister television and radio stations for shared content. When Web-based multimedia news works, it can be stunning in the same way that live television takes us to the scene of breaking stories.

Some newspapers, particularly those dailies in small and medium-sized markets, have begun to use live and archived audio on the Web to enhance coverage and add to their competitive edge. One of the best at using multimedia is the *Topeka Capital-Journal* (CJ Online, http://www.cjonline.com) in Kansas. It offers exceptional access to video and audio clips for users. Because of this and other high quality site content, the *Capital-Journal* was honored by *Editor & Publisher,* a newspaper industry trade magazine, as the Best Online Newspaper in 2002 (Richardson, 2002). Among the other early multimedia users have been the *Indianapolis Star-News* (IndyStar.com, http://www.starnews.com), and the *St. Paul Pioneer Press* (PioneerPlanet, http://www.pioneerplanet.com). These sites use sound from press conferences, interviews, and even voice reports from their own staff members in the field (Lasica, 1999). One media analyst, in fact, suggested that online newspapers use a database management model—in which stories, images, and other content are stored in a database and Web news pages are built on demand—for distributing their news instead of the broadcast, or television station style, model that had been common throughout the past decade (Blankenhorn, 1999).

Online news organizations are often restricted by their budgets and staff resources. While they may seek to use interactive and multimedia resources and to develop original content, the number of people, their experience and skills levels, and the money available to pay the bills often sets stark parameters in which the online edition operates (Fouhy, 2000; South, 1999). In recent years, budgets of online newspapers and their staff sizes have been impacted by worsening economic conditions—increasing operational expenditures and decreasing revenues—that have led to numerous layoffs and curtailed services (Fouhy, 2000; Houston, 1999; Palser, 2001a; Rieder, 2001). Survival became the focus of many online news sites as the new century began (Farhi, 2000b). Some analysts have argued that such decisions were short-sighted and not in the future interests of online news (Palser, 2001a; Rieder, 2001). Cutbacks have led to greater dependence on automated Web page production using software that builds HTML pages from news system databases, including those of the wire services.

Online managers use research about their audiences on a regular basis. It may not be conventional readership research in the same way newspapers or broadcast stations have conducted it over the years. Instead, Web-based news sites depend on traffic reports generated by software that monitors the Web site's server and visits to and from the site and its pages. Online news sites are also using information about Web site users, such as how long they spend reading a site's pages, to help sell advertising and to target audiences more carefully (Nicholas & Huntington, 2000).

LEADING ONLINE NEWS SITES

It is important to note, at this point, that some of the most popular online news sites are operated by broadcasting companies. This underlines the popularity of television as an entertainment and news source during the past four decades in the United States. Since the 1960s, the largest proportion of Americans has used television more than newspapers, magazines, or radio for daily news and information and that brand success has carried over to the broadcast companies' presence on the World Wide Web.

Just as they have done in the print and broadcast world, a handful of online news sites have dominated the short lifespan of online journalism. Recently, this has meant a combination of sites based on the nation's leading newspapers and television networks, as shown in Tables 1 and 2. These top sites have been typically identified by their visitor traffic per day or other time period, but also because of their traditional content quality.

Table 1 Top 10 U.S. Online News Sites during 2001

Average Monthly Visitors		
1. MSNBC	MSNBC.com	10.7 million
2. CNN	CNN.com	9.5
3. New York Times	NYTimes.com	4.8
4. ABC News	ABCnews.com	3.9
5. USA Today	USAToday.com	3.3
6. Washington Post	WashingtonPost.com	3.1
7. Time magazine	Time.com	2.0
8. Los Angeles Times	LATimes.com	1.9
9. FoxNews	FoxNews.com	1.4
10. Wall Street Journal	WSJ.com	1.4

Source: Barringer, 2001, (p. C6).

Table 2 Leading U.S. General Online News Providers in 2002

Web Site Reach at Home/Work		
1. MSNBC	MSNBC.com	18.0%
2. CNN	CNN.com	14.1
3. Yahoo! News	Yahoo.com	13.1
4. AOL News	AOL.com	11.9
5. ABC News	ABCnews.com	6.7
6. New York Times	NYTimes.com	6.0
7. USA Today	USAToday.com	5.5
8. Washington Post	WashingtonPost.com	4.0
9. FoxNews	FoxNews.com	1.8

Source: MSNBC rated #1 (2002).

ABC News

ABC News, the television and radio broadcast network, has emerged as one of the leading U.S. broadcast news sites on the Internet. Drawing on its rich base of radio and television content from the network and its affiliates, the site (http://abcnews.com) features updated news content and news from its own news programs ("Good Morning America," "World News Tonight," "20/20," "Downtown," "Primetime, "Nightline," "World News Now," and "This Week").

The site, which went online on May 15, 1997, utilizes a large amount of interactivity. Site visitors can view network audio and video as well as webcasts. Users can visit community message boards, chat with other visitors about current public issues and breaking news stories, subscribe to e-mail news alert services, and send news content feedback, make story or coverage suggestions, and provide other comments to the newsroom. Visitors to the site as well as network television viewers are frequently encouraged to interact.

The site combines the news and information content from ABC News and The Walt Disney Internet Group. The site is updated in an ongoing fashion throughout the day and night and seven days a week. It describes itself as "up-to-the-minute, engaging, informative, and interactive coverage of a range of issues and events" (ABC Newsroom, 2001). ABC News is also affiliated with ESPN, the cable sports network and channels, and often offers its content on the site.

ABC News television correspondent Sam Donaldson is frequently seen on the Web site, anchoring his Internet-only live webcast called "SamDonaldson@ABCNEWS.com." ABC News investigative reporter Chris Wallace produces and anchors an Internet-only edition of "Internet Expos." The Wallace program has covered issues and events involving the Mafia, UFOs, and the impact of pornography on the Internet.

The award-winning site attracts about 4.3 million unique users per month. This creates about 23.9 million visits and 103.9 million impressions (or deliveries of advertising messages to site readers). Most users are about 40 years old, male (59%), and are in their offices (54%) or at home (41%) when visiting the site (Advertising Opportunities, 2002).

Cable News Network

The Atlanta-based Cable News Network, known worldwide as CNN, has become a dominant international and national news site (http://www.cnn.com) on the Internet. The television news network, which debuted in 1980, arrived early on the World Wide Web and established itself as one of the leading multimedia news sites in the world, just as its broadcast parent has become in the past decade. Taking advantage of its existing video and audio content and nearly 4,000 news employees worldwide, the site combines content from the numerous television and radio services and networks with print sources to produce a news site that often functions as a news portal for many users.

The site offers three editions. The U.S. edition is geared to cnn.com's largest audience. However, the site also offers an International edition (http://edition.cnn.com) as well. Content on each edition's site is similar in that it focuses on breaking and daily news and events, weather, sports, business, politics, law, science/technology, space, health and medicine, travel, and education. While the Asian and European pages are in English, the European page also offers readers connections to local language sites produced in Arabic, German, Italian, Spanish, and Portuguese. The CNN sites offer a variety of interactive opportunities for visitors such as the CNN polls about current events and issues, viewer feedback, and e-mail newsletters and alerts about breaking news. Content is also available for mobile telephone services.

The site's strongest attractions are its international and national coverage of news, but it also provides limited local market news through its affiliated network stations. The site has made the latest multimedia technologies, from live video streaming to audio packages, available to users. It also provides users access to searchable archives of news features and background information on the site. Like other major news sites, CNN's site is updated continuously throughout the day. The CNN partnership with corporate cousin *Time* and its well-established stable of international publications has led to extensive content from such prominent magazines as *Time, Sports Illustrated, Fortune,* and *Money* on the CNN site.

In 2002, cnn.com began distributing much of its video content by subscription only. Described as premium content, it offered the service with more video content and higher quality than that of the freely distributed video clips and packages from its networks. The package included CNN, CNNSI, and CNN Money video content by the month or annually at the basic level and viewers may access unedited video and "quickcasts" at the advanced level of service. CNN's sites, of course, also emphasize user interactivity through the polls, multimedia content, e-mail, and other features.

Fox News Channel

Fox News Channel (http://www.foxnews.com) is a relative newcomer in television news in the United States when compared to the well-established ABC, CBS, NBC, and CNN networks, but it has come on strong in the past few years to become a popular news site for online news users. It was ranked in the top ten of U.S. news Web sites in terms of site traffic in 2002.

The site features content related to the Fox News network news programs, the top news stories of the day, business news, sports, politics, weather, lifestyles and living, and opinion. Much of the content is generated by Fox News reporters, correspondents, and others in the Fox News Channel newsrooms around the world. The site also uses a large amount of Associated Press content.

The main sections of the site include FNC Shows, which provide listings information about Fox News programs; Top Stories, which provides headlines and links to full stories about major news of the day; Politics, which focuses on international, national, and state political news from the world's capitals; Business, which offers headlines and stories on current business news and information from Standard and Poor's, financial and investment advice, and stock performance information and price quotes; and FoxLife, which concentrates on leisure activities such as movies and music, celebrities news, and other personal lifestyle news.

The Views section of the site offers opinion essays, commentaries, and columns from Fox News correspondents, news program hosts, and contributors from outside the newsroom. The Weather section provides local forecasts and conditions searches, AccuWeather dangerous weather alerts, and other almanac and historical information. The Fox News site also has a strong Sports news and information presence that comes from/through integration with the Fox Sports Web site (http://www.foxsports.com). The Sports site is organized generally by sports and offers up-to-date partial and complete scores, sports programming listings, and headlines and stories of the current news cycle.

Fox News provides an enhanced Web experience for its registered users. Registration is free, but it gives participants newsletters and discounts for fee-based services such as video archives access as well as other benefits.

Los Angeles Times

The *Los Angeles Times* is one of the world's leading newspapers and the dominant news source on the West Coast. Although it is internationally known, the newspaper seeks to serve Southern California with its online presence. Nonetheless, it has a significant audience of Californians as well as a national and international following, according to Richard Core (personal communication, May 2, 2002), editor of latimes.com. Core has responsibility for overseeing the editorial production of the site. Core said about half of the site's visitors come from the print edition's primary circulation area of Southern California, the remaining half is worldwide.

The site went on the Web in April 1995, but an earlier electronic edition known as Times Link was available in 1993 and 1994 on the proprietary Prodigy information system. The site has built a considerable amount of daily traffic. Core said there are about 1.7 million page views per day (the number of pages, not single advertisements, seen by users).

The site offers news, sports, business, and entertainment news stories that originate from the newspaper. This is supplemented with wire service content and video from KTLA-TV, an affiliated Los Angeles television station. However, the strength of the site is the newspaper's daily and weekend coverage. "We offer the whole package that you would get from a metropolitan newspaper," Core explained. The *Times'* entire print edition is not included on its Web site, however. This includes display advertising, tables such as Business section CD interest rates or Sports section results and statistics, and syndicated content such as comics or advice columns. Almost all of the *Times'* news stories are posted except in extremely rare cases. Some puzzles and some photographs are also not included, but the site producers plan to improve on this.

There is free content on the site that is not found in the printed editions. It includes Calendar Live! Web-only entertainment coverage, guides on Southern California, specially organized special reports, a searchable restaurant guide, searchable classifieds, and audio and video content. Original content also includes multimedia such as interviews with reporters on major local stories, documents to support reporting (such as Adobe Acrobat pdf files of indictments and lawsuit filings), and wire content. The site also presents affiliated television station video and audio clips, and it offers graphics such as Flash presentations.

MSNBC

MSNBC is the combined effort of Microsoft Corporation in Redmond, Washington, and NBC News in New York. The joint product of the computer software giant and one of the major news networks in the United States created both a widely watched cable television network, MSNBC, and what has become the nation's leading online news site (http://www.msnbc.com). In 2002, the site was independently rated as the leader in the general news category, according to Jupiter Media Metrix (MSNBC rated #1, 2002). Much of its recent popularity with online news users was due to its coverage of the 2002 Olympics, the war in Afghanistan, and coverage of terrorism threats during 2001–02.

MSNBC is a leader in creation of content solely for the Web. MSNBC content producers do not offer users as much repurposed content as newspaper sites do, for example. MSNBC.com utilizes the resources of NBC's network television and radio operation, its business-oriented channel CNBC, and NBC affiliate stations and news services. MSNBC on the Web provides content from the collection of NBC News news programs, MSNBC TV news programs, CNBC news programs, and NBC Sports broadcasts. CNBC offers extensive business content to the site, including live stock market coverage and frequent news updates that are included on the Web site. The site prepares coverage of major daily national and international stories, updates on breaking stories, and in-depth reports on issues. MSNBC is also partnered with *Newsweek* magazine and often shares information with readers on the site.

The national award-winning site is searchable and keeps stories on the site for about a week. The site also offers webcasts featuring hourly video news updates and extensive streaming video. The site has a reputation as a leader in breaking news and original journalism on the Internet. MSNBC.com also has developed a series of partnerships that generate site content as well. These include arrangements with *The Washington Post, Newsweek magazine, The Wall Street Journal,* ZDNet, MSN Money Central, *The Sporting News,* Expedia, E! Online, Pencil News, *FEED* magazine, *Science* magazine, BET.com, inc.com, Space.com, Inside.com, The Motley Fool, RedHerring.com, Internet Broadcasting Systems, and Pioneer Newspapers.

In addition, MSNBC.com programs interactive content for MSNBC TV and NBC News on WebTV Plus. The breadth and depth of the coverage is combined with high levels of personalization. MSNBC.com content is also available through wireless services.

The site permits a large degree of user customization. These features include the ability to design a personal front page to display predetermined categories of headlines and news, current stock quotes, and sports scores. Users can also find NBC local affiliate stations' local news content on a daily basis.

Interactivity options include chat sessions and voting opinions on controversial topics and issues on specially produced "Dateline Interactive" programs. Results are shown live on-air. Readers may also use WebTV Plus to access MSNBC enhanced television via MSNBC TV 24-hours a day as well as NBC News programs:"Dateline NBC," "Nightly News," and the "Today" show.

San Jose Mercury News

The San Jose Mercury News online edition is part of a larger portal site serving the San Francisco Bay region known as BayArea.com. The site went online in 1995, but its online roots of the early 1990s Mercury Center are even deeper. *The Mercury News* site is dependent on the printed edition of the newspaper. About 95% of the *Mercury News* online site's text content comes from the newspaper, noted Rob Neill (personal communication, May 2, 2002), program manager for BayArea.com. Other content originates from wire services and newspapers such as *The New York Times* and *Washington Post.* The site does not have individuals devoted solely to reporting for the online edition. "I don't think it makes sense having a specific team writing for online. It throws up a wall," Neill explained. "The newsroom is best suited for producing news and online producers are best suited to put news online."

The recently redesigned site, which debuted in early 2002, does not require registration or subscription. It is searchable and information is easy to find, Neill (2002) says. To improve service to readers, the online production staff constantly refines site usability, Neill added. "You can always work on usability. I don't think that's something that's ever perfect."

BayArea.com and the *Mercury News* managers and staff are searching for the right news model, Neill said. "Newsrooms have to redefine what they are, what a news story is," he said. "I am a big fan of newspaper reporters. I used to be one. I don't think we need to be sending other people to do their jobs."

The New York Times

The New York Times is one of the world's preeminent newspapers. Its online edition, called New York Times on The Web (http://www.nytimes.com), debuted on January 20, 1996. The site truly serves the world. Its readers come from all over the globe. Editor Bernard Gwertzman, who is in charge of the entire editorial operation online, determines the home page content each day (personal communication, April 10, 2002). The site draws about one-quarter of its audience from outside the United States, another quarter of the audience from the New York metropolitan area, and the rest are from other parts of the United States.

The site draws high audience numbers, partly because of its print edition content but also because of original content. There are about 1.5 million users per day visiting the site and about 300 million page views per month. All of which, of course, leads Gwertzman to be concerned about the servers' capacities at the newspaper. The content of the site varies, but reflects the general news standards of the company, Gwertzman said. All of the newspaper's National Edition is offered on the online site, although the Web site pays, Gwertzman said, "a little more attention to entertainment." During the first two months of the war on terrorism in late 2001, for example, *The Times'* site carried a photographer's journal from Afghanistan narrated by photographer Vincent Laforet. The site relies on newspaper reporters and wire services for most of the online content, but content at any given moment is a function of the time of the day.

The Times requires online users to register at no cost and it uses the demographic information it obtains to understand its market and attract advertisers. Users can subscribe to e-mail-based news alerts for major or breaking stories. These alerts were particularly valuable to readers in the aftermath of the attacks on the World Trade Center.

Major stories are a way of life at the site. The site's producers displayed the stories that won the newspaper a record seven Pulitzer Prizes in April 2002 for its coverage of the September 11, 2001, attacks, even though it is difficult to determine whether online coverage had any impact on the Pulitzers (Kramer, 2002). The newspaper's Web site has also received considerable attention for coverage of the Enron scandals, the 2000 presidential election, the Monica Lewinski scandal and investigation, and the government's seizure of Elián González in Miami to return him to his father in Cuba.

The Washington Post

The Washington Post (http://washingtonpost.com) has an international and national audience. Built upon the reputation of the print brand of the national capitol's primary daily newspaper, the site has rapidly built a strong following since its debut in June of 1996. Douglas Feaver, executive editor of washingtonpost.com, oversees the content of the site. He said (personal communication, April 10, 2002) the site does well providing online news and information to the Washington metropolitan area, including suburban Virginia and Maryland, as well as a large international market.

The Washington Post online, like *The New York Times,* has opted to present itself as a national and international online newspaper. The results have been impressive with 4 to 4.5 million unique visits a month and about seven million page views on a typical weekday and about half that amount on weekends.

Approximately 20 to 25% of the site's content is original. Feaver said (personal communication, April 10, 2002) about 60 to 75% of news content is original and about 33% of features are original to the site. Content is developed mostly by the newspaper's extensive staff, but also from online staff writers and Associated Press, Reuters, and Agence France-Presse wire services. Although about 240 persons are employed at washingtonpost.com, only five writers are devoted solely to developing online content. A number of other part-timers contribute news, features, business, health, and entertainment content. "We have the entire *Washington Post* and additional information that compose our content," Feaver stated. "This includes wire services, exclusive Internet stories, restaurant reviews, and audio."

The Post's online edition has attracted considerable international attention because of a number of Washington-based stories in recent years. Coverage of the September 11, 2001, attack on the Pentagon is among the most recent. Despite the incredible demands for online news that day that crashed servers around the United States, *The Post* site remained online 95% of the time. Not only did the site develop its own content as the story unfolded throughout the day, but it also published the newspaper's multiple Pulitzer Prize-winning stories throughout the fall. The site also drew attention with its coverage of the Monica Lewinsky and Bill Clinton scandal and investigations in 1998 and it was among the first online sites to provide coverage of the seizure of Elián González in Miami-thanks to coverage from the newspaper's Miami bureau chief.

USA Today

Gannett Corporation is one of the world's largest news businesses. It is the largest newspaper group in the United States with 95 newspapers and a combined daily circulation of about 7.7 million. It also owns and operates 22 television stations. Its flagship newspaper is *USA Today,* which has a daily circulation of 2.3 million and is available in 60 countries worldwide. The company also publishes *USA Weekend,* a weekly magazine of 24 million circulation in almost 600 newspapers (Gannett, 2002).

The newspaper is internationally and nationally circulated five days per week and is one of the world's largest. Gannett is known for its corporate sharing and this benefits all of its newspapers, broadcast stations, and online sites. The online edition of *USA Today* (http://www.usatoday.com) has won numerous national awards for its online journalism (Accolades, 2002).

The Web edition is organized similar to the newspaper, with main sections devoted to news, sports, money, and life. News includes daily coverage from the newspaper as well as columns, opinion pieces, U.S. Supreme Court decisions, health and science, and local city guides. One of the newspaper's widely recognized strengths is sports and the online site mirrors that for Web readers. Baseball is one of the top attractions, drawing on Gannett's *Baseball Weekly,* but all other professional and amateur sports are also covered and updated regularly. The Money section offers business and financial world news as well as interactive opportunities to create and manage investment portfolios online. Small business tips are offered and consumers seeking information about automobiles find the site useful. The site's Life section reflects celebrities, television news and listings, movies and reviews, arts news, travel information, and a collection of columnists.

In addition to these sections, the site offers comprehensive and current weather information, a technology section, and a shopping section. Site visitors can take advantage of several interactive components of the site. These include "Talk Today," a chance for readers to chat online with experts and to subscribe to e-mail newsletters. Readers can subscribe to five different e-mail newsletters. These focus on the daily news briefing, daily technology report, weekly book publishing news, weekly automobile news and reviews, and weekly retail shopping deals and special offers. USAToday.com, like many other online news sites today, also provides wireless Internet users access to news and information through wireless telephones, PDAs, and handheld devices, and with pagers. Site users may also search archives of the newspaper dating back to 1994.

The Wall Street Journal

The Wall Street Journal introduced its online edition in 1995. It has been designed to serve the business and financial communities in the United States and abroad, but it also focuses on general news, says managing editor Bill Grueskin (personal communication, April 18, 2002). Most of its content originates in the printed edition of the newspaper, but 2 or 3%, he said, is original. The original content, Grueskin described, is mostly additional feature content. From the printed *Journal,* the site uses news from its various editions. While much comes from the U.S. edition, it also depends heavily on its European, Americas, and Asian editions. The site's producers also place news from the Dow Jones news service and other wires on the site.

Subscriptions include access to the site's market news, research and charting, current and historical stock quotes, use of site personalization tools, e-mail and news alert services, access to the *Barron's* online edition, and access to the most recent 30 days of news archives. The online site is organized somewhat differently from the print edition. For example, Marketplace, the second section of the printed edition, does not appear as an online section. Instead, articles from that section appear online by subject. The staff of about 80 full-time reporters and editors put the next day's edition of the newspaper on the site at about midnight (ET) the evening before publication. Early looks at coming stories are available as early as 7 p.m.

The site provides audio and video multimedia content and interactivity such as discussions and reader feedback. The site can also be personalized for users to select priorities and preferences for content display. Its ultimate strength, in the view of editor Grueskin, is the exclusive online content. Two years ago, the site gained attention for its coverage of the America Online-Time Warner merger. The site was the first news organization to break the story, posting it on its pages in the middle of the night.

ONLINE NEWS SERVICE MODELS

News media often follow a market model on the World Wide Web based on geography despite the fact that the Web and Internet have disposed of such limitations. This is often the case because the original print edition of a newspaper or television station is defined by a particular geographic orientation or signal limitation. There is much uncertainty about the economic and financial future of online news. At the least, the success or failure of online news depends on highly complex economics. Experts are unclear about whether online news, such as that from online newspapers, will become an economically viable business (Chyi & Sylvie, 2000). Some of this uncertainty and confusion is market related. What market does an online news organization serve? Is a particular online news site local? National? Global? There are four to five common levels of service: national/international (these may also be separated), regional, local, and specialized-niche (Chyi & Sylvie, 1999; Dibean & Garrison, 2001). These market models are summarized in Table 3.

There are at least four models for online news functions. These include the 24-hours-a-day news model, the community bulletin board site model, the supplementary news site model, and the exclusive news site model: These models are neither distinct nor mutually exclusive and two or more of them are generally found to exist within a single news site. In fact, many news sites at the national and international level in particular often use all four models at a given time. For example, if a weather threat or other natural disaster occurs, many online news sites update constantly, offer interactivity such as bulletin boards and other outlets for flow of personal information, offer supplementary coverage and information as it becomes available, and even create specialized sites devoted to coverage of that news story. This certainly happened in many communities in the United States and other nations during the U.S. war against Iraq in 2003.

The 24-Hours-a-Day News Model

Like the wire services and cable networks, online newspapers and broadcast television stations can be operated on an always-on-deadline or never on deadline circumstance (Farhi, 2000a, 2000b; Smith, 1999). Some authorities argue that for an online news operation, especially traditional newspapers that have gone online, to survive in the new century, they must adopt a twenty-four-hours-a-day news radio or wire service approach. News organizations that offer constantly updated content must not only invest heavily in their Web sites, but provide both depth and ongoing effort to keep content current. They must be responsive to the peaks and valleys of their region (usually 10 a.m. to 5 p.m. local time) and update most intensely during the heaviest of traffic periods (Farhi, 2000a, 2000b). Some online news organizations take this approach only when there are major breaking news stories, such as when the September 11, 2001, terrorist attacks occurred.

Table 3 Online Markets

Model	Description	Examples
National/International	Coverage focuses on national-international news and serves an international audience.	*USA Today* *Washington Post* *New York Times*
Regional	Coverage centers on large geographic region, such as a metropolitan area, with emphasis on local and regional news. Coverage includes substantial national and international news.	*Chicago Tribune* *Miami Herald*/Miami.com *Boston Globe*/boston.com *Atlanta Journal-Constitution*/ Access Atlanta *Houston Chronicle* *Arizona Republic* *Indianapolis Star/News* *Minneapolis Star Tribune* *Newark Star-Ledger*/New Jersey Online
Local-community	Serves small towns, communities or counties usually defined by smallest geographic area. Content is centered on local-area news, but other levels also.	Charlotte, Fla. *Herald-Sun* Fort Myers, Fla. *News-Press* Key West, Fla. *Citizen* Weekly newspapers Regional-city magazines
Specialized-niche	Coverage devoted to highly specialized or unique subjects and their audiences.	*Wall Street Journal* (business) *Florida Today* (space program) Financial Times (finance) Salon.com Subscription newsletters

Sources: Chyi & Sylvie, 1999; Dibean & Garrison, 2001; Grimes, 2002; Lasica, 2001; "Missed What the Top Analysts . . . ," 2000; Shaw, 1997.

The Community Bulletin Board Site Model

Many online newspapers have done more than rehash their print copy. Some have opted for the *community board site* model. *The Boston Globe* (http://www.boston.com/globe/), for example, was among the first major newspapers to design its online version to provide community information about events, featuring news about Boston's arts, weather, and commerce. University of Iowa online journalism scholar Jane B. Singer (2001) argues that, despite the Web's worldwide reach, most online newspapers have their best chance to succeed by emphasizing local news content.

The Supplementary News Site Model

Some print and broadcast outlets also take advantage of the limitless space on the Web to add additional material to online news stories that appeared in their print and broadcast versions. Many local newspapers have followed this model serving as local equivalents of CNN, taking advantage of online's timeliness by reporting local breaking news stories which, if not covered online, would be reported on local radio or television before the newspapers (Heyboer, 2000).

The Exclusive News Site Model

One model that print and broadcast news organizations have not adopted is the exclusive news site model. Exclusive news sites offer news not found in other forms, such as those other news media operated by the same news corporation or company. To date, however, few broadcasters, newspapers, or other news outlets have posted exclusives online and this model remains more a conceptual model than one commonly used on the Web.

GLOSSARY

Alert Notification about a breaking news event or story by an online news service, usually by e-mail.

Banner Form of advertising that usually contains a graphic product image or message on a Web site. The size of the ad may vary, but often spreads horizontally across the top or other part of the page. Some banner ads, however, are designed in a vertical shape.

Brand The recognizable name of a news company that has been given to a news product on the World Wide Web or other forms of communication on the Internet.

Convergence Implementation of multimedia combining text, graphics, audio, and video in online news presentation as well as the combination of divergent media technologies to produce a new medium, form, or method of communicating with audiences.

Flash Software that produces animation and video for Web pages and browsers. Content producers can quickly and easily prepare special effects through animation and other techniques, interactivity, sound, and video.

Headline For an online news site, brief one sentence or short phrase summaries of the news stories of the day, often with brief abstracts accompanying them. These features are linked to longer stories and coverage elsewhere on the site.

Hypermedia Links among any set of multimedia objects such as sound, animation, video, and virtual reality. Usually suggests higher levels of interactivity than in hypertext.

Hypertext Organization of news and information into associations that a user can implement by clicking on the association with a mouse. The association is called a hypertext link.

Independent An online news organization not grounded in traditional, branded news media and only found on the World Wide Web or Internet.

Interactivity For online news services, features such as e-mail feedback, instant opinion polls, chat rooms, and message boards that involve the site user in some form.

Market The audience served by an online news organization. The grouping may be determined by founding or originating media markets, by geographic boundaries, specializations, or defined in other ways.

Online journalism News presentation on the World Wide Web or other Internet-based formats. This includes news offered by traditional news organizations (e.g., newspapers, television stations and networks, and magazines) as well as nontraditional sources such as Internet service providers (e.g., America Online and CompuServe) and bulletin board services, Web "zines," and news groups.

Online news Collection and presentation of current events information on the Web or other Internet-based formats, including traditional and nontraditional sources such as those discussed above under "online journalism."

Package The combination of audio, video, graphics, and hypertext components that create a news story on a Web site.

Portal Entry points, or "doorways," to the Internet that often serve as home pages for users. These are offered by online newspapers, radio stations, and television stations and provide much more than just news content. These sites provide activity and entertainment listings, community calendars, government reference material, online interactive features, and multimedia content.

Producer For an online news site, this is an individual who develops news content for a Web site. These individuals work with all forms of news media to prepare the coverage for a story.

Shovelware Software that produces content of an online news site taken from the published content of a newspaper, television broadcast, and/or wire services.

Web edition The online edition of a traditional printed newspaper. These editions often include breaking news as well as extra, and often extended, content not found in printed editions.

Webcast Live news broadcast or other video program distributed on the Web by an online news or media organization.

Weblog Personal and noncommercial Web site prepared in dated log or diary format. The site is usually specialized and focuses on a single topic or subject area and offers information and observations from the author, contributors, or from other sites. The sites are often also called blogs.

Zine Independent publication on the Web that often offer idiosyncratic themes, irregular frequency of publication, ephemeral duration, and a noncommercial financial base.

CROSS REFERENCES

See *Convergence of Data, Sound and Video; Internet Navigation (Basics, Services, and Portals); Multimedia; Webcasting.*

REFERENCES

ABC Newsroom (2001). Retrieved May 13, 2002, from http://disney.go.com/corporate/press/wdig/abcnews/

Accolades (2002). Retrieved April 15, 2002, from http://www.usatoday.com/a/adindex/omk/accolades.htm

Advertising opportunities (2002). Retrieved May 13, 2002, from http://abcnews.go.com/ad/sponsors.html

An evolving medium (1999, April 26). *Online Newshour.* Retrieved April 28, 1999, from http://www.pbs.org/newshour/bb/media/jan-june99/lasica_qa.html

Barringer, F. (2001, August 27). Growing. *The New York Times,* pp. C1, C6.

Blankenhorn, D. (1999, March). Publishers seek ways to win the Web news war. *Boardwatch Magazine, 13*(3), 88–90.

Burgess, J. (1993, February 14). Firms face questions of technology's timing, cost. *The Washington Post,* p. H1.

Chyi, H. I., & Sylvie, G. (1999, August). *Opening the umbrella: An economic analysis of online newspaper geography.* Paper presented to the Association for Education in Journalism and Mass Communication, New Orleans, LA.

Chyi, H. I., & Sylvie, G. (2000). Online newspapers in the U.S.: Perceptions of markets, products, revenue, and competition. *Journal on Media Management, 2*(11), 69–77.

Dibean, W., & Garrison, B. (2001). How six online newspapers use Web technologies. *Newspaper Research Journal, 22*(2), 79–93.

Donatello, M. (2002, May). What consumers tell us about paying for news online. *EContent, 25*(5), 36–40.

Farhi, P. (2000a). The dotcom brain drain. *America Journalism Review, 22*(2), 30.

Farhi, P. (2000b). Surviving in cyberspace. *America Journalism Review, 22*(7), 22–27.

Fee, F. (2002, March). *New(s) players and new(s) values? A test of convergence in the newsroom.* Paper presented to the Newspaper Division, Southeast Colloquium, Association for Education in Journalism and Mass Communication, Gulfport, MI.

Fouhy, E. (2000). Which way will it go? *American Journalism Review, 22*(4), 18–19.

Fredin, E. S. (1997). Rethinking the news story for the Internet: Hyperstory prototypes and a model of the user. *Journalism & Mass Communication Monographs, 163,* 1–47.

Fredin, E. S., & Prabu, D. (1998). Browsing and the hypermedia interaction cycle: A model of self-efficacy and goal dynamics. *Journalism & Mass Communication Quarterly, 75*(1), 35–54.

Gannett (2002). *About Gannett: Company profile.* Retrieved April 19, 2002, from http://www.gannett.com/map/gan007.htm

Gates, D. (2002, May 1). The future of news: Newspapers in the Digital Age. *Online Journalism Review.* Retrieved May 2, 2002, from http://www.ojr.org/ojr/future/1020298748.php

George, W. (2000, January 11). The AOL/Time Warner merger: What it really means to the markets and Apple. *The Mac Observer.* Retrieved August 21, 2002, from http://www.macobserver.com/news/00/january/000111/aoltimewarner.shtml

Grimes, C. (2002, February 6). Premium payments may boost Web revenues. *Financial Times,* p. 3.

Heyboer, K. (2000). Going live. *American Journalism Review, 22*(1), 38–43.

Higgins, J. (2000, July 14). Staying afloat on WEBLOGS: Sites are water wings for surfers sinking in sea of cybermadness. *Journal Sentinel Online.* Retrieved August 11, 2002, from http://www.jsonline.com/enter/netlife/jul00/weblogs16071400.asp

Houston, F. (1999, July/August). What I saw in the Digital Sea. *Columbia Journalism Review, 38*(2), 34.

Internet growing as news medium, at times exceeding traditional media usage. (2002, January 7). Global News Wire and Business Wire, *Financial Times,* n.p.

Keefe, R. (1996, October 13). New *Tribune* president arrives during time of transition. *St. Petersburg Times,* p. 1H.

Koerner, B. (2001, July/August). Click here for Britney! *The Washington Monthly, 33*(7/8), 25–30.

Kramer, S. D. (2002, April 9). The elephant in the jury room: Maybe it's time to be as visionary as Joseph Pulitzer. *Online Journalism Review.* Retrieved April 13, 2002, from http://www.ojr.org/ojr/workplace/1018398261.php

Langfield, A. (2001, December 21). Net news lethargy: Most sites fail to make use of the medium's main strength-speed. *Online Journalism Review.* Retrieved January 2, 2002, from http://ojr.usc.edu/content/story.cfm?request = 676.

Lasica, J. D. (1999). Online, papers can speak volumes. *American Journalism Review, 21*(2), 66.

Lasica, J. D. (2001, September 20). A scorecard for net news ethics: Despite a lapse related to the terrorist attack, online media deserve high marks. *Online Journalism Review.* Retrieved January 7, 2002, from http://ojr. usc.edu/content/story.cfm?request = 643.

Marlatt, A. (1999, July 15). Advice to newspapers: Stop the shoveling. *Internet World.* Retrieved from http://www.iw.com/print/current/content/19990515-shoveling.html

McMillan, S. J. (1998, September). Who pays for content? *Journal of Computer Mediated Communication, 4*(1). Retrieved December 6, 2001, from http://www.ascusc.org/jcmc/vol4/issue1/mcmillan.html

Miller, R. (2001, September 14). From niche site to news portal: How Slashdot survived the attack. *Online*

Journalism Review. Retrieved December 6, 2001, from http://ojr.usc.edu/content/story.cfm

Missed what the top analysts are saying about our digital future? Catch up here. (2000, February 21). *The Guardian*, p. 62.

MSNBC rated #1 online news provider for February 2002. (2002). Retrieved May 13, 2002, from http://www.msnbc.com/m/info/press/02/0319.asp

Nicholas, D., & Huntington, P. (2000). Evaluating the use of newspaper Web sites logs. *Journal on Media Management, 2*(11), 78–88.

Noack, D. (1998, May 1). Fees fail. *Editor & Publisher, 132*(28), 20.

Outing, S. (Fall 2001). *Crisis notes from the online media, Poynter Report* (pp. 48–49). St. Petersburg, FL: Poynter Institute for Media Studies.

Palser, B. (2000). Not quite ready for prime time. *American Journalism Review, 22*(6), 84.

Palser, B. (2001a). Retooling online news. *American Journalism Review, 23*(2), 70.

Palser, B. (2001b). Pay-per-click. *American Journalism Review, 23*(8), 82.

Perlman, J. (1999, May 6). Print sites still wary of chatting it up. *Online Journalism Review*. Retrieved April 12, 2002 from http://www.ojr.org/ojr/business/1017968634.php

Picard, R. (2000). Changing business models of online content services: Their implications for multimedia and other content producers. *Journal on Media Management, 2*(11), 60–68.

Pryor, L. (2000, June 30). Some guidelines from one of online news' walking wounded. *Online Journalism Review*. Retrieved December 6, 2001, http://ojr.usc.edu/content/story.cfm

Quint, B. E. (November 1994). "Extra! Extra!" *Wilson Library Bulletin,* 69, 67–68.

Rauch, J. (2002, August). Hands-on communication: The rituals of Web publishing in the alternative zine community. Unpublished paper, Association for Education in Journalism and Mass Communication, Miami Beach, FL.

Rich, C. (1999). *Creating online media*. Boston: McGraw-Hill College.

Richardson, T. (2002, February 10). CJOnline nets two EPpy Awards; was finalist in four categories. *CJOnline*. Retrieved August 20, 2002, from http://www.cjonline.com/stories/021002/com_eppies.shtml

Rieder, R. (2001). At the crossroads. *American Journalism Review, 23*(2), 6.

Samoriski, J. (2002). *Issues in cyberspace: Communication, technology, law, and society on the internet frontier*. Boston: Allyn and Bacon.

Scasny, R. (2001, November 23). "Online news users have to pay: Change the product, give it more value and paid access will work," *Online Journalism Review,* accessed December 6, 2001, http://ojr.usc.edu/content/story.cfm?request = 663

Shaw, D. J. (1997, June 17). The media.com: Revolution in cyberspace. *Los Angeles Times,* Home edition, p. A1.

Singer, J. B. (2001, Spring). The Metro Wide Web: Changes in Newspaper's Gatekeeping Role Online, *Journalism & Mass Communication Quarterly, 78*(1), 65–80.

Smith, S. (2002, February). The free lunch is over: Online content subscriptions on the rise. *EContent, 25*(2), 18–23.

Smith, S. (2002, April). Content at your service. *EContent, 25*(4), 44–45.

Smith, T. (1999, April 26). A new source of news. *Online Newshour*. Retrieved April 28, 1999, from http://www.pbs.org/newshour/bb/media/jan-june99/inews_4-26.html

South, J. (1999, June 11). Web staffs urge the print side to think ahead. *Online Journalism Review*. Retrieved April 12, 2002, from http://www.ojr.org/ojr/business/1017968570.php

Stepp, C. S. (1996, April). The new journalist. *American Journalism Review*. Retrieved August 20, 2002, from http://www.ajr.org/Article.asp?id=833

Stone, M. (2002, February 1). Breaking stories and breaking even. *Online Journalism Review*. Retrieved April 13, 2002, from http://www.ojr.org/orj/workplace/1015015509.php

Strupp, J. (1999, July 3). Welcomed visitors. *Editor & Publisher, 132*(27), 22–28.

The impact of the Internet. (2000). *American Journalism Review, 22*(2), 4–5.

Trombly, M. (2002, May). Looking for online dollars: News providers are finding ways to make their Web sites profitable. *Quill, 90*(4), 18–21.

Walker, L. (2001, June 24). Moving from logs to blogs: Technology and you. *Spokane Spokesman-Review*. Retrieved August 10, 2002, from http://www.spokesmanreview.com/sections/sports/stateb/2002/stateb.asp?ID=2002/blog/blog-primer

Weise, E. (1997). Does the Internet change news reporting? Not quite. *Media Studies Journal, 11*(2), 159–163.

Whittaker, J. (2000). *Producing for the Web*. London: Routledge.

Online Public Relations

Kirk Hallahan, *Colorado State University*

INTRODUCTION

Online public relations involves the application of Internet technologies by organizations to communicate and build relationships with key publics: customers and consumers, employees, investors and donors, community members, government, and the news media. Beginning with the popularization of the Internet in the mid-1990s, public relations units within corporations, non-for-profit organizations, and public institutions have embraced the Internet widely. The result has been a dramatic change in how public relations is practiced.

Nature of Public Relations

Public relations is a management staff function that facilitates communication between an organization and its key constituencies. Public relations can go by many names, including *corporate communications, public affairs, development,* and *public information.* Importantly, public relations is the not only the organizational unit responsible for managing communications between an organization and key publics because communication is the constitutive process of management. Nonetheless, public relations units facilitate organizational communication and relationship building two ways: through counsel and communications.

Counsel

Public relations provides *counsel* to an organization on problems and opportunities that confront it. This counsel includes feedback based on formal and informal research about public opinion, as well as recommendations on communications strategies that can be pursued across the organization to respond to audience needs, concerns, and interests and to speak effectively with a "single voice."

Online technologies have enhanced the capability of PR practitioners to conduct research and counsel organizations. *Environmental scanning* is the generic term for the process used to identify relevant concerns being discussed in the public, media, or public policy arenas of opinion. *Cyberscanning* adapts this technique to the surveillance of topics being raised in Web sites, discussion groups and chat rooms on the Internet. Whereas scanning is a more general form of intelligence gathering, *environmental monitoring* and *cybermonitoring* involve tracking specific, searchable topics. Public relations professionals routinely review Internet content themselves by tracking specific sites or key terms using search engines. Organizations also hire professional *Web monitoring services* that compile daily reports about Web and discussion group content.

Communications

Public relations also conducts strategically planned communications programs to reach key audiences. Today these programs integrate an array of both online and offline activities. These programs are grounded in clearly articulated organizational goals and communication objectives, involve carefully crafted strategy, and emphasize the objective evaluation of results. Some communications and relationship-building efforts are intended primarily to make audiences more *knowledgeable* about an organization, products and services, candidate(s), or causes. Other programs strive to shape, reinforce, or change *attitudes* among key publics about an organization's policies or practices or to enhance an organization's overall reputation. Finally, many public relations programs are intended to prompt specific *actions*—to influence how people buy, invest or donate, work, vote, or care for themselves or others by avoiding risks.

Traditional Public Relations

Six principal specialties comprise the PR practice. These specializations are organized around the typical organization's key constituencies and include *employee relations, customer or consumer relations* (sometimes referred to as *marketing public relations), investor relations* (corporations) or *donor relations* (not-for-profit organizations), *community relations*, and *government relations*. In addition, almost all public relations departments are responsible for *media relations* (incorporating publicity), which operates as a conduit to reach all of these constituencies. Some manufacturing organizations also have special programs directed to *suppliers or vendors* (supplier relations).

Within each of these specializations, programs fall into four broad categories. *Promotional programs* are short-term efforts to influence actions, such as purchasing a new product or voting for a particular candidate. *Relationship-building* programs are longer term efforts undertaken to solidify patterns of behavior that benefit both the public and the organization. *Crisis communications* programs entail short-term responses to extraordinary triggering events that create uncertainty in an organization's operating environment. Organizations strive to minimize such disruptions and assuage concerns by providing timely and factual information. Finally, *issues management* programs focus on anticipating trends in public opinion and on responding to disputes or challenges by activists to an organization's philosophies or practices.

Online Public Relations

The advent of the Internet has dramatically altered the composition of PR programs and the mix of communications tools used by practitioners. As a result, many practitioners have been required to quickly learn about Internet technologies. Online or interactive media now constitute one of five major groups of communications media used in public relations (Figure 1).

Public relations traditionally relies on publicity coverage in the news and entertainment portions of *public media* (newspapers, magazines, radio and television) to communicate with the public-at-large and with business-to-business audiences (also known as trade audiences). Practitioners often augment these publicity-generating efforts with *controlled media,* which are tools manufactured and distributed to target audiences, such as brochures or annual reports. Public relations also depends heavily on *events,* or gatherings of key constituents such as speeches, conferences, awards ceremonies, and rallies. Finally, public relations practitioners employ *one-on-one communications,* where a representative of an organization interacts directly with individual members of key publics. One-on-one communication is prevalent when comparatively small numbers of people are important to reach. Examples include lobbying, fundraising among major donors, consumer affairs and customer relations, and community outreach.

Online and interactive media fill a critical need in many public relations programs by providing the ability to respond to queries and by providing a potentially endless amount of information about an organization and its product and services, candidates, or causes. Online media also can enhance involvement with an organization through education and entertainment made possible by online media's interactive character. Online information can be updated by organizations instantaneously and accessed by users 24 hours a day, 7 days a week from anywhere in the world where Internet access is available. In addition, information can be *personalized* to the inquirer using *cookies, user profiling,* and *database* technologies.

Interactive media have reshaped the use of other public relations media as well. Online media allow organizations to be less dependent on public media to supply information to large audiences. Online media probably will not replace newspapers, magazines, radio, or television, however. Instead, public media will remain important tools when it is necessary for organizations to create broad public awareness, to communicate with the public quickly, or to respond to news events.

Interactive media also make organizations less dependent on controlled media. Web sites, for example, have reduced the direct and incremental costs of distributing printed matter. These materials otherwise must be given out through literature racks, personal visits, postal mail, or package delivery services. Interactive media are not likely to replace these communications tools, however. Similar to public media, controlled media will continue to used in situations when it is necessary to "push" information into the hands of audiences instead of waiting for audiences to be "pulled" into a Web site, discussion group, or online chat.

Online communications can replace some events by creating opportunities for people in remote locations to come together online. These include webcasts, seminars and workshops, bulletin boards, discussion groups, online chats, and Web conferencing (two-way video and audio online exchanges between participants). Such applications are not yet fully developed, however, and their potential is less understood.

The Internet provides additional options for one-on-one communications between organizational representatives and individual members of constituencies. In particular, personal e-mails can complement personal meetings, telephone calls and postal mail to facilitate and speed up exchanges. This might be especially important as more people adopt wireless phones and personal data appliances that can deliver Internet content.

PUBLIC RELATIONS WEB SITES

Web sites represent the most visible and widely used Internet tool in public relations. When organizations first experimented with Web sites in the early 1990s, their first crude attempts were usually produced by "techies" in computer operations departments. Many computer

Media Group	Key Strategic Uses	Special Challenges for Successful Use
Public Media		
Newspapers, magazines, radio, television. Also out-of-home media, yellow pages and directory advertising, venue and specialized media	Build broad public awareness	Competing for attention Promoting newsworthiness
Interactive Media		
Web sites, CD-ROMs, automated telephone response systems, data bases (allowing file transfers), computerized bulletin boards, news groups and chat rooms, e-mail	Respond to inquiries Enhance involvement through education, entertainment.	Promoting access and use Facilitating easy navigation and location of information
Controlled Media		
Brochures, video brochures, annual reports, newsletters and other periodicals, books, direct response, point-of-purchase displays, advertising specialties	Provide detailed, promotional information.	Designing attractive and engaging materials Assuring distribution to targeted audiences
Events		
Speeches, trade shows and expositions, exhibits, conferences and meetings, demonstrations, rallies and protests, sponsored events, observances, sweepstakes and contests, recognition and awards presentations	Reinforce existing beliefs and attitudes	Guaranteeing attendance by targeted audiences Heightening participation and involvement levels
One-on-One Communications		
Letters, telephone responses (inbound), telemarketing (outbound), soliciting and personal selling, consumer affairs, community outreach, lobbying	Negotiate disputes, resolve conflicts. Obtain commitments.	Delegating authority to negotiate on behalf of organization Fostering positive personal dynamics between parties

Figure 1: Five major groups of communication media. From Hallahan (2001a).

specialists—and their bosses—quickly realized that technical staffs were ill-equipped to deal with content issues. It was then that public relations units became involved. Many managers recognized it was important that online content be managed so that Web sites reflected the organization's branding, positioning, and corporate identity. Many PR professionals approached the Internet without much knowledge but soon began to integrate online communications into PR program planning,

Most organizational Web sites have gone through several incarnations since the commercialization of the Web. Between 1994 and 1996, organizational Web sites were mostly *informational sites.* Web content resembled electronic brochures that provided descriptive information about the organization. Between about 1996 and 1998, organizations recognized the value of *interactivity.* Organizations added e-mail links, fill-in forms, interactive games, and the ability to conduct non-secured transactions. Finally, about 1998, the *transactional* potential of e-commerce was recognized. At that point, many organizational Web sites were transformed from primarily information channels to become dual-function channels that perform both communications and distribution functions.

Organizational sites are *sponsored Web sites* that are dedicated to the interests to a single organization that underwrites the costs of operations in other to promote its own interests. Most sponsored sites do not accept third-party advertising but can use advertising space to promote the organization's own products, services, or causes. Most organizational sites operate as communication utilities that that serve a variety of purposes, including the needs of marketing, human resources, and public relations.

Corporate Web Pages

Today almost all large and medium-sized organizations operate Web sites to communicate and to build relationships with key external publics. Web sites are important platforms to establish corporate identity and to supply information to media, customers and consumers, and investors.

Online News Rooms

One of the most important uses of the Internet in public relations is to allow journalists to access current news releases, background information, fact sheets, and other materials about the organization. Most major organizations now devote a special section of their public Internet sites to *newsrooms* containing these company announcements and other information. In addition, news archives can be outsourced on the sites of one of the major public relations news wire distribution services, such as Business Wire (http://www.businesswire.com) or PR Newswire (http://www.prnewswire.com).

In the mid-1990s, many large organizations began to post news releases online on a regular basis. These announcements previously had been distributed by mail, messenger, facsimile, or private wire services. News releases previously were not archived electronically and were not conveniently available to the public-at-large.

Today journalists routinely download Web-supplied releases (in addition to news received via e-mail and the private wire services). Journalists can quickly transfer materials to a newsroom's word processing system, thus making it easier to incorporate supplied materials in stories.

In addition to text, reporters and editors can download other types of files, including photos, audio, and video. Photos reproduced with 300 or more dots-per-inch (dpi) can be printed with sufficient resolution by most newspapers and magazines (600 dpi is preferred). Audio actualities or "sound bites" similarly meet ready-to-broadcast standards. Because of bandwidth restrictions, the use of video files is more limited, although streaming video is being used by organizations with increased frequency. Most editors and reporters now are limited to watching streaming video for their own education, or to preview available video footage that can be downloaded from a satellite. Direct broadcast-quality video downloads are likely to be used extensively in the future as bandwidth availability increases.

Journalists have adopted the Web and e-mail as basic reporting tools (Hallahan, 1994). Reporters today spend as many as three hours a day on the Internet accessing Web materials and processing e-mail. More than 90% of journalists use the Web regularly. Most larger media organizations provide news workers with online access, although Internet accessibility flags at some smaller media outlets. Significantly, little evidence suggests that the Web will replace traditional reporting tools. One study suggested that corporate Web sites ranked only fourth in preference as a news gathering tool—behind the telephone, personal interviews, and news releases (Hachigian & Hallahan, 2003). In part, this is because online newsrooms often still lack materials that journalists want, including basic contact and telephone information.

One problem with online newsroom is that information is accessible to people other than journalists. Early efforts to restrict access by requiring journalists to register as users and to use a password met with resistance and subsequently have been abandoned. Notable exceptions are the newsrooms operated by large public relations agencies concerned about documenting usage for clients. One advantage of controlled access is to allow embargoed materials to be released selectively to reporters who confront early deadlines.

Investor Relations Sites

Investor relations is a PR specialty that has especially benefited from the Internet's ability to deliver large volumes of information on demand. Investor relations professionals produce detailed annual reports, quarterly reports, company profiles, and so on. These materials can be posted online then read or downloaded by investment professionals and individual investors in HTML (hypertext markup language) or .pdf formats. In addition, various government filings, such as 10-K, 10-Q, and other reports to the Securities and Exchange Commission, can be posted online or made accessible with link to the SEC's financial document site (http://www.edgar.gov).

Investor relations practitioners also can link or post analysts' research reports and electronic presentations before investment groups. Investors also can access scheduled live Webcasts or view them on-demand later, or participate in audio investment discussions (online versions of telephone conference calls) sponsored by organizations. Such participation is consistent with the SEC's mandate in Regulation FD, adopted in 2000, that organizations make information more widely available to all investors. Organizations are experimenting with the distribution of proxy materials online, and with conducting annual meetings online. The traditional investment "road show" tour undertaken when an initial public stock offering is made is also being replaced with a "virtual road show" where regional analysts and registered representatives can learn about soon-to-be public companies through Webcasts.

Consumer Promotion and Education Sites

Many organizations operate sites targeted to consumers. Not surprisingly, these sites often bear characteristics that

that meld public relations (information and relationship-building) and marketing (purely promotion).

Product Promotion Sites

A growing number of PR- and marketing-sponsored customer sites promote the benefits of products and services by educating consumers about how products might be used. Food companies, for example, feature sites with recipes, cooking and other household tips. Other sites, such as for soft drink manufacturers, feature prize giveaways and redemptions, contests, and interactive entertainment. Many of these especially target young people.

A particular genre of consumer sites is the *fan site*. Movie and other large media companies now routinely post sites about forthcoming productions. Fans can preview scenes or book chapters, download photos and music, read background about the performers or authors, play specially created games, send fan mail, or chat with others who share their infatuation. Some of these sites are grassroots sites created by fans themselves. Fan sites have become integral parts of promotional campaigns for movies, television shows, and books. The goal of most of these sites is to generate "buzz" or word-of-mouth promotion.

Education and Advocacy Sites

Other public relations sites are intended to educate the public on topics of particular concern to organizations. Young people and adults alike can learn about current issues or problems, such as the dangerous of environmental pollution or the value of conservation, through games, exercises, quizzes, and streaming video and audio. Although these sites might provide a valuable learning experience in some instances, the motives of the sponsors are often self-serving. Thus the accuracy and objectivity of these sites are suspect.

Health Sites

Web sites related to health issues are among the most popular types of consumer sites. Health-related sites are sponsored by a wide range of organizations—from commercial health care providers and pharmaceutical and medical device manufacturers to government and advocacy groups devoted to the eradication of particular maladies. Health-related sites have shown themselves to be effective tools for patient education and self-diagnosis, and for the delivery of social support from others through Web-based e-mail exchanges, discussion groups, chats, and so on. The accuracy and objectivity of information on many health sites has been challenged, however, leading to calls for self-regulation and possibly governmental oversight.

Fundraising/Development Sites

Many not-for-profit organizations have incorporated Web sites in their promotional and fundraising efforts to solicit funds. These sites capitalize on the Internet's secured transaction capabilities. Web sites, for example, allow disaster relief agencies to post information on "how to help" almost immediately after devastation occurs. Many advocacy groups also use their sites to promote their causes, enlist members and solicit contributions. Membership recruitment and fundraising represent two applications of e-commerce technology in traditional public relations functions.

Employee Relations Sites and Intranets

Online communication has dramatically changed employee relations and internal communications, particularly within organizations in which large numbers of employees can access computers. Many employers begin by using the organization's public Internet site as an employee recruitment tool. Once an employee is hired, organizations use closed-access Web-based systems or *intranets* to communicate a wide array of employee relations-related information.

Typical public relations-oriented content on an Intranet includes employee newsletters and updates, memoranda, streaming audio and video clips on organizational developments, employee benefits materials, policies and procedures manuals, guides to employee counseling and assistance programs, employee Q&A bulletin boards, and promotional bulletins on recreational and social activities.

Intranets clearly are not substitutes for one-on-one or group communications at work. But Web sites can provide employees with quick and easy access to frequently asked questions traditionally answered by public relations or human resource departments using other media. Organizations also use intranets as part of knowledge management programs to share organizational intelligence and thus improve productivity (Weitzel & Hallahan, 2003) The advantages are particularly important for organizations that operate in multiple locations (including global organizations) and for employees who travel frequently or telecommute.

Supplier or Distributor Sites and Extranets

One of the newest forms of Web sites in public relations is specifically directed to vendors of materiel as well as wholesalers, distributors, and retailers in an organization's channel of distribution. *Extranets* are extended controlled-access systems similar to an intranet. Most extranets began as outgrowths of earlier data-based inventory control systems intended to facilitate just-in-time deliveries and supply other operations-related information to suppliers or distributors. The advent of Web-interfaces, however, enables organizations to provide business partners with a broader range of content. Examples include information on corporate developments, recognition of performance, and other relationship-building messages.

Government and Political Sites

Government institutions, individual lawmakers and candidates for political office have embraced Web sites as a way to communicate with constituents and prospective voters. Government institutions post a wide range of citizen information online to facilitate their work and to inform the public about government activities. Examples include agendas and minutes for public meetings; notices of impending hearings or rule making; enacted laws, regulations, and ordinances; forms and applications; and basic questions and answers about government operations. Many use feedback mechanisms, such as e-mail and fill-in forms, to monitor citizen feedback and help citizens resolve problems.

Similarly, lawmakers and candidates have found the Web to be an effective tool to articulate their positions on political issues, to inform supporters about their activities, and to solicit public comments. Ironically, not all lawmakers or candidates necessarily relish their vulnerability to constant constituent challenges or criticisms. Lawmakers are often inundated with e-mail lobbying them to vote a particular way on legislation and often find it impossible to wade through the volume of e-mails received. Campaign organizers also have begun to use the Web as a fund-raising tool in the same way as not-for-profit organizations.

OTHER INTERNET APPLICATIONS

In addition to Web sites, public relations uses a variety of other Internet tools to disseminate information and to build relationships.

E-mail

E-mail allows organizations to send messages to key constituents, both individually and in groups. Broadcast e-mail lists can be purchased from brokers, drawn from organizational databases, or developed through solicitations seeking permission to send information to prospective addressees (permission marketing). In addition to text and graphical messages, audio e-mail and video e-mail allow recipients to link to Web sites to hear or view messages.

Media relations specialists use e-mail to distribute news releases as well as story ideas or "pitches" to reporters and editors. Journalists similarly have found e-mail a valuable way to seek information from sources and to conduct interviews where sources provide written responses to questions.

An important new extension of basic e-mail are *e-newsletters,* or compendia of news and information items distributed periodically to an e-mail list in lieu of a printed publication. Newsletters can be distributed in either text or HTML formats, with links that provide supplemental information. Recipients can scan the contents, select items of interest, and discard others.

Listservs provide another useful group e-mail technology where organizations can send messages to a controlled-access list of recipients. An addressee then can respond to all other recipients.

Discussion Groups, Chats, Online Seminars

Some organizations encourage the use of asynchronous bulletin boards and synchronous (real-time) chat rooms to facilitate discussions on topics relevant to the organization. Companies ranging from software manufacturers to motorcycle manufacturers encourage users and customers to come together online to discuss topics of common interest, to share experiences, and to solve problems collectively. Bulletin boards also are being employed in health communications to link sufferers of various maladies or their caregivers. Some employers also sponsor chat sessions to solicit and encourage online discussions by employees about new ideas and suggestions for improvements and so on. Sponsored discussion groups and chats are supplemented by a huge number of user-originated groups and chats without official sponsor sanction. About a dozen discussion groups serve the public relations community alone.

An extension of sponsored and user-organized chats are *chat tours* on major Web portal or media sites, such as Yahoo!, AOL, ESPN.com, and MSNBC.com. PR representatives can arrange for clients to "appear" online in the same way clients might make guest appearances on television and radio talk shows. Publicists can book a client, then promote the appearance via online and traditional media. Clients are "trained" on how to conduct the tour and work with a ghost typist, who actually transcribes the guest's oral responses.

Alternatively, *Web events,* such as speeches, presentations, seminars, and workshops, are being used by public relations and marketing departments to educate targeted groups on selected topics. Many scheduled *Webinars* and online workshops employ real-time chats, combined with Webcasts and electronic presentations using software such as Microsoft PowerPoint. Such conferences are even used to train media spokespersons. Alternatively, events can be held at designated host sites, where participants use banks of computers in a room also equipped to receive video feeds via satellite or audio conference calls over phone lines.

File Transfers

Public relations officers for government and other large institutions can supply large files of data, text, or graphics using *file transfer protocols*. FTP permits computer-to-computer delivery of government agency databases, large reports, corporate financial information, photos and multimedia, and promotional games. Anonymous FTP sites (which do not require user identification) can be located by searching specialized search engines. Although traditional Internet and limited-access extranets allow viewing the information, FTP downloads provide savvy researchers with the ability to manipulate and analyze data to their specifications.

Remote Kiosks and CD-ROMs

Organizations also can engage in online exchanges by stationing terminals in kiosks in public facilities. Kiosk terminals normally rely on phone or dedicated connections and allow users to interact with a central system by using a keyboard or touch screen. Kiosks also can operate as stand-alone personal computers.

CD-ROMs allow organizations to circumvent the problems of slow Internet downloads by delivering large files of information on an easy-to-insert compact disk that can be played on a personal computer. Although CD-ROMs are not used extensively, CD-ROMs are ideal to deliver digital information that requires large storage capacity—books, directories and catalogs, annual reports, multimedia presentations, video clips, and interactive games. Although expanded bandwidth might reduce the need to rely on them in the future, CD-ROMs can be reproduced and delivered inexpensively to "push" information into the hands of users. Miniature CD-ROMs fit in a ordinary business envelope or coat pocket and substitute as electronic business cards for some organizations, complete with multimedia presentations.

ORGANIZATIONAL–PUBLIC RELATIONSHIPS ONLINE

Although the distribution of information is important, the Internet's more important contribution to public relations is its ability to establish relationships with users. An *organizational–public relationship* can be defined as a routinized, sustained pattern of behavior among individuals related to their involvement with an organization. Some people would argue that relationships established or maintained primarily online are less robust than relationships established through direct, personal contact. Nonetheless, online relationships can be strategically important to organizations. Many online relationships operate in tandem with offline relationships and thus are part of a total organizational–public relationship.

Potential for Creating Relationships Online

Media research suggests that some users of computers (and other media) do not fully recognize that organizations are the ultimate sources of online communications, that is, people confuse the Internet or a Web site as the creator or source of information. Today's increasingly sophisticated Internet user understands how the Internet operates, however, and that organizations of all types (in addition to individuals) produce Internet content.

Online users constitute important new publics for organizations in cyberspace, and thus have redefined organizational–public relationships and the very nature of organizations themselves. Similarly the inherent interactive features of online communications can foster interaction. This makes online communications not merely a means toward an end (transmitting information), but an end in itself (a relationship builder).

A central concept in the PR field today is the notion that public relations ideally is practiced as a *dialogue* involving two-way symmetrical communication between an organization and its publics. The Internet's inherent interactivity provides the potential for balanced exchanges, the equalization of power relationships in society, and the development of a sense of community. This potential reciprocity is readily evident in online tools such as e-mail and discussion groups. Many organizational Web sites have yet to fully capitalize on online's potential for feedback and two-way exchanges, however. In fact, organizations often fail to respond effectively to even the simplest online inquiry.

Factors Shaping Organization–Public Communications Online

Figure 2 is a useful model for understanding how organizational–public relationships are created online. The antecedents in the left column represent factors that shape the process, particularly the motivation and ability of organizations and people to enter into organizational–public relationships online.

Organizational Factors

These include the organization's commitment to, purpose, and knowledge about using online media. Organizations have adopted online communications in varying degrees to perform different communications functions. They have also demonstrated varying degrees of expertise and sophistication in adopting new online techniques. More advanced organizations are more likely to use online communications to their advantage in building relationships.

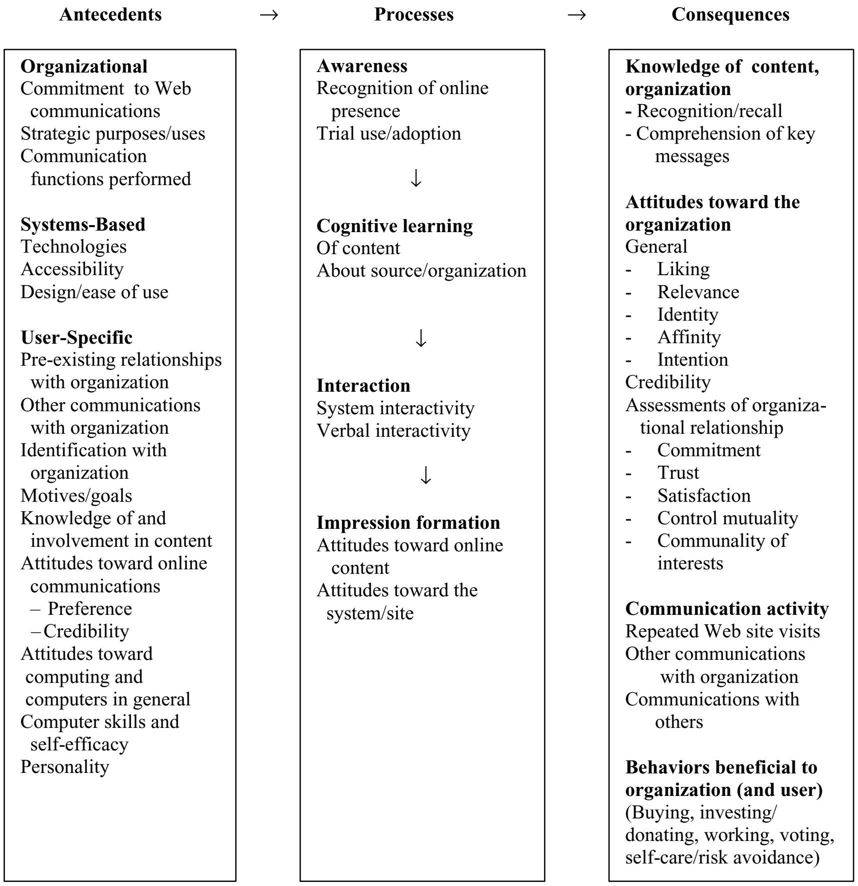
Systems Factors

These include the nature of the technologies used by organization and the hardware and software available to the user. User accessibility to online media is a particularly important factor. As of mid-2002, only about 60% of the U.S. population had access to the Internet. Meanwhile, access worldwide is limited for various political, economic, technological, and cultural reasons. The resulting "digital divide" suggests that online communications are not appropriate, nor effective, for organizations to reach all audiences. System usability also moderates the ability of online communications to nurture relationships. Systems with poor designs, poor navigation, or deficient content are less likely to be used effectively by members of key publics and thus dissuade relationship building.

User Factors

These include a wide range of variables that users, as individuals, bring to the online communications experience. Among factors that moderate organizational–public relationship building online are a person's

- Extent of use of online technology;
- Preexisting relationship, role, and identification with the organization;
- Concurrent, offline communications with the organization;
- Motivation sought and gratifications received (e.g., whether one's purpose is to conduct transactions, interact with others, escape reality, or obtain social support, etc.);
- Skill and confidence (computer expertise, computer self-efficacy, and lack of computer anxiety);
- Knowledge of content (expert vs. novice);
- Involvement in content (relevant vs. inconsequential);
- Attitudes toward the Internet in general, including preference for and assessments of the Internet's credibility as an information source;
- Attitudes toward computing and computers in general; and
- Personality, including but not limited to cognitive ability, cognitive style, personal innovativeness; also age and gender.



Antecedents → **Processes** → **Consequences**

Organizational
Commitment to Web communications
Strategic purposes/uses
Communication functions performed

Systems-Based
Technologies
Accessibility
Design/ease of use

User-Specific
Pre-existing relationships with organization
Other communications with organization
Identification with organization
Motives/goals
Knowledge of and involvement in content
Attitudes toward online communications
– Preference
– Credibility
Attitudes toward computing and computers in general
Computer skills and self-efficacy
Personality

Awareness
Recognition of online presence
Trial use/adoption

↓

Cognitive learning
Of content
About source/organization

↓

Interaction
System interactivity
Verbal interactivity

↓

Impression formation
Attitudes toward online content
Attitudes toward the system/site

Knowledge of content, organization
- Recognition/recall
- Comprehension of key messages

Attitudes toward the organization
General
- Liking
- Relevance
- Identity
- Affinity
- Intention
Credibility
Assessments of organizational relationship
- Commitment
- Trust
- Satisfaction
- Control mutuality
- Communality of interests

Communication activity
Repeated Web site visits
Other communications with organization
Communications with others

Behaviors beneficial to organization (and user)
(Buying, investing/donating, working, voting, self-care/risk avoidance)

Figure 2: Creating organizational-public relationships online. From Hallahan (2003a).

Processes of Building Relationships Online

Awareness and Adoption

Online relationships cannot be created until people recognize that an organization has an online presence and actually access or try the system. Relationship building thus begins by organizations' promoting and attracting users. Importantly, many users have come to *expect* that organizations have an online presence. A person's acceptance of a new technology such as online communications is determined by perceived usefulness and ease of use—two factors that organizations must promote. Other important factors include the relative advantage provided, simplicity, compatibility with other systems, ease of trial, and ready observability.

Cognitive Learning

Cognitive learning involves processing a minimal level of content and making judgments about the content and the source.

Content learning is facilitated by the simplicity and organization of content and the system itself—the simpler and more intuitive, the better. Source learning involves understanding who the source organization is and making attributions about the organization's intent. For example, is the purpose to *tell* valuable information, or to *sell* a product or service? Infusing messages with a strong sense of *social presence*, that is, the use of people's faces and names, will facilitate learning and relationship-building processes.

Media richness theory suggests that online communications tools vary considerably in the degree to which they contain nonverbal cues that facilitate understanding. Web sites containing explanatory text and graphics, for example, are comparatively *rich* media that provide contexts for content learning. By contrast, an FTP file containing only data would be considered *lean* because few cues for interpretation of meaning are provided.

Users are posited to develop cognitive schemas or knowledge structures for how online communications should work, based on experience. As a result, users develop expectations about content and rules for how to process information. Violations of these rules by content originators diminish relationship building.

Interaction

Interaction heightens levels of user involvement and can facilitate cognitive learning. *System interactivity* involves the ability of users to modify the format and content of online communications or messages. Hypertext and links on Web sites, for example, allow people to search for information, to play games, to take quizzes, to complete surveys, or to customize screen content. Similarly, customization features allow users to select and to focus on content of particular personal interest. *Verbal interactivity* involves the ability to create and send messages to others. Users thus are not merely receivers or manipulators of content but become full-fledged *producers* of content through e-mail, bulletin boards, and chats.

Impression Formation

Affective responses involve physical arousal or emotional reactions. Affect is distinct from cognition but combines with beliefs to create attitudes or predispositions, including attitudes toward online content or attitudes toward on online system (such as a Web site). Considerable research in the advertising field suggests that attitudes toward messages (sometimes referred to as attitude toward the ad) moderate the attitudes formed about the topics or objects that constitute the subject of the message (attitude toward the brand). Moreover, users form attitudes toward the online system, such as Web site, referred to as *attitude toward the system (site)*. These reactions or attitudes can moderate relationship formation. A positive experience, which also might be defined as *user satisfaction,* is critical to positive relationship building online.

Outcomes of Relationship Building Online

How can online relationship building be measured?

Knowledge

Successful relationship building online results in users becoming more knowledgeable about online information or the sponsoring organization. Evidence of learning effects can include rudimentary awareness and comprehension measures such as recognition and recall of content.

Attitudes

Relationship building effectiveness alternatively can be measured in terms of positive *attitudes toward the organization* or content sponsor. Attitudes can be measured as the degree to which a user *likes* the organization, thinks the organization's actions are *relevant* to them, *identifies* with the organization, feels an *affinity* (desire to belong), and *intends to act* in keeping with the organization's suggestions. Organizational *credibility* is another important outcome, i.e., the degree to which the organization is perceived as expert, trustworthy or independent. Finally, attitudes can be measured in terms of perceptions about the relationship that exists between the user and the organization. These assessment measures used in public relations in recent years include perceptions about the organization's *commitment to people, trustworthiness,* and willingness to share power (*control mutuality*). Relationship quality also can be measured as *satisfaction* and a sense of *communality* with the organization.

Communications Activities

Other important indicators of successful relationship building online include *repeated use* of online sources (such as returning to a Web site because the information or experience is valuable). Effects can be measured in terms of a user's willingness to communicate with the organization, whether through feedback online or offline. Finally, an important indicator of successful relationship building is the extent to which a person shares information or one's experiences with others—as an advocate for the organization, whether online or offline.

Routinized Behaviors

The ultimate measure of success in relationship building online is the degree to which users engage in routinized behaviors beneficial to the organization. In order words, do users' actions help the organization meet its mission or goals? Do people buy, invest, donate, work, vote, or avoid risky behaviors consistent with the organization's objectives for being online?

CRISIS COMMUNICATIONS AND ISSUES MANAGEMENT

Two special types of PR programs in which the Internet has become increasing important involve crisis communications and issues management. Both involve organizational responses to situations that originate outside the organization's control and focus on preserving organizational-public relationships.

A *crisis* is the uncertainty created following an extraordinary event. An event that could trigger a crisis be as inconsequential as a public official's embarrassing criticism of an organization. By contrast, catastrophic occurrences could include product tamperings, industrial accidents, sabotage, terrorism, or natural disasters. Crises erupt suddenly and often involve concerns about personal safety or loss. In most cases, organizations must respond quickly.

By contrast, an *issue* is a controversy or dispute that impugns an organization's reputation and often involves calls for changes in organizational or public policy. Unlike many crises, issues originate with people or organizations that make claims and seek changes (and possibly restitution) on behalf of a group that claims to be victims of a situation. Most issues are not life threatening per se, although an issue might involve potentially hazardous situations. Issues usually take a long time to incubate before they burst onto the public, media, or public policy agendas.

Crisis Communications

The Internet in Crisis Planning

Crises cannot be predicted or prevented entirely. Organizations try to reduce the prospect of problems by maintaining quality standards in their operations. In addition, organizations develop *contingency plans* for how they would respond to unavoidable crisis situations that might be particularly probable or damaging. These plans include a detailed communications component for which public relations units are responsible.

Online communications now play a central role in most crisis communications plans because Web sites and e-mail are well suited to disseminate timely and consistent information to a wide range of audiences. Indeed, these publics now *expect* organizations to disseminate crisis information online.

Online Crisis Response Strategies

Among the response actions now commonly found in online crisis communications plans are the following:

- Create a crisis Web page template in advance and immediately substitute it for the organization's regular home page (which should remain accessible through a link on the crisis page).
- Include both the organization's content and technical webmasters on the organization's crisis response team.
- Post online the same information released to the press at the location of a crisis event so that announcements are accessible to other media as well as the public at large.
- Create separate sections devoted to key audiences: employees, customers, investors and donors, neighbors and communities at large, and government officials.
- Develop, in advance, a list of useful links to online resources.
- Establish an employee bulletin board for staff to post information. Bulletin boards can be useful for locating missing employees, spotting rumors, and tracking incidents or problems. (This system should be on the public Internet, not on a restricted-access intranet, so that employees can communicate from any location.)
- Make arrangements in advance to access or create e-mail lists to communicate with company offices or officials, as well as employees, customers, investors, community leaders, government officials, or media personnel.
- Provide for increased processing capacity during peak periods of online queries, especially during the first 24 to 48 hours following a triggering event.

Major disasters involving the loss of life or facilities pose particular problems. In such instances, communities often lose electricity or telephone or cable service, effectively disabling Internet use. Prudent organizations contract with disaster recovery services firms to operate a *hot site,* or backup computer facility. Even when loss of service is temporary, restoration of Web and e-mail capabilities should be given priority.

In major disasters, an online *victim resource center* can be created online to provide information to survivors, as well as the immediate and extended families of all victims. Links can provide help from disaster relief agencies, medical facilities, insurance companies, employee benefit specialists, and posttrauma or grief counselors.

Issues Management

Whereas crisis communications usually involves responses after a triggering event has occurred, public relations can use the Internet to monitor developments and manage its responses throughout the typically long life cycle of an issue. (For discussion of how issues originate and how organizations respond, see Hallahan, 2001b).

Uses of the Internet by Activists

The Internet has been a boon to the creation of issues by activists and activist organizations. An *activist* is an individual who identifies a problem and then organizes others to seek its resolution. Remedies can include direct negotiations with an organization identified as the cause of a problem as well as indirect influence attempts such as boycotts or lobbying for legislative or regulatory changes. The Internet provides a pervasive, relatively low-cost forum for activist groups to access a worldwide audience. Indeed, almost any group can create a Web presence and obtain e-mail access for the cost of retaining a Web hosting service and registering a domain name–only a couple hundred dollars a year.

Activists often use discussion groups and chats to advance their causes. They also can purchase or otherwise obtain e-mail lists of people who might be sympathetic to their cause. E-mails and e-newsletters pertaining to political issues have a somewhat higher rate of user acceptance than commercial messages. Nonetheless, unsolicited messages from activists generally are still considered junk e-mail (spam) by unsympathetic recipients.

Activists use Web sites to post position papers, notices of events, and reports on recent accomplishments in order to reinstill commitment and solidarity among existing members and to recruit new supporters and volunteers. Effective activists encourage online feedback from supporters. Activists also use Web sites to raise money for their cause. More sophisticated activist groups have mastered how to combine database technologies with the Web for grassroots organizing activities. Activist Web sites, for example, can feature maps showing legislative districts so that supporters and sympathizers, acting as voter-constituents, can write letters to appropriate government officials. Sample letters are frequently provided that can be sent as-is or adapted by the sender. E-mail links allow messages to be sent directly to lawmakers' offices or to corporate officials who might be the targets of organizing efforts. In addition to e-mail addresses, smart organizers also provide postal mail addresses and telephone numbers.

Complaint, Attack, and Rogue Web Sites

Activists use a variety of imaginative online techniques to press their causes. Many activist Web sites adroitly use emotionally charged graphics, audio, and video to dramatize the plight of victims and the need for action. Some activists create *complaint sites* or *complaint domains* that are intended to allow people with a gripe against an organization to air their frustrations. Complaints are posted for public viewing or forwarded to the targeted organizations. Most notable among these are various *sucks.com* sites (http://www.sucks500.com).

Alternatively, disgruntled former customers, employees, or investors create *attack sites* intended to impugn the reputation of a targeted organization or political candidate. A special variation of an attack site is a *rogue site* in which the site sponsor obtains a domain name deliberately intended to be confused with a targeted organization's site. For example, http://www.ford.org was once a thinly veiled and deceptively titled site that attacked Ford Motor Company (http://www.ford.com).

The registration of names can be tracked by conducting *whois* searches on one of the major domain registration sites or domain name search engines (e.g., http://www.namedroppers.com).

Organization Responses Online

Besides monitoring the development of an issue online, organizations can use the Internet and other technologies to communicate their positions on issues or controversies. One technique involves direct participation in discussion groups and chats, where an identified representative of an organization articulates an organization's position. In some cases, the use of a moderated chat with a sysop might be appropriate, similar to a chat tour.

Other techniques include the use of broadcast and individual e-mails. Broadcast e-mail involves the creation of a tailored list of opinion leaders and or others who have expressed an interest in an issue. E-mail provides a simple but effective way to maintain contact and to provide updates on a situation, such the status of negotiations or pending legislation. Individual e-mails can provide a valuable way for organizations to respond. Nonetheless, it is important to avoid ingenuine, formulaic answers or perfunctory automated response messages that only reinforce perceptions about organizational indifference.

When confronted with an ongoing controversy or issue, many organizations use their corporate Web sites to articulate their position on issues, chronicle their record of performance, and solicit feedback or support. In addition to providing news releases and position statements in their online newsrooms, organizations address issues in separate, permanent sections of their Web sites devoted to specific topics (such as environmental protection) or to the broader subject of corporate responsibility. An increasing number of these pages invite and post public responses and comments.

MANAGING ONLINE PUBLIC RELATIONS

Public relations managers face a variety of concerns in maximizing the effectiveness of online public relations.

Content Control

One of the most difficult problems deals with the decentralized *control* of Internet and Web operations in many organizations. Responsibility for day-to-day operation of the system, including specification of hardware and software, usually rests with computer operations departments. This often results in the appointment of dual Webmasters with separate responsibilities for technical support (domiciled in computer operations) and content development (domiciled elsewhere, usually in a user department such as public relations).

Meanwhile, content decisions are often shared by public relations, marketing, human resources, knowledge management and other units that rely on the Internet. These units often have different objectives and priorities, and the resulting competition often leads to conflicts over content control. In particular, what messages will appear in the precious space of organization's front page?

Public relations units must be vigilant to develop a good working relationship with the technical staff. Organizations must establish procedures to ensure timely updates of sites, preferably direct updating by the public relations department itself. Similarly, PR units must be vigilant to ensure that public relations needs and objectives are not relegated to secondary positions by other organizational priorities.

Branding and Promotion

Effective Web sites need to reflect the organization's branding and corporate identity and must be promoted effectively. Many public relations units assume responsibility for maintaining Web content standards as part of their oversight of corporate identity. In particular, public relations must ensure that elements of the corporate identity system—logos, logotypes, colors, type fonts an so on—are used properly and consistently in all online communications.

Branding includes the selection of domain names when new sites are established and obtaining the rights to domain names through a *site registration service*. A good domain name should be short and easy to remember and to spell, and should communicate a key organizational message, if possible (e.g. www.flytoanywhere.com). To avoid problems of rogue sites or *cybersquatting,* organizations often purchase the variations of their domain names, including other suffixes (.org, .net, .info, .biz, etc.) and international extensions (.uk, .fr, .nz, .de, etc.). Organizations with frequently misspelled names also can purchase the misspelled names so errant inquirers are automatically forwarded to the correct Web address.

Most organizations seek to maximize awareness of Web sites by including their Web addresses on printed materials and print ads. One advantage of a very simple address is that the domain name can be used in broadcast promotions and told to people orally without confusion. Effective promotion also includes registration of domain names with the major search engines and effective use of *meta-tags* in all Web documents. These imbedded site descriptors include key words likely be used by prospective users and search engine agent software to locate matching information.

Message Quality

Beyond consistency of branding, PR units often must be concerned with the *quality* of materials that appear on organizational Web sites and standard e-mail correspondence, including materials generated by public relations and by other units. A systematic content review and control process is required to ensure consistency and accuracy of information, as well as conformity to organizational communication strategies.

Other quality issues include assuring that content is well written and follows standards for writing for the Internet. Web and e-mail copy ought to be brief, precise, and vivid because much Web content is scanned, not read completely. Writers must also recognize the importance of linking and layering content to address the varying levels of interests of users. Other quality concerns include making sure that content is accessible by the disabled.

Compliance with disability standards is required for Web sites supported by U.S. government funding.

Organizations also need to be sensitive to *multiculturalism.* Many organizations should consider creating sites in more than one language and to make language options readily available. Using a domain name with a national extension (such as .uk, .fr, .de, etc.) enhances acceptance by users in those and other nearby nations. Similarly, many organizational Web sites created by American or European designers are not well suited for use worldwide because these sites fail to take into account important cultural nuances related to the use of images, page organization/layout, and color. To communicate effectively and to build effective relationships, Web sites must be culturally appropriate.

Usability and User Satisfaction

In addition to content, PR units must be concerned with the functionality of organizational Web sites, discussion groups, and e-mail systems. *Usability research,* an extension of ergonomics, examines whether users can use technology to accomplish desired outcomes easily and with a minimum of errors.

Simple design is a key to usability. Generally, organizations err by employing cutting-edge techniques, such as animations, that distract users or reduce download time. Users should be able to locate information on a Web site intuitively. Pages should be organized simply, logically, and hierarchically. Hypertext links should allow users to find related information quickly and with a minimum of click-throughs. A search function should be included on all sites.

Among key usability concerns are ensuring that systems are compatible with the typical users' systems, speed of use and decision making, ease of navigation, and accuracy-of-use and success-of-search rates. Users must have a positive experience using online technology or communications efforts will fail. Dysfunctional systems generate negative opinions about a site and the sponsor and can lead to abandonment of site use together (Hallahan, 2001c).

Integration

The Internet represents one of many ways that organizations are deploying technology to manage relationships with important constituencies, particularly customers and employees.

Database technologies allow organizations to track transactions and other information. Database interfaces enable organizations to tailor direct mail, e-mail, and Web-based communications using profiling information.

As public relations practitioners examine future uses of Internet technologies, one of the greatest potentials is ability to provide *personalized* information that draws on organizational knowledge about the user's characteristics and past interaction with the organization. Marketers have recognized the value of this kind of relationship building, but the concept of personalization is only beginning to be applied in public relations.

Integration also includes the use of online communications in tandem with the full range of other communications tools available to an organization. For example, many organizations err by not prominently displaying e-mail, postal address and telephone numbers that can be used by users to reach them. Effective communication cannot be limited to a particular technology. Indeed, public relations communicators must use interactive media in concert with public media, controlled media, events, and one-on-one communication. More needs to be known about how these various tools can be deployed in tandem with online communications.

Finally, integration involves being prepared to adapt current Internet strategies as new devices are added to the personal computer in the online media mix. The advent of wireless telephones with Internet capabilities and other new personal data appliances (PDAs) will require that organizations adapt their online strategies and tactics. Indeed, the devices through which online public relations will be conducted in the second decade of the 21st century are only now in the development stage.

Security

Enlightened public relations principles encourage frank and open communications by organizations. Similarly, the trend toward open-systems organizations emphasizes making information freely and widely available. Organizations run risks by making too information prominently available, however. Public Web sites, for example, are one of the major tools used in competitive intelligence gathering by other organizations. Similarly, release of corporate and personal data can encourage or facilitate corporate espionage and computer hacking. This is why a large portion of online corporate information is protected within the firewalls of private Intranet systems. Organizations must constantly balance the need for user access and openness with the protection of the organization's digital assets (Hallahan, 2003b).

Legal and Regulatory Compliance

The Internet poses various new wrinkles on the plethora of legal and regulatory concerns that already confront public relations practitioners. Examples include avoidance of defamation in cyberspace and of the electronic invasion of customer and employee privacy. Other problems include digital infringement of intellectual property (copyright and trademarks) and the electronic misappropriation of people's images and words (in violation of the person's right of publicity).

Regulatory concerns include truth in advertising, compliance with rules related to political activities (contributions, lobbying, and representation of foreign clients), communications during labor disputes, and prompt and full disclosure of material financial information by publicly traded corporations. Government practitioners also must be concerned with the implications of new technology for making government records and meetings accessible to the public under freedom of information and "sunshine" laws, respectively.

Although the Internet's impact is far reaching, legislative, regulatory, and judicial mechanisms are slow to respond to changes. Online communicators must be sure their activities are in compliance with new regulations

Table 1 Six levels for assessing effectiveness of public relations Web sites.

1. Production
Qualitative assessments to judge compatibility with stated objectives; the adequacy of research or preparation, the appropriateness and completeness of content, and the production quality of messages (writing, design, visual appeal)

2. Usability
Technical configuration of the system, speed of use and decision making by users, ease of navigation, accuracy of use and success-of-search rates.

3. Exposure
Number of users impacted. System-generated data include the number of URLs or pages created, number of files accessed (hits), number of user sessions, click-through rates to secondary pages, number of unique visitors, and number of registered or potential users. Also can be assessed by examining user access logs or downloads based using domain name, domain type, or demographic data.

4. Awareness
Users' learning of content. Proxy measures of awareness might be imputed from tracking users, based on length of visits and downloads/transfers of content (presumably because of interest or value of content). Other measures include familiarity (recognition) with the existence of the site or of specific Web content, recall of use or specific content (user's ability to retrieve information from memory), comprehension (user's ability to explain meaning), and content retention (user's recall of information over time).

5. Attitudes
Changes in people's predispositions toward a topic or organization. As with awareness, length of visits and numbers of downloads or transfers provide a rough measure of positive attitudes because few people return or download materials they do not like. Other measures include the number of people who recognize the importance of topic as relevant to them, who express positive attitudes toward the site, and whose attitudes are altered (formed, reinforced or changed).

6. Actions
Effect on actual behaviors. Actions are relatively easy to measure in an e-commerce context because transactions are conducted online. Possible measures include assessing people's past actions (based on self-reports), or stated intent to take a specific actions, or actual observation of buying, investing, donating, working or voting behavior.

and must be in a position to counsel clients when others engage in activities that might be questionable. Many other legal questions and regulatory questions need to be addressed. These include the ownership and regulation of content in private e-mail and other systems, electronic records retention requirements, the legal status of electronic messages and signatures, and efforts to assess taxes on online activities. These legal issues are particularly difficult because cyberspace crosses legislative and regulatory jurisdictions.

Assessment and Measurement

Public relations must develop viable ways to measure the success of its Internet endeavors to ascertain the effectiveness of their work and understand the contribution of online activities to the organization. Unfortunately, particularly in an integrated campaign context, it is difficult to segregate measures and to isolate the impact of Internet and Web site activities from other organizational communications.

Figure 2 and the section on organizational–public relationships online examine measures of relationship building. Alternatively, six possible measures of the effectiveness of public relations Web sites as a specific types of online communications are illustrated in Table 1. Three of these measures directly relate to the site's content and use: qualitative assessments of the work product, usability of the system, and exposure generated. More sophisticated measures rely on surveys or otherwise obtaining measures of user awareness, attitudes and actions (see Traditional Public Relations earlier in the chapter).

Somewhat more difficult to compute, particularly when responses may occur offline instead of online is the *return on investment* (ROI) of conducting online public relations. Yet public relations must address the question of how online information and relationship-building efforts contribute to attainment of the organization's goals (financial, sales, production, etc.). Practitioners must be able to demonstrate the cost-effectiveness of such efforts and the consequences if these activities were not undertaken.

SOCIETAL ISSUES

The advent of online public relations poses a variety of professional and societal concerns for contemporary public relations practitioners.

User Concerns About the Internet— A Looming PR Issue

As the Internet becomes more pervasive and people become more dependent on online communications, people's expectations about and dependence on Web sites, e-mail, and other Internet technologies will rise. Organizations need to make sure that their Internet support is both dependable and responsible. Dependability involves the system being accessible to users by providing sufficient system capacity and software and system reliability.

Problems with downtime or reliability of information can create public relations crises in their own right.

More broadly, however, users expect organizations that communicate and conduct business online to act in an ethical and socially responsible manner. Successful organizations need to be vigilant in observing cultural values, norms, and mores. Among key values are respect for the individual, truthfulness, and freedom of choice. Organizations must be sensitive to issues such as privacy, accuracy, security, and the importance of giving people options for how they interact with an organization. To do otherwise is to misuse online technology. In 2001, the Arthur Page Society, an organization of top corporate PR practitioners, outlined four principles for ethical communications online that have been endorsed by about a dozen professional organizations. These tenets include presenting fact-based content, being an objective advocate, earning the public's trust, and educating the profession on best practices.

Quality of Organizational–Public Relationships in the Internet Society

Perhaps the bigger challenge is determining the right mix of tools for how organizations should interact with employees, customers, investors and donors, community leaders, government officials, and the media. Clearly, the Internet is a powerful and effective medium to communicate information, to respond to queries, to facilitate the exchange of ideas, and to build relationships.

Some organizations view the Web and related technologies as cost-efficient channels that can reduce the costs of transactions and control exchanges with customers and others. Yet as the Web and other Internet tools gain prominence, organizations must carefully consider the proper use of online communications. For example, pressing people to interact *only* with organizations online is not prudent, witnessed by the failure of banks to force customers to use automated teller machines. Similarly, the Internet makes it clear that organizations, at least from the audience perspective, can no longer *control* communications. Rather, organizations now *guide* communications involving key constituents, who are increasingly empowered to be come active participants in the conversation.

As suggested earlier (in the section titled Online Public Relations), the Internet is not a panacea for all communications or public relations needs of an organization. Organizations also run the risk of becoming slavishly committed to technology. Online communications must take its proper place in the communications activities of organizations. Misuse of the technology poses the risk of reducing, not enhancing, the quality of communication between an organization and its key publics.

Advocates point to the ability of the online communications to build community and to collapse boundaries created by time and space. Yet critics suggest that the Internet has isolated individuals by making people content to interact with others in cyberspace while forgoing important social contacts. Although the debate will undoubtedly continue, online communications are changing both structural and functional relationships in modern society. Public relations professionals and their client organizations must be learn more about the process and effects of online communications and be sensitive to these changes in order to nurture organization–public relationships and to take maximum advantage of online technology.

GLOSSARY

Attack site A Web site set up by a disgruntled activist, employee, customer, or investor that features negative content about an organization, political figure, or cause.

Branding Development of a unique identity that enables audiences to identify an organization, product, or Web site. A critical foundation for Web site promotion and traffic building.

Cyberactivism Use of online communications by organized groups to effect changes in public or organizational policies or practices.

Cyberscanning Process of environmental scanning applied to the Internet to identify potential concerns or controversies that might affect an organization.

Cybersquatting Purchase of the domain name registration of a Web site by an opportunist who later seeks to sell the rights to a party willing to pay for the use of the name.

Chat tour Use of public chat rooms to promote particular clients, such as book authors. Guests on tour answer questions from online participants.

Crisis communications The response to an unexpected triggering event that interrupts an organization's normal operations. Internet technology now plays a central role in crisis responses.

Fan site Web sites that build consumer interest in cultural or recreational activities, such as movies, television programs, pop music, or professional sports. Sites can be sponsored by the producers or started by enthusiasts themselves. Typical content includes information, photos, multimedia, discussion groups and chats, and links.

Interactivity Ability of users to make choices, select or customize content, play games, respond to surveys, or communicate actively with others while online in real time. Interactivity heightens levels of message processing. System interactivity involves interaction with the technology only; verbal interactivity involves producing and sending messages to others.

Issues management Responses by an organization to disputes or controversies that challenge the organization's principles or practices.

Media richness The degree to which a communication provides nonverbal cues for understanding a message. Media that provide a lot of context for message understanding are deemed rich media; those with few cues to facilitate understanding are considered lean media.

Online newsroom The section of an organization's Web site intended for journalists. Typically contains news and feature stories, background information, and downloadable multimedia.

Organizational-public relationship A routinized, sustained pattern of behavior among individuals related to their involvement with an organization. Public relations strives to build positive relationships between organizations and its publics.

Public relations Management staff function that facilitates communication between an organization and its key publics: employees, customers and consumers, investors and donors, community at large, and government.

Rogue site A Web site with a domain name similar to the URL (uniform resource locator) of another entity intended to distract unwitting users; sometimes a special form of attack site.

Usability The functionality of a Web site that enables users to find information quickly and accurately in a manner users find satisfying. Site usability is a necessary condition for messages to be effective.

Sponsored Web site A site that is subsidized by an organization to provide information about the organization, its products, services, candidates, or causes, in contrast to a *portal site,* which is operated as a gateway to the Internet, and a *news site* that suppliers users with information from various services. Sponsored Web sites do not charge for access; portal and media sites can charge for their services.

Webcasts Use of streaming video and audio to show a live presentation to online users in remote locations.

CROSS REFERENCES

See *Legal, Social and Ethical Issues; Online Dispute Resolution; Usability Testing: An Evaluation Process for Internet Communications; Webcasting.*

REFERENCES

Hachigian, D., & Hallahan, K. (2003). Perceptions of public relations Web sites by computer industry journalists. *Public Relations Review, 291,* 43–62.

Hallahan, K. (1994, Summer). Public relations and circumvention of the press. *Public Relations Quarterly, 38*(2), 17–19.

Hallahan, K. (2001a). Strategic media planning. Toward an integrated public relations media model. In R. L. Heath (Ed.), *Handbook of public relations* (pp. 461–470). Thousand Oaks, CA: Sage.

Hallahan, K. (2001b). The dynamics of issues activation and response. An issues processes model. *Journal of Public Relations Research, 13,* 27–59.

Hallahan, K. (2001c). Improving public relations Web sites through usability research. *Public Relations Review, 27,* 223–240.

Hallahan, K. (2003a, May). A model for assessing Web sites as tools in building organizational-public relationships. Paper presented to Public Relations Division, International Communication Association, San Diego.

Hallahan, K. (2003b, August). Protecting an organization's digital assets. Presentation to Public Relations Division, Association for Education in Journalism and Mass Communication, Kansas City.

Weitzel, D., & Hallahan, K. (2003, May). Adoption of an Intranet-based performance reporting system. Paper presented to Communication and Technology Division, International Communication Association, San Diego.

FURTHER READING

Web Sites

PR-education.org (http://PR-education.org) A portal to public relations education on the Web. Includes several online guides to books, articles, Web sites, and other resources on all phases of public relations. Includes a comprehensive bibliography on Public Relations and Technology.

Online Public Relations Course (http://www.online-pr.com) An introductory online public relations course developed by James L. Horton of Robert Marston Associations, New York.

The Institute for Public Relations (http://www.instituteforpr.com) A growing collection of research briefs and reports related to the Internet and management issues in public relations.

Pew Internet & American Life Project Online (http://www.pewinternet.org) Reports about Internet research, plus links other studies and Internet statistics.

Books

Holtz, S. (2002). *Public relations on the net* (2nd ed.). New York: AMACOM.

Horton, J. L. (2001). *Online public relations: A handbook for practitioners.* Westport, CT: Greenwood.

Marlow, E. (1996). *Electronic public relations.* Belmont, CA: Wadsworth.

Middleberg, D. (2001). *Winning PR in the wired world.* New York: McGraw-Hill.

O'Keefe, S. (2002). *Complete guide to internet publicity. Creating and launching successful online campaigns.* New York: Wiley.

Phillips, D. (2001). *Online public relations.* London: Kogen Poge.

Sherwin, G. R. & Avila, E. N. (1997). *Connecting online.* Grants Pass, OR: Oasis Press.

Shiva, V. A. (1997). *The Internet publicity guide.* New York: Allworth Press.

Witmer, D. F. (2000). *Spinning the Web.* New York: Addison-Wesley Longman.

Reports and Monographs

Middleberg, D., & Ross, S. S. (2002). *The Middleberg/Ross media survey. Change and its impact on communications.* Eighth annual national survey. New York: Middleberg+Associates. Retrieved September 2002 from http://www.middleberg.com

Pavlik, J. V., & Dozier, D. M. (1996). *Managing the information superhighway: A report on issues facing communications professionals.* Gainesville, FL: Institute for Public Relations Research & Education.

Wright, D. K. (1998). *Corporate communications policy concerning the Internet: A survey of the nation's senior-level corporate public relations officers.* Gainesville, FL: Institute for Public Relations.

Wright, D. K. (2001). *The magic communication machine. Examining the Internet's impact on public relations, journalism and the public.* Gainesville, FL: Institute for Public Relations.

Online Publishing

Randy M. Brooks, *Millikin University*

INTRODUCTION

Publishing is the process of preparing a text for public distribution. The traditional means of publishing is to design and print text on paper for physical distribution. Online publishing is the process of designing a text for public distribution and viewing on a computer screen. Whereas a printed newspaper arrives at your doorstep in the morning with a thud, the online newspaper arrives on your computer screen as a favorite link in your Web browser, a software program for displaying Web pages. In both cases, the publishing process includes (a) the design, (b) the public distribution, and (c) the reader's interaction with a published text. This chapter examines how shifting from paper to the computer screen changes all three elements of publishing.

Moving from paper pages to the computer screen significantly changes all elements of the publishing process—designing, distributing, and using texts. Online publishing changes the process of designing texts based on the limitations and new opportunities of the computer as a public display of information. Online publishing changes the process of distributing and acquiring texts to your computer screen—depending on standard formats, yet retaining the flexibility of being instantly reconfigured and displayed on a variety of electronic devices. Online publishing also changes the process of reading texts, transforming our habits of reading printed pages into multimedia habits of computer users. In fact, some would suggest that "computer user" is a more appropriate term than the concept of a reader, given the multiple opportunities and choices of control available to the reader of electronic texts. Online publishing builds on conventions and traditions of print media publishing, but it also creates new processes and new experiences of blending text, graphics, video, audio, and computer capabilities into multimedia experiences. In the book *Remediation*, Jay David Bolter and Richard Grusin (1999) characterized this shift as a remediation of print traditions into new approaches to online media. Online publishing remediates our habits as readers into new conventions and experiences only available through the electronic devices so common in our contemporary living and work space.

PUBLISHING ONLINE TEXTS

Designing texts for the computer screen instead of paper changes most aspects of the publishing process, including (a) our conception of an online text, (b) the acquisition and preparation of digital content, (c) organizing content for flexible online use and reuse, (d) launching and promoting the online publication through online channels of search and distribution, and (e) maintaining or updating new versions of the text—not in new editions but through ongoing dynamic updates.

What Is an Online Text?

A traditional publication is a string of words and graphics organized by arrangement on paper pages—usually bound with staples, thread, or glue in a protective cover. When we think of a book or a magazine, we conceive of these publications as fixed editions with static content and physical reality. When we pick up a traditional publication, we feel the texture of its bindings, we feel the weight of it in our hands, and we consider the size and scope of the text itself. Unlike a traditional paper text, an online text has no clear boundaries of size, scope, or weight based on the physical structure of bound pages. Instead of being conceived as a physical object that can be placed on a shelf, an online text is a collection of digital files organized into computer directories with invisible, embedded codes linking the files together for display on a screen of the user's computer. Instead of receiving the entire publication at once, users pull only the files and information they need or want at the moment of viewing.

An online publication is an organization of computer files, with flexible means of arrangement on a variety of computer screens and devices. Instead of being fixed and arranged by placement on paper pages, the words, graphics, audio and video information in online publications are all organized as layers of available computer screen displays. The online text is a performance, designed by the online publisher but controlled or reconfigured by the readers (computer users). The online text offers guided access to the computer files and links available as part of the experience of the electronic publication.

Online publications have a beginning screen (often referred to as a home page or index page), which serves multiple functions. The index page welcomes the user to the publication, often explains the primary purpose of the online text, establishes the overall tone of the publication, provides an overview of available content, and presents navigational links into various sections of the electronic text. Online publications that do not use common display software may need special instructions or "online guides" to help the beginning user start interacting with the publication. Once the user has entered into specific information screens, the online publication usually provides an ongoing means of navigation "back to the home page" or to other areas of the publication by means of additional links, "next" and "back" buttons, or a cluster of navigation links present on every screen.

So what is an online text? To the publisher, an online text is a collection of digital files, prepared and packaged for delivery through the Internet. To the designer (or writer), an online text is a collection of information carefully organized and linked for a variety of users. To the user, an online text is an opportunity to explore and experience available information and to answer questions by following navigational links through the computer screen displays.

Acquisition and Preparation of Digital Content

The acquisition of digital content has become extremely simple through recent technological innovation. Existing documents can be quickly scanned and converted into word-processing files. Graphics and photographs can be scanned and saved for online distribution. Music and video may be digitally recorded and edited for quick electronic distribution. Digital cameras and digital recording devices simplify the process of creating new content with the click of a button. Even children have learned how to create Web sites using various Web design editors or basic HTML (hypertext markup language) coded text files. Getting content is easy, whether it is coming from existing media or from original means. The difficult part is preparing digital content for effective distribution as an organized, well-designed online publication, and the sticky part is getting legal rights to that content for electronic distribution.

Preparing digital content for online publications requires knowledge of standard formats, file-naming conventions, and the strategy of "modularization" of your information into small, discrete chunks. Breaking your digital information into small chunks (separate files) allows you to manipulate each chunk easily, modifying only small parts of the larger online publication and thus maintaining the quality of each separate piece of information. Many organizations attempt to reduce workload and text development efforts by establishing a "single source" system of breaking information into separate chunks, which can be repurposed into a variety of online or printed texts. Because an online publication is a compilation of these small chunks of digital data, it is important to establish clear file-naming conventions to manage these pieces easily. The users may never see these discrete chunks and their file names, but the publisher and designer will need to be able to understand the pieces as they continually modify and repackage them into new compilations. Universal access formats such as rich text format (RTF) and HTML provide better access to the information from a variety of computer platforms and software. Graphics need to be prepared and compressed in standard formats for high-quality but small file size—GIF format for high-contrast graphics and JPEG for photographs and shaded graphic files. Small chunks of digital information can be individually compressed, making each file small and quick to distribute through Internet connections. So there are several reasons for breaking online publications into a large number of small modularized chunks of information.

Copyright law is currently under revision to address the ease of acquiring and redistribution of texts as online publications. The 1998 DMCA (Digital Millennium Copyright Act) attempts to address issues of intellectual property and new media. Through provisions concerning the circumvention of copyright protection systems, fair use in a digital environment, and online (Internet) service provider (ISP) liability, the DMCA addresses the technology of online distribution of content. Even in light of the new DMCA, existing copyright law assumes a more static, fixed publication or compilation of information in a dated public document. Under current law, copyright protection lasts for the life of the author plus 70 years. After that time period, works are considered to be in the "public domain" and may be freely distributed. Current technology makes it easy to acquire digital content: If it is pulled to the user's computer for display, it can easily be stored and modified by the user for future redistribution to others. For example, the current technology for playing music CDs on a computer also makes it easy to save the files in the compressed MP3 format for easy file sharing through the Internet. Illegal copies of new movies are recorded with digital cameras in theaters and distributed as digital video files before the industry releases DVD versions. Of course once the DVD versions are released, it is technically much easier to reproduce and distribute illegal copies. However, acquiring the necessary rights for electronic distribution is a responsibility of the online publisher. The copyright laws may be struggling with the concept of a text that is amorphous and dynamic in nature, but court cases and revisions to copyright law will eventually clarify the process of acquiring legal digital content.

Attempts to employ technological means of restricting the distribution of digital information usually result in a crippled means of distribution, which violates one of the primary benefits of online publication (users get to pull the desired or needed information to their computer screens). Every technological attempt to restrict the ease of acquiring digital information is quickly "hacked" by determined users and often has resulted in a crippled experience for the average user. Users want easy access to desired information, and the best solutions to legal rights need to be addressed through better laws and social processes, not through proprietary formats or technological restraints. To acquire electronic rights legally for online publishing, publishers follow the conventions of asking permission and purchasing these rights from the creators of the information, even the smallest chunks of information included in the online publication.

ORGANIZING CONTENT FOR FLEXIBLE ONLINE USE AND REUSE

With a large number of small modularized chunks in an online publication, users will be overwhelmed with too much information if they are simply given large lists of available files. Therefore, one of the essential elements of online design is to categorize information into related clusters and to provide an overview of the organizational strategy of available digital content. Users appreciate the added value of guides to content. The art and science of organizing content for online publications is referred to as information architecture, an analogy championed by Louis Rosenfeld and Peter Moreville (1998), authors of *Information Architecture for the Web: Designing Large-scale Web Sites*. Information architecture is the process of ordering and structuring of your basic content to shape the user's experience with the available information.

Rosenfield and Moreville (1998) described two approaches: top-down and bottom-up information architecture. Top-down information architecture is the process of developing an information architecture based on an understanding of the context of the content and the user needs. This involves determining the scope of the site and the creation of blueprints and mockups detailing the grouping and labeling of content areas. Bottom-up information architecture is the process of developing an information architecture based on an understanding of the content and the tools used to leverage that content (e.g., search, indexes). This involves the creation of building blocks, the databases to contain them, and the procedures for their maintenance.

Digital content may be organized in a multitude of organizational schemes (alphabetical, chronology, file type, content) depending on user needs, but organizational guides basically come in two options: (a) directories or (b) search engines.

A directory is an organized listing of content and topics created by the author or editor, attempting to predict common questions and content information users are expected to seek. Directories take the form of table of contents, indexes, site maps, or graphical previews of information. The advantage of a directory is that it provides a conceptual preview of the entire online publication (although the view may be limited by the number of categories that can be displayed on the screen at one time). Yahoo is an example of an online directory, which has been created and updated by humans categorizing the information available on the Web.

Computer search engines provide the other type of organizational guidance. A database or "indexed" record of available files is quickly analyzed by the search engine to seek matches to the terms the user enters into the search field. The user does not get a preview of available information, so the results of the search depend more upon the user's ability to anticipate useful key "subject words" for the search. Sometimes search engines offer a more restricted search through focused "content categories," and many search engines provide additional tutorials and tips for "advanced searching," which help the user find relevant information quickly. Some contemporary search engines organize results in a variety of ways, chosen by the user, but essentially this method of organizing the content of the online publication leaves the process entirely to the search strategies of the user. Search engines reduce the need to organize and structure the information available, allowing the user to reconfigure the related information quickly into a screen of links.

Many online publications provide multiple means of searching for desired digital content—a search engine for key-word searches *and* an index of contents or a site map. With both directories and search engines, once you have found the desired content, you can pull it to your display by clicking on links to the information.

Launching and Promoting the Online Publication

After the publisher has acquired, prepared, and organized the digital content for users, the next step in the process is to launch and promote the existence of the new online publication. Unlike traditional publications, which require prepublication copies to be sent to the major reviewers far in advance of the actual publication date, the online publication is launched quietly by becoming available to users through a unique Internet address, the publication's URL.

Once the online publication is launched, the publisher submits the URL to available search engines and directories. These search engines, such as Lycos, Altavista, and Google, automatically send a "spider" or "Web crawler" program to the Web site, which adds details about the pages and content from that publication to their database of available information online. In the case of directories, such as Yahoo, a person may check the existence of the online publication, consider the placement of the Web site in the appropriate directory category and send an indexing program to the URL for its database.

The web designer may include an invisible code in the HTML file called a "meta-tag," which provides indexing information for search engines and Web directories. The "key words" meta-tag includes a list of words users are likely to use to search for this particular publication or page. The "description" meta-tag is a descriptive sentence or short phrase about the online publication. Many of the search engines and directories will include the description from the "description" meta-tag in their listing of online publications or Web pages for the user.

In addition to the online means of promoting the existence of the online publication, some publishers advertise and promote their online publication through traditional means such as press releases, celebrations, e-mail notifications, print media advertisements, and inclusion of the URL on related products or public displays. The driving force behind online publications and the reason the web has become such a revolutionary approach to publishing, however, is the fact that users pull information when it is needed and desired. If the publisher pushes unwanted advertising and information at users, this violates the principle of user control of the online reading experience. Users have vigorously resisted previous attempts to build "push" technologies and automatic delivery of online publications. The best promotion of online publications is simply to be ready when the user comes to it and to have your online publication show up in the search engines and directories as soon as the user is looking for it.

Maintaining or Updating the Online Text

Once users come to your online publication, they will not stay long unless there is rich, up-to-date content. If you launch a new online publication, but the content and design just sit there unchanging, users will assume it provides only old information, and there will be no reason to return to your site. Users will not return to your publication unless you maintain, update, and add to the content of the online text on a regular basis. An online publication is valuable because it offers rich layers of information that are continually being updated. The user can always get static, old information from books. From an online publication, the user expects dynamic, constantly improving information.

Of course not all information goes stale instantly, so the publisher can design the online publication with a mixture of static and dynamic information. By targeting certain sections for scheduled, routine updates, users will quickly find the new information but also appreciate the depth of some of the static information, which retains its value even though it is not being reconfigured as often. For example, the Web version of the *New York Times* provides free access to daily news stories, but archives are only available for a fee. Academic online journals, such as *Kairos,* provide current and archived hypertexts as a means of distinguishing between dynamic and static areas of the publication. The dynamic nature of an online publication assumes an ongoing "Webmaster" or editor is watching over the publication and continually making adjustments and improvements to the compilation of information available. An online publication may continue for several years with this dynamic growth of content, but it requires periodic long-term revisions or "makeover designs" to keep it fresh after a few years. These makeover designs are rarely considered "new editions" as they would be in traditional books, and the old versions rarely remain available. The fact that earlier versions disappear as they are made over into new versions of the online publication is disturbing to some people, especially researchers who like the stability of traditional books that retain a permanent record of editions.

Online publishers gather digital content and organize it for maximum user control of the reading experience of the electronic text. The online publisher prepares this information into small chunks, with each saved as a separate file with links and guides to the organization of the content. These online publications are valued for rich amounts of content, carefully organized so that the reader can quickly find particular data or answer specific questions by navigating to the needed content. Through the use of directories and search engines, the publisher gives the user access to the elements of the online publication. The online publisher promotes the electronic publication through existing computer search engines and Internet directories, adding their online texts to the millions of other documents on the Internet. One of the primary values or advantages of online texts is the ease of updates and edits. Current, up-to-date information is expected from quality online publications, so instead of publishing in editions, electronic texts are expected to be dynamic—to be maintained, updated, and changed as an ongoing, evolving publication.

DISTRIBUTING TEXTS ONLINE

Distributing texts online depends on conventional Internet protocols and standard formats for file exchange and display. These file exchange protocols are essentially a means for two computers on the Internet to connect, "shake hands" to establish their identities, then share files based on the rules and restrictions of the server settings.

File Exchange Protocols

Authors, editors, and publishers create digital information from a variety of local computers and then send this information to a "server" computer using a file transfer protocol (FTP) program. Most Web design software includes the FTP protocol for sending files to a remote server, and there are many programs dedicated to file transfer. Sometimes this process of transferring files to a server computer is referred to as "publishing" the files. Because FTP programs do not include content display functions, they are a fast means of simply moving files from one computer to another computer.

A user finds the location of the desired online publication with an Internet browser, such as Netscape Navigator or Internet Explorer (software designed for both transfer and display functions). The user connects to the server computer using hypertextual transfer protocol (HTTP) to download a copy of the desired information to his or her own computer. The "server" offers access to available files, and the "client" computer pulls desired files to the user. These files are stored in a temporary cache for viewing on the local computer, or they may be downloaded and saved in a more permanent file format (such as PDF [portable document format], RTF, or HTML) for immediate and future use. Large computer files that will take longer to download are often compressed and saved as "zipped" or "stuffed" files for quicker downloads. These documents have to be "unzipped" or "unstuffed" to be viewed on the user's computer.

The Web is built primarily with HTML pages, so this is now the most common format for online publications.

HTML is a convention for embedding invisible tags into a simple text file so that browsers may display the information according to the conventions of those tags. For example, a mark up tag for bold typeface will be a pair of tags for starting and ending the bold text—for example <B>bold text</B>. The characters between the tags would be displayed as bold by the browser software, such as Netscape Navigator or Internet Explorer. Every HTML document must have certain tags, such as a beginning <html> and ending </html> tags to let the browser know that this document is an HTML document. Here is an example of the minimum HTML tags necessary for any Web page:

```
<html>
<head>
<title>Untitled Document</title>
</head>
<body>
</body>
</html>
```

Of course, additional tags can add display details such as tables, alignment, location of graphics, hypertext links to other files, and additional information in the HTML file. The user sees only the display of words and graphics on the computer screen, but the source code is always available for viewing if the user wishes to see the underlying HTML tags.

HTML pages are often enhanced with JavaScript programming or other types of code that may be activated as users come to certain web pages. These scripts extend interaction with the user and provide special effects not available through standard markup tags. For example, a JavaScript might add a rollover effect to a graphic so that when the user's cursor moves over the graphic it is swapped with another graphic.

Standard Digital Content Formats

Other formats for distributing online information include database publishing, which employs Active Server Pages (ASP) technology to generate Web pages on the fly. Extensible markup language (XML) is a growing format language that allows content to be reformatted or reshaped quickly for a variety of output screens such as cell-phone displays, automobile computer screens, and computer monitors. Another format for designing Web sites as animated, scripted movies using Macromedia Flash has been created with Shockwave Format (SWF) movie files. These movies can be scripted for user control but are placed on a simple HTML page for compatibility with standard browsers.

The most common formats for online content have been developed for quick downloads through Web browsers. Graphics are saved for the standard screen resolution of 72 pixels per inch. This reduces the file size considerably yet maintains a high-quality graphic if properly saved in the appropriate format. There are two main formats of graphics on the Web: GIF and JPEG formats. GIF graphics are high-contrast graphics compressed by saving a small index table of colors needed in that particular graphic. JPEG graphics are usually photographs or graphics with a wide range of shades and colors. A color table does not contain enough colors to provide a quality display of these graphics, so they use the local computer's color display system. JPEGs may be compressed with a sliding scale of compression that allows the artist to save the graphic at the most highly compressed level that is still acceptable for the required visual quality. A more recent graphic format is PNG, a format similar to the GIF format that allows for transparency and features an index table of colors for each graphic file. Animations, video, audio, and other types of digital content each have been developed with standard formats for exchange and activation on the user's computer. Many of these elements require a special "plug-in" for the browser to extend the capabilities of the browser, but such software extensions tend to become standard features of browsers over time. The goal of each format is to provide high-quality files with maximum compression for quick downloading to the user's computer.

Electronic Books

Although online publications often blend into Web sites without clear boundaries of a distinct text, traditional publishers have attempted to develop the concept of electronic books (often referred to as e-books) that represent a more limited notion of a complete text (with title pages, table of contents, chapters, page numbers, and illustrations) based on conventions of printed books. These e-books are issued with a publication date and clearly designated author and are not usually updated with dynamic information in the same way expected of Web-based online publications. Often these are distributed for a fee as one complete downloaded file for viewing on an electronic device. An e-book is essentially a computer file tagged for display in a variety of electronic devices. Although electronic books are often marketed, sold, and distributed online, they may be displayed or read on an electronic device not connected to the Internet.

E-books have been produced in several formats for display on a variety of electronic devices.

Formats of Electronic Books

E-books come in several file formats, including HTML, PDF, RTF, Palm operating system-compatible, Windows CE-compatible, or other proprietary file formats for electronic reading devices. Windows CE is an embedded operating system for consumer electronic devices. E-book reading software is often available as a free download. Although e-books are not always available online, they are marketed as being designed for better on-screen reading quality (with improved anti-aliased typefaces), better management of your collection of e-books, special book marking and annotating capabilities, and better searching within the e-book file.

ASCII Format Books

Although there are several libraries of text-only publications, an example of this approach is the Gutenberg project. Since 1971, Project Gutenberg has been saving public domain documents in ASCII text format. This basic text-only format preserves only the letters, numbers and punctuation. All bolding, italics, and other formatting are lost. Any word processing software can read ASCII files, and the Project Gutenberg Reader Web site (http://pdreader.org/) allows people to read Gutenberg titles and other public domain titles online. With simple ASCII format, the user can easily search an entire library or online publication, without a proprietary software program. The library can be searched with an ordinary search command within the browser or using any word-processing program. The Project Gutenberg Library is organized into three large categories of online titles, described in the project overview as (a) light literature, such as *Alice in Wonderland, Through the Looking-Glass, Peter Pan,* and *Aesop's Fables*; (b) heavy literature, such as the Bible or other religious documents, Shakespeare, *Moby Dick,* and *Paradise Lost*; (c) references, such as *Roget's Thesaurus*, almanacs, and a set of encyclopedias and dictionaries. Like open-source software development, the Gutenberg project is a collaborative effort with thousands of contributors, so literary scholars have decried the production errors and varieties of editions found in many of the Gutenberg texts.

Rich Text Format (RTF) Books

RTF preserves some of the text formatting such as bold and italics, tabs, indents, and typeface choices. These simple texts are read in word-processing software that displays the text formatting, and most browsers can display them, although they will lack the integration of text and graphics common in HTML documents.

CD-ROM Books

CD-ROM books are published in a wide variety of formats, including proprietary software for specific computer systems. As a static collection of computer files, CD-ROM books tend to blend multimedia elements such as video, music, and text into a unified electronic interaction. Despite the popularity of certain CD-ROM books for children, such as the Living Books series by Random House/Broderbund, this format has not gained widespread acceptance because CD-ROM publications tend to alter the user's computer automatically. To use a CD-ROM book, the user often has to "install" electronic resources such as fonts or viewer software so that the multimedia elements will function. If you install an old CD-ROM book's resources on your new computer, you may accidentally downgrade your computer to old versions of video or audio output. As a static form of publishing, the CD-ROM book is not easily adjusted to the variety of computer systems without such installations. Although CD-ROM books were promising before the growth of the Web, a quick glance at one of the leading CD-ROM publishers, such as Voyager, reveals that most CD-ROMs were published in the early to mid-1990s and few new titles are being published or developed in this format.

Adobe Acrobat PDF Format Books

Acrobat's PDF displays e-text exactly the way it appears on printed pages, maintaining typeface choices, graphics, and arrangement of elements on a page. Many publishers like PDF because they can carefully control the way text looks. The user cannot resize the document nor easily display such files on a variety of electronic devices. The disadvantage is that a PDF document is designed for a particular-sized screen and doesn't flow onto other sizes of screens.

Like HTML, Acrobat PDF documents come in a simple format that allows easy access to the digital content. Read-only versions of files can be created, but like Web pages the content is readily available to the determined user. Customers download a PDF text and display it with free software called Acrobat Reader. Another variation of the PDF format, Modified PDF, can be displayed by a free Adobe Acrobat eBook Reader. This software is designed to make it easy to purchase and download Modified PDF electronic books to desktops, laptops, or notebook computers and eventually on Palm Pilots. This reader software displays e-books with the pictures, graphics, and rich fonts you've come to expect from printed books.

HTML Format Books

HTML flows text on computer screens with an easy integration of visual and verbal elements. Using the common navigation system of the browser software, it is easy for users to navigate and orient to the information in these texts, especially if they have been designed with user control in mind. This "flow" gives flexibility for text to look tolerably well on any size computer screen based on the user's preferred settings of text size, and display resolution. These e-texts can be read online or downloaded and read offline with any Web browser. HTML publications are the standard format for most free virtual libraries and for many publishers selling e-texts for downloading. Large collections of online publications are available on the Web in HTML format (or a combination of HTML, RTF, and ASCII formats).

Virtual Libraries

Virtual libraries offer free books, including literature, humanities, and social sciences texts, for reading online in HTML format. Virtual libraries frequently provide key-word searching of the full-text of entire collections. Some providers of e-text offerings through the Web (as of June 18, 2002) include the following:

- Bartleby.com (http://www.bartleby.com/)
- Bibliomania (http://www.bibliomania.com/)
- Black Mask (http://www.blackmask.com/)
- Classic Bookshelf (http://www.classicbookshelf.com/)
- Electronic Text Center (http://etext.lib.virginia.edu/)
- Internet Classics Archive (http://classics.mit.edu/)
- Internet Public Library (http://www.ipl.org/div/books/)
- National Academy Press (http://www.nap.edu/)
- On-line Books Page (http://digital.library.upenn.edu/books/)
- WWW Virtual Library (http://www.vlib.org/)

Some publishers sell new e-texts on the Web for downloading and reading offline. Because the entire e-text is saved as a single file, downloading is quick. Some virtual libraries are building collections of online books available through the Web through subscription or library association memberships. For example, *netLibrary* licenses e-books to libraries and library consortia. Its e-books are current academic books from the top academic publishers in the country. The e-books are housed on netLibrary's Web servers. Libraries and schools purchase access to entire books. Users can key-word search across the entire collection purchased by the library. Web access allows library patrons to see purchased books at home or on dispersed campuses. NetLibrary has numerous free e-book versions taken from Project Gutenberg and other public domain sources (http://www.netlibrary.com/).

Other Web collections provide samples or portions of books for viewing, but payment is required for the entire publication; for example, *ebrary* provides an online business and economic book collection through the Web. Users can search the entire collection and browse its contents for free. Payment is required for copying or printing any amount of text (http://www.ebrary.com/).

The ibooks Web collection provides e-book versions of books on computers, electrical engineering, and business. Users can key-word search the entire collection of books or continue to access those books electronically purchased by the user. The full text is only shown to those who have purchased a book (http://www.ibooksinc.com/).

Books24X7 provides access to e-book versions of books about computers and related areas. Users can browse texts or key-word search full text. Only subscribers or those purchasing books can see the text (http://www.books24x7.com/).

Questia is a term-paper writing tool with a 250,000-volume library of most used academic books attached. *Questia* selects the books most cited by other academic books and most used in undergraduate curriculums. Students can use key word searches of the full-text of the entire library for relevant passages and quickly cut and paste the text and automatically create citations and bibliographies into their term papers (http://www.questia.com/).

Variety of Electronic Display Devices

E-books may be read on desktop and laptop computers, handheld computers, personal digital assistants (PDAs), and specialized e-book reading devices, such as the REB1100 sold by RCA. As specialized electronic devices for reading electronic text offline, e-book readers feature portability, backlighting, and screen display customization by the user. These electronic devices are usually battery powered, with limited battery life given the demands of quality displays. Although the Web contains a vastly larger collection of text and online publications, these e-books and their special reading devices are marketed as better than the Web. Handheld readers are promoted by emphasizing design features such as providing more readable texts, including turning pages with a button, portrait orientation of text, high-resolution display of text and graphics, and lightweight portability. Online texts displayed with the Web tend to feature scrolling navigation, more often follow the landscape navigation of computer screens, may be limited by the computer's display resolution, and are less portable than handheld readers.

Dedicated Electronic Reading Devices

E-book reading devices are designed specifically for reading books and emulate the print book experience. Screens are backlit and usually about the size of a large paperback book. The devices use page-turning systems rather than scrolling and use touch screen technology for searching, book marking, highlighting, annotating, and font size changes. You can store approximately 10 to 20 e-books on most handheld readers, more if you upgrade the memory. Dedicated e-book reading devices need specific file formats. E-book readers have numerous parts (e.g., reader, stylus, screen cloth, AC adapter, and zippered container), so people buy additional cases to contain all the parts. Efforts to establish standards for e-book file formats are often modifications of HTML or XML and PDF with a variety of attempts to prevent illegal sharing of copies (but allowing for conversion between various devices).

Rocket E-Book. Although technically not the first e-book reading device, the discontinued Rocket eBook is responsible for creating the modern era of the e-book. It was the first device to gain widespread recognition and a fair amount of sales. A simple device with a bright, small display enticed readers to try taking e-books offline, allowing them to use the system on the train, in cars, and at the beach. The Rocket eBook's manufacturer, NuvoMedia, was purchased by Gemstar-TV Guide in January 2000. Late that year, Gemstar and manufacturing partner Thomson Multimedia released the Rocket eBook's successor, the RCA eBook REB1100.

RCA REB 1100 and 1200. The REB1100 and REB 1200 are handheld devices designed specifically for reading books and emulate the experience of reading a printed book. Screens are backlit and usually about the size of a paperback book. The devices use page-turning systems rather than scrolling, and use touch-screen technology for searches, bookmarks, and annotations. REBs read e-books in their specific format and include a simplified modem to download e-books from online eBook stores.

Hiebook. A third-generation e-book device, the hiebook is a multifunction device that features an e-book reader, MP3 player, PDA functions, audio recording and games. Made in Korea, the hiebook is similar to the Franklin eBookMan and combines an e-book reader (OeB-compliant) with personal information management applications, an MP3 player, and an open architecture for third-party development.

Palm Computing Devices (PDAs). Using software to become e-book readers, PDAs can display e-books in addition to their other functions (address book, scheduler, calculator). Some users criticize PDA screens for being too small to display much e-text, making the reading experience frustrating or annoying. Others argue that PDAs are the future of e-book readers because people want to carry one device that does many things rather than carry many devices, each specializing in one task. Palm Pilots use software to turn them into e-book readers. AportisDoc and TealDoc are the best known e-book reader software for PDAs. Adobe is also creating software for displaying e-books in PDF format on Palm Pilots. Software, called conversion tools, is available to convert many files for display on Palm Pilots.

eBookman From Franklin (PDA/eBook Reader). Franklin sells eBooks with content built in for various reference areas such as medicine or law with new content areas available as snap-in, domino-sized hardware cards. Or users can purchase a BOOKMAN Writer, which allows them to download titles and create their own portable Book Cards. The eBookMan can read aloud to the user with software called Audible.com. Audible has thousands of audio books available for download.

Audio Books. Audio books are available in a variety of formats including digital formats for handheld devices and laptop computers. For example, Audible Otis is a 64-MB digital audio player that plays audible content in several formats, including MP3 and Windows Media audio files. Audible ready players include all PocketPC makes and models, Handspring Visor with Audible Advisor Springboard Module, Audible Otis, Rio500, Rio600, and Rio800 MP3 players, Iomega HipZip, Digisette Duo-Aria MP3 player, and the Franklin eBookman PDA.

Standards for digital audio books have been created for the Library of Congress program Talking Books, which creates editions for blind, visually impaired, physically handicapped, or otherwise print-disabled readers. Digital audio books provide users the ability to move to any part of the audio book including jumping to footnotes, the ability to highlight, bookmark, and annotate the original audio text and to synchronize the audio and text display so readers can read portions and listen to portions simultaneously.

The vast majority of online publications are distributed as HTML pages employing the navigation and display advantages of a Web browser on desktop or laptop computers. Users have access to a vast library of interrelated information; they can quickly copy and recreate online publications for their own use with software simultaneously available on the display computer. They can bookmark Web pages for future use and create annotations in a favorite word processor. The Web-based online publication integrates graphics, video, and sound elements into a coherent unified experience. The user can adjust the display of the online publication on his or her own computer screen, setting the size of typeface, sometimes the choice of typeface, the size of the display window, the resolution of the monitor, and other elements of user control (such as the size of the browser cache and e-mail settings).

Handheld electronic devices such as PDAs are simply too small and limiting for a majority of online publications. They are fine as electronic Post-it notes, providing access to extremely small chunks of text such as addresses, phone numbers, brief e-mails, and to-do lists, but when the user attempts to read a larger electronic file such as a newspaper, a report, a long e-mail, or an e-book, the small display limits the user to two or three sentences at a time. Graphics become so small that they are seldom more than icons representing content more than delivering visual content. The quality of the small display can be especially frustrating in sunlight or when the battery is running low. High-quality, anti-aliased type doesn't matter when the screen goes dim. Clearly these devices are designed for brief bursts of looking for small, discrete bits of information.

Specially designed electronic readers are better than PDAs because they are designed for displaying text, but they are still limited by the size of the screen, the small amount of memory in the device available for storing electronic texts, and battery life. They are the electronic equivalent of cheap paperbacks, which are viewed in the book design industry as low-cost, cheaply designed disposable editions. In other words, they do not include significant amounts of graphics, the information is crammed into the small display space available, and they are best suited for long, continuous reads such as novels, biographies, and some nonfiction works. Unlike Web-based online publications, the screen cannot be resized, and all hypertext links in the electronic text are limited to elements of the e-book currently stored in the user's reading device. In addition, users do not have simultaneous access to other software on their computer while interacting with the electronic text. Like paperbacks, the idea of e-book readers is to download the computer file, read it anywhere with the portability of batteries, then dispose of the file (or at least get it off the reader to make room for new electronic texts). The file could be stored on a computer for future reloading into the reader, but it is not expected to have the dynamic rich content of truly online publications.

Perhaps the best portable version of online texts is the audio book, which does make sense in a small handheld electronic device, because it does not depend on a large display. Adding more user controls for quickly navigating within the electronic audio file is a significant step beyond a simple cassette or CD recording that only allows the user to play, fast-forward, or rewind the entire text. These audio books seldom have the interactivity and dynamic content we have come to expect in Web-based online publications.

DESIGNING ONLINE TEXTS FOR EFFECTIVE USABILITY

Online publishing changes the process of reading texts. Our habits of reading are primarily derived from experience with printed pages, but these habits are changed as users move to reading texts on the computer screen. New habits and conventions are emerging from the expectations and experiences of computer users, who are used to the multiple opportunities and choices of control available when interacting with an electronic text. Users of online publications bring existing strategies for reading to the electronic text but do not expect the online publication to behave the same as a conventional book. Therefore, creators of online publications need to understand how to design their electronic text for effective, high-quality usability, exploiting existing reading strategies while simultaneously providing the new interactive capabilities of the new media environment.

Remediation of Print Publishing Conventions

Traditional printed publications include conventional elements such as folios (page numbering), headings, chapters, tables of contents, prefaces, and indexes that help orient and guide the reader through the text. Print media design conventions employ "white space" on the page to help frame information, suggest relative importance, and to help the reader understand the organization of information on the page. Moving from the printed page to the computer screen alters all of these design conventions.

The transition from the printed page to the computer screen is not a complete replacement of conventions, but rather a remediation of existing print media conventions in the new publishing space of the computer screen. For example, printed publications usually include page numbers, but screens rarely have numbers, preferring a sense of scrolling or displaying multiple "windows" of information for access to additional content. Both pages in a book and "Web pages" in an online publication need headers and titles, however, to let the reader know where they are at any moment of the reading process. Even the concept of "Web pages" is a remediation of the idea of pages in a book, although an electronic window also has capabilities of scrolling and displaying moving or shifting content on the screen and usually includes ongoing links to other screens of information.

The computer screen is not a page, and layered windows of digital information are not text columns. In general, printed pages are more vertical and computer screens more horizontal in design space. This "portrait" orientation of printed pages reinforces reading habits, such as the strategy of "top-down" scanning for significance and, for Western readers, the process of reading across columns of text from the left to right side of lines of text. The relative narrowness of printed pages has reinforced conventions of column length, so readers are most comfortable reading relatively short columns of text. Graphics are carefully integrated with text on the printed page because of their power to overtake the eye, becoming a focal point of attention and emphasis. Often graphics go beyond the boundaries of columns or across the gutter between pages in a print publication.

The design space of the computer screen forces us to remediate our publication conventions into a horizontal orientation. Graphics and text are placed side by side, and readers look across the computer screen as well as from the top down for quick overviews of content. Columns of type stretching across the entire horizontal space are too wide for comfortable reading, so online publication designers usually break these columns into smaller chunks, or limit the width of columns into a set-width area of the page. The top of the computer screen continues to function as an important orientation space containing Web page titles and organizational identity logos.

The computer screen is inherently more cluttered and busy with competing elements of interaction—the web browser's navigation bars, the computer system menus, graphics and text on the Web page, and other elements all compete for the user's attention. Whereas a print publication has a fixed position (branding and headlines on the cover, table of contents near the front, index in back, ads dispersed throughout), online publications require that each page includes all elements, as users may follow "deep-linking" and get to a page from a search engine. So every Web page usually includes the logo, links, ads, and content. Links to related information and other Web sites attempt to draw the reader away from the information on the present page, so one of the main remediations from printed pages is to simplify the design space. Instead of white space, the online publication designer needs to manage available screen "real estate" carefully, leaving some areas of the screen empty or open for emphasis and communication of essential content. The prime real estate of the horizontal screen for orientation includes the top of the screen and the left-hand side (which is quickly becoming an area associated with navigation within the online publication). The prime area for content is the upper middle portion of the screen. Advertisements or extra information is, therefore, usually placed to the right side or bottom of online publications, where they are visible but less distracting to the user's primary experience with the text.

Graphics, video and animation have a powerful draw for the short-term attention, and therefore these must be used carefully to attract but not to disrupt the user's experience with the content of the online publication. Every graphic has implied motion, so that inherent motion can be designed to lead the eye to important content or to appropriate starting points for more detailed verbal content. In the same way, animations and video have a powerful attraction and because of their inherent movement can help lead to significant content or possible navigation to desired content. If these run on nonstop loops and remain visible on the user's screen, animations and videos become annoying distractions instead of helpful guides to meaning.

Common tools for designing online publications have brought online publishing capabilities to a large number of people who have never been involved in publishing traditional print media documents. As readers, these novice designers know traditional publishing conventions only as subliminal reading habits, not as explicit strategies for designing publications for other readers. Novice online publishers attempt to use strategies for writing that they learned in schools—following manuscript conventions of double-spaced lines, indents before paragraphs, and two spaces between sentences. They employ school-based manuscript conventions of all caps and underlining for titles and headings while enjoying the wild choices of typeface now available on the computer. These beginners have no concept of an ideal column width for reading. The online publications they create employ the concept of 1-inch margins to the edges of the available space. These manuscript conventions open up space in the text for editors or teachers to add comments, annotations and notes to the typesetter. The novice designers do not realize that online publications are designed not for editors and teachers, but for readers.

Instead of remediating their manuscript design strategies for the Web, beginning online designers merely transfer their text production habits. Web design software will not allow them to follow these manuscript conventions easily—for example, not allowing more than one space between characters and automatically adding space between paragraphs instead of employing indentation. Although these limits frustrate novice online designers because it forces them to give up certain habits, they still fail to remediate other elements such as column length, so their resulting pages are difficult to read and enjoy online. They do not understand that the goal of online publication design is to prepare an electronic text for users.

Usability and User-Centered Design

An effective online publication is designed for its intended users—their needs, wishes, and expectations. Designing an online publication for users means that the design presents an established pattern of elements and options, especially for orientation and navigation. This online publication style consists of several elements of Web design, including a common background color or graphic, consistent use of typeface and headings, consistent placement of navigation elements, and options for interaction on different levels by different types of users. A user-centered design approach anticipates the range of user personalities and the variety of purposes they will bring to an online publication. Instead of expecting all users to conform to a set pattern of interaction (more common to print media), the online publication lets users interact with the publication according to their own preferences.

There are several ways to anticipate and design for a variety of users. The development of complex online publications may call for user testing to observe a variety of users attempting to interact with the publication. User testing considers personality differences of users and their approaches to interacting with the online publication. Some users prefer a playful trial-and-error approach to interacting with a new online publication. Others prefer detailed guidance, Web site maps, prompted actions, and feedback to their choices. Some may simply want to find a specific bit of information as quickly as possible, with all other elements of the publication being merely a distraction to that goal.

User-centered design also considers the variety of purposes users bring to the online publication. Is the user seeking quick information, an opportunity for thoughtful contemplation, an experience of goofy play, or a slightly interested learning opportunity? Each purpose may require a slightly different means of orienting and navigating in the online publication. How does an effective online publication design meet these usability needs? Instead of limiting users to one means of entering and interacting with the text, most online publications provide multiple means of orientation, navigation, and interaction with the text. Three online design strategies help provide these multiple means of interaction: (a) clear orientation labels of content, (b) consistent navigation prompts, and (c) careful layering of information from simple overviews to more detailed content.

User testing attempts to observe a variety of users' reactions and responses to a prototype design of the online publication, resulting in rapidly revised versions of the online publication (which can be further tested through repeated user testing). Often user testing is conducted with talk-aloud protocols of interaction. Users are asked to talk out loud about what they are doing as they make choices and use the online publication, sharing their thinking (or at least some of it) while viewing and using the text. Sometimes users are asked to respond to the opening screens of main sections. They describe anticipated content and share their expectations of the publication. This initial orientation process is important, especially given the invisible nature of the rest of the online publication. After the user has expressed initial responses and expectations, the user may be asked to perform a specific task with the online publication, such as looking up a particular fact or area of information. This prompted task helps test the effectiveness of the organization and navigation of the online publication. The user testing session usually closes with impressions about the publication's overall orientation, ease of navigation, and the aesthetic value of its design.

Rhetorical Purpose and Continuity of Online Publication Style

While designed for a variety of users, each online publication declares a primary rhetorical purpose through the publication style choices presented. The opening page sets the atmosphere or tone of the whole electronic document, announcing or suggesting the purpose and appropriate forms of interaction available. If the opening page contains flashy graphics, lots of animation, music, and gamelike navigation, it clearly announces that this is a space for playful relaxation. On the other hand, if the opening screen includes a mission statement, a list of questions, or an extensive table of contents, it clearly asserts that this is a space for seeking information. Most online publications can be placed on a continuum of competing rhetorical purposes, ranging from entertainment to information values. On the entertainment end of the continuum online publications become interactive games, with role-playing and a great deal of music, animation, and video included for the online experience of the publication. On the information end of the spectrum, graphics may disappear altogether, few colors may be employed, and key design elements become the search engine or lists and charts of available content.

The opening "home" page is a preview of the rest of the online publication; therefore, it orients the user not only to the available content, but also to appropriate user attitudes for interaction with that content. The opening screen says "come play" or "here's the information you are looking for." The user anticipates the subject of the Web site and the expected type of user interaction available. Home pages are designed to meet the needs of several kinds of users and are most effective when they can direct each group to the right information. The home page provides a brief overview or a surface level version of the available content, instead of an in-depth immersion into the details of the publication. It is a preview of purpose, design, and appropriate user interaction.

Online publication style is established through continuity of design elements on pages following the opening Web page. The continuity of page elements (backgrounds, consistency of headings, placement of navigation links, and consistency of typeface and type colors) provides an ongoing orientation and identity for the publication. Publication style maintains the image of the publication's sponsoring organization. Users know that they are still using the same online publication (and haven't moved to another online publication by accident). It maintains the overall attitude and purpose of the electronic document. Publication style helps maintain the appropriate user's role of being a player or information seeker. Consistency in the use of design elements helps the user navigate forward and back through the available information and creates a pattern of expectations for future navigation.

The continuity of publication style simplifies the design process because it allows the designer to make certain rhetorical choices in the development of a few key pages and then simply save versions of that page and change the content for new pages. This process of using one page as a template for other pages is a standard development process for building online publications and is rhetorically effective because it helps establish the publication style that will maintain a common attitude and purpose throughout the online publication. One additional benefit of following a publication style template is that it creates an expected norm for the reader, so if a designer wants to BREAK OUT of the pattern of expectation for EMPHASIS the user will immediately notice it. If every page were designed with different design elements, it would be difficult to provide that moment of surprising difference.

Orientation

Orientation is an ongoing process of knowing where you are, what you are reading, how much is here, and where you can go from here. One of the complaints about early hypertext publications was that users could quickly follow hypertext links and connections to a point where they became disoriented. Users become lost in virtual space—they don't know where they are, they don't know how they got there, and they don't know how to get back to a starting point. Current conventions of Web design are addressing this problem so users will be able to orient themselves immediately from the opening screen and maintain this sense of orientation throughout the online publication.

With a traditional print publication, readers orient themselves to the book by considering its size, title, and table of contents; by reading the book chapter titles and headers; and locating bookmarks or dog-eared pages to maintain a sense of where they are in the process of reading the book. Online publications help the user orient to the publication's scope by providing an overview of the whole book on opening screens or on site map pages. The primary principles for maintaining orientation are (a) label every page and chunk of information clearly and (b) provide a means of previewing things before viewing. The online publication orients the user to his or her current location within the book through page titles and variations in the navigation options. The design principle can be summarized as "context before details." An example of previewing before viewing is the use of thumbnails of large photographs. The user sees the miniature preview before deciding to download the more detailed larger graphic. In the same fashion, navigational links should be descriptive enough that users anticipate what they will get before they actually visit the linked page. Maintaining orientation means allowing users to know where they are and allowing them to make intelligent choices before going to new places within the online publication.

Web browser features have also been designed to help with the problem of orientation in online publications. Browsers include backward and forward buttons, a history panel, favorites or bookmark pages, and titles of Web pages at the top of the browser's window frame. Even if the publication does not contain internal orientation elements, these browser features help users maintain a sense of where they are, where they have been, and where they can go next during their online experience.

Navigation and Layering Depth of Online Content

The design assumption of most printed publications is that the reader will become interested in the text and read from start to finish. Of course the reality is that readers start, jump around within the text, skim the graphics, and more often than not rarely finish reading a text completely, with the possible exception of novels. Traditional print media offers several means of aiding readers in the process of jumping around within the text—table of contents, chapter titles, page headings, indexes, or physical means of marking certain sections of text such as the letter tabs on the pages of some dictionaries. How do online publications allow readers to navigate to desired sections or specific information within the text?

The primary means of navigation is through clearly labeled links that simultaneously orient the user to available categories of information as well as providing the means of going to that information. The default indication of links in Web browsers is to underline words that are links to other pages. Graphics, navigational image maps, and special buttons also may serve as navigational links. Web designers employ several strategies to help users notice navigational options, such as simple text prompts, guidelines or introductory tutorial pages suggesting paths of navigation, or rollover interaction with the user's cursor. Once users have navigated to one or two pages, the navigational strategy of the online publication should be apparent by the continuity of links and repeated arrangement of navigational prompts. If users become disoriented, a navigational link back to the "home" page commonly lets them return to their starting point in the electronic file.

Most online publications are designed to begin with overviews and brief samples of more in-depth content available. For example, an online newspaper may include only the headline and lead sentence to several news stories on the opening screen, with a closing navigational link for each story prompting the user to seek the complete story. This layering of content is the fundamental navigational strategy of reading online publications—a back-and-forth movement from overviews and introductions to detailed information. The user navigates out from familiar online pages such as the home page and back to start another search for information. Each major section of the online publication takes on the character of a home page for that section, presenting an overview with links to more in-depth coverage of its topics. Good Web designers let users know that they are leaving the main site by coding the link to spawn a new browser window for each outside link. As users move into deeper levels of content, the links may take them out of the original Web site into archived information, databases, or other related Web sites, but most online publications retain a means of allowing the user to navigate quickly back to their starting point.

Interactivity

The primary interaction with online publications is navigation to various available screens to read or obtain available information, but additional levels of interactivity are possible because online publications may use all of the available resources of the computer. The level of interactivity built into an online publication is closely related to its rhetorical purpose—entertainment publications are going to provide more opportunities to play with the text, whereas reference publications are more likely to emphasize searching for information as the primary user interaction available.

Placing these levels of interaction on a continuum grid, the two axes of interactivity would consist of (a) the extent of user control of navigation and (b) the extent of reconfiguration of the text. If users have no control of navigation through the online publication, then they are merely passive receivers of the information. Like television, they can merely turn it on and watch the information as it appears on the screen or they can turn it off. Few online publications restrict the user to a limited linear presentation of the content.

A middle level of interactivity would allow the user more control of navigation through the text, with multiple means of controlling the navigation. The user is allowed to shape the paths of interaction, although the content remains relatively unchanged.

At the highest level of interaction, users not only choose their own paths of navigation through the online publication, but also reconfigure or change the content itself, essentially becoming a co-author of the text, which remains following the interaction. Each online publication establishes the parameters of interaction allowed—controlled presentation, multiple means of navigation, or "building block" approaches of altering the text in the process of interacting with it. The highest level of interactivity allows the user to add new content, to create new paths or navigational guides through the electronic text, or to remove and change existing content. Most online publications reside in the middle zone of interactivity, providing multiple means of navigational control and allowing users to reconfigure the display of the information on their own computer screen, but limiting the ability to reconstruct the online publication beyond feedback and threaded lists of responses to the publication. Online publications often include e-mail addresses and provide forms that generate e-mails from readers. A few online publications actually allow the reader to reshape or continuously change the existing publication in a collaborative spirit of co-creation.

Designing for Dynamic Maintenance

Online publications are valued for up-to-date information, which suggests that someone is routinely maintaining and changing the pages. If someone e-mails the Webmaster with a complaint or correction, the prevailing assumption is that the online publication will be corrected immediately. The responsiveness of the online publication to its readers and to changes in knowledge depends on the dynamic maintenance provided by the Webmaster or Web editors watching over the information.

To simplify the process of updating the information in online publications, certain areas need to be designed to feature these updates or new information links. Often the home page includes an area reserved for promoting new content and guiding returning users to the latest information and changes available. For rapid updates, these dynamic areas need to be designed for simple edits, that do not require complicated animation or graphic changes. The modularization of information into small chunks and discrete files also helps make changes to small parts of the online publication possible without having to revise the layout and design of the entire Web site. Quality online publications retain their value as reliable, up-to-date sources only if they are continually updated and changed. If they are simply designed and placed online, they soon resemble the ghosts of old texts printed on yellowed, brittle paper.

CONCLUSION: POSSIBILITIES FOR THE FUTURE OF ONLINE PUBLISHING

Online publishing will continue to grow in scope and variety, especially through the simple standards of the Web such as HTML, graphic formats, and improved compression techniques that make multimedia more effective on the Web. Portable electronic devices will be more easily connected to the Web through wireless or satellite transmission, and the limitations of battery devices will be overcome as electronic books begin to resemble electronic sheets of paper. The real struggle will be intellectual property issues—the clash between dynamic, collaborative technology for easy distribution among users versus static, officially published documents controlled by information content owners.

Database publishing will grow as owners of large bodies of content develop ways for users to create custom publications from the archived collections of available content. For example, Pearson Custom Publishing (http://www.pearsoncustom.com/database/merc.html) now offers *The Mercury Reader,* an anthology of literature based on nearly 500 classic and contemporary essays and works of literature that professors can customize to meet the needs of a particular course. The content industry will continue to seek ways to constrain the technology, but the only lasting means of competing is to provide the added value of quality production. Users will enjoy, appreciate, and pay for quality online productions, especially as they have more and more access to a large quantity of choices.

On a more personal note, the individual writer continues to gain the increased ability to become a publisher with online technology. In 2002, computer magazines and the news media celebrated the explosion of interest in blogs—personal journals and commentaries updated on a daily basis. In terms of structure, blogs are a sequential series of Web site entries arranged in reverse chronological order. Blogs can be updated easily through access to a Web browser, and they have the voyeuristic appeal of reading someone's personal diary. Through individually published blogs, the writer can reach a broad readership, especially if that writer has a special area of expertise or a unique perspective of interest to a group of readers. Those who understand the technology of online publishing can become their own publishers with daily editions and updates to their own blogs.

Effective online publishers will know how to gather digital content and organize that content into effective information architectures and how to design interactive experiences for a wide range of users. They will know how to maintain publications for immediate use and archive valuable static information for future use. And they will understand strategies for distributing these texts online with the conventions and standards of the global network, the World Wide Web.

GLOSSARY

ASCII A standard text-only format used since the earliest days of personal computing, with minimal text format beyond characters spaces and punctuation.

Audio books Publications designed to be listened to, through digital or online devices.

Blog or weblog A personal journal of commentaries published online with frequent updates organized in a sequential series of Web site entries arranged in reverse chronological order.

Cache An area of the user's computer memory in which downloaded files are temporarily stored for viewing in a Web browser.

Directories Lists of available information, traditionally organized as guides to large categories of information.

Digital content Text, graphics or other information prepared and stored as a computer file.

Dynamic content Information that frequently changes.

Dynamic maintenance Updating or changing information on a regular basis.

Electronic book (e-book) An electronic file (a complete text equivalent to a paperback book) that can be downloaded from the Internet and viewed on a variety of electronic devices.

Electronic-book reader (e-book reader) A hand-held portable electronic device designed to display electronic books.

Extensible markup language (XML) A file format containing tags for marking content for display on a variety of browsers, with the ability to be quickly reconfigured for different viewing devices, such as a cell phone, PDA, or desktop computer monitor.

File transfer protocol (FTP) A standardized procedure for exchanging computer files from one computer to another through the Internet.

Hypertext A type of electronic text characterized by links within sections of the text or between related chunks of text.

Hypertext markup language (HTML) A text-only file format containing tags for marking content for display by a variety of web browsers.

Hypertext transfer protocol (HTTP) A procedure for exchanging computer files from a server computer to a user's computer through the Internet for viewing in a Web browser.

Index or home page The opening page or screen of an online publication.

Information architecture The art and science of organizing information to help people effectively fulfill their information needs.

Interactivity Refers to the range of interaction available within an online text—the extent of control of navigation and reconfiguration of the text by the user's actions.

Landscape view The horizontal orientation of most computer screens (and most online pages).

Layering content The design strategy of providing overviews or condensed versions of information with underlying rich layers of in-depth details.

Meta-tags Hidden content, within an HTML document, often used by search engines to help target the document to appropriate readers with key words or a description of the text.

Modularization The process of breaking information into small modules of text and separate computer files.

Navigation The process of moving within the available content of a publication and selecting alternative paths through an online publication.

Online publication style The continuity of design choices (typeface, arrangement, graphic treatments, backgrounds) carried from page to page throughout an online publication.

Online publishing The process of preparing a text for public distribution through the Internet for viewing on a computer screen.

Optimizing A process of compressing files (especially images) to download more quickly.

Orientation 'Users' understanding of the scope and structure of a publication, allowing them to know where they are within the document and where they can go from there.

Personal digital assistant (PDA) A small handheld electronic device for storing small amounts of information, often also used as electronic book readers.

Portable document format (PDF) A format developed by Adobe Corporation to let users download or view carefully formatted documents online without losing the typeface and graphic formatting of the original.

Portrait view The vertical orientation of most books and printed pages.

Project Gutenberg An extensive, collaborative effort to build an online library of shared texts of public domain publications available in simple text-file format.

Publishing The process of preparing a text for public distribution, traditionally on paper.

Pulling content The process of bringing selected information to the user's computer screen.

Pushing content The process of sending information to the user's computer even though it was not requested.

Remediation of print The process of transforming existing technologies and strategies of printing and publishing into new approaches for the computer screen.

Rhetorical purpose The global aim of a publication—to entertain, persuade, or inform.

Rich text format (RTF) A universal text format that allows files to be shared between a variety of systems and word-processing programs, with limited paragraphing, tabs and typeface options.

Search engine A software program that compares content in a database with parameters of a user's requested search.

Standard generalized markup language (SGML) A markup language for coding various elements of text including styles and formats.

Single-source publishing The strategy of developing and storing only one copy of small modules of information that may be used and reused for a variety of publications instead of re-creating it within each document.

Site architecture The design of a Web site; not referring to the layout of pages, but to the structure of the entire site.

Site map A listing of available pages in a Web site, providing a visual representation of the online publication with links to all content.

Static content Information or publications that remain unchanged with multiple uses (such as a traditional print media publication).

Thumbnails Small graphical representations of larger graphics or pages available for viewing.

Uniform resource locator (URL) The address of any file on the Internet.

User A person interacting with an online publication.

User-centered design The process of designing for the known needs of intended users.

User testing The process of testing the usability of a publication with intended users.

Usability The goal of designing an online publication for maximum effectiveness for its users—matching rhetorical purpose with functional operations of the electronic text.

Virtual libraries Collections of electronic texts available only through the Internet.

Web site A collection of Web pages and files connected by navigational links.

CROSS REFERENCES

See *Digital Libraries; Extensible Markup Language (XML); File Types; Multimedia.*

REFERENCES

Bolter, J. D., & Grusin, R. (1999). *Remediation: Understanding new media.* Cambridge, MA: MIT Press.

Rosenfeld, L., & Moreville, P. (1998). *Information architecture for the Web: Designing large-scale Web sites.* Sebastapol, CA: O'Reilly.

FURTHER READING

Aarseth, E. J. (1997). *Cybertext: Perspectives on ergodic literature.* Baltimore: Johns Hopkins University Press.

Birkerts, S. (1994). *The Gutenberg elegies: The fate of reading in an electronic age.* Boston: Faber and Faber.

Bolter, J. D. (1991). *Writing space: The computer, hypertext, and the history of writing.* Hillside, NJ: Erlbaum.

Deibert, R. J. (1997). *Parchment, printing, and hypermedia.* New York: Columbia University Press.

Dvorak, J. C. (2002, February 5). The blog phenomenon. *PC Magazine.* Retrieved October 2, 2002, from http://www.pcmag.com/print_article/O,3048,a = 21865,00.asp

Fidler, R. F. (1997). *Mediamorphosis: Understanding new media.* Thousand Oaks, CA: Pine Forge Press.

Fidler, R. F. (1998). Life after 2001: Redefining print media in the cyber age. *The Future of Print Media.* Retrieved May 15, 2002, from http://www.futureprint.kent.edu/articles/fidler02.htm

Landow, G. P. (1997). *Hypertext 2.0.* Baltimore, MD: John Hopkins University Press.

Lanham, R. A. (1993). *The electronic word: Democracy, technology and the arts.* Chicago: University of Chicago Press.

Marcus, A. (1992). *Graphic design for electronic documents and user interfaces.* New York: ACM Press.

O'Donnell, J. J. (1998). *Avatars of the word: From papyrus to cyberspace.* Cambridge, MA: Harvard University Press.

Stim, R. (2000). *Getting permission: How to license and clear copyrighted material online and off.* Berkeley, CA: Nolo Press.

Sutherland, K. (Ed.). (1997). *Electronic Text: Investigations in Method and Theory.* Oxford: Clarendon Press.

Thurston, T. (1999). New questions for new media: scholarly writing and online publishing. *American Quarterly, 51,* 250–253.

Wearden, S. (1998). Landscape versus portrait formats: assessing consumer preferences. *The Future of Print Media.* Retrieved June 13, 2002, from http://www.futureprint.kent.edu/articles/wearden01.htm

Online Religion

T. Matthew Ciolek, *The Australian National University, Australia*

INTRODUCTION

Since the late 1970s religion has entered and colonized even the newest of all frontiers: the Internet and its information services. Over the last quarter of the century, the link between religion—that is, people's "relation to that which [they] regard as holy, sacred, spiritual, or divine" (*Encyclopaedia Britannica*, 1987, p. 1016)—and digital communications technology became well established and, according to statistical indices, becomes stronger with every month. Around 2000, there were more than 1 million online religion Web sites in operation worldwide (Brasher, 2001, p. 6). Furthermore, a U.S. survey from 2001 reveals that no less than 25% of the studied Internet users (i.e., 28 million people) had accessed the Net in search of spiritual information and to liaise with others about religious issues. This is a marked increase from the figure of 21% the studied users (i.e., 19.5 million) that was recorded a year earlier (Pew Internet and American Life Project, 2001). Other sources expect that 90% of North American churches are likely to be connected to the Net before the end of the decade (Thumma, 2002). Accordingly, experts predict that by 2010 more than "100 million Americans will rely upon the Internet to deliver some aspects of their religious experience" (Barna Research Online, 2001b).

This chapter reviews the last 25 years of major electronic developments among religions and spiritual communities from all parts of the world. By doing so, it will extend and complement the more detailed, and culturally and geographically localized studies of Barna (2001), Barna Research Online (1998, 2001a, 2001b, 2002), Bedell (2000), Hayes (1999), Bunt (2000a, 2000b), Miller and Slater (2000), and the Pew Project (2001). Because of the Net's most uneven distribution and accessibility across the globe, however, it is unavoidable that the majority of the cited examples—although being selected from both the core and fringes of major spiritual traditions as well as from the realm of "New Religions" and fledgling cults—will be derived from networks situated in the United States, Canada, and Europe. Should any of the electronic resources mentioned be relocated or closed down, they may be retrievable from the Internet Archive (http://www.archive.org).

This chapter only examines spiritual and social uses of computer-mediated communications, and it will tackle "religions-that-have-moved-online" as well as the emerging universe of the "religions-that-were-born-online" (see also Hadden & Cowan, 2000b, pp. 8–9; Helland, 2000, p. 207; Herring, n.d.). Because of the lack of space, I refrain from commenting on the connections between online religion and contemporary politics, construction of ethnic and cultural identities, or the intercommunal conflicts.

Finally, although the chapter focuses on the world of Internet resources, it also covers aspects of ARPANET (the first computer network in the world), as well as Usenet, Fidonet, bulletin boards systems (BBSs), and Bitnet systems, because it takes a macroscopic and long-range point of view.

THE INTERNET AND RELIGION—EARLY DEVELOPMENTS

Establishment of the First Religious "Communities" and Discussion Groups

The meeting of the Net and religion has been a complex and protracted affair. Digital computers, which have been in use since 1944, were first put together in the form of a network in October 1969. Slowly at first, then more energetically, the network grew in size and capability. In addition to its original computational uses, the Net unexpectedly became a handy tool for long-distance communicational exchanges via e-mail (since 1971), Usenet (since 1979), and mailing lists (since 1981). Additional innovative software—such as BBSs (since February 1978)—in which messages are distributed over phone lines between users with personal computers equipped with modems (invented in 1977)—and the file transfer protocol (FTP; since 1985) were also conceived and constructed. These enabled the online storage and dissemination of large volumes of software, numeric data sets, and text files. In the hands of hundreds and then thousands of enthusiastic users scattered across North America and Northern Europe, the Net has evolved into a rich medium suitable for purposes ranging from the strictly technical and professional, to the more private, more informal, and more emotional social encounters.

In June 1978, all these gradual changes culminated in sections of the CommuniTree BBS forming a vibrant virtual community (Rheingold, 1994, p. 134), or more accurately, a virtual special interest group designed and built around a religious practice. Independently, the ever-increasing sophistication of the networking technologies, combined with the rapidly growing number of people with access to the Net led to the creation, in March 1983, of a Usenet discussion group called net.religion. In retrospect, these two parallel and pioneering online developments proved to be highly seminal. In addition to serving the immediate needs of their members, they were also the first working (and replicable) examples of *free public access* electronic environments supporting: (a) regular online social interactions between like-minded strangers, (b) spontaneous sharing of digital offerings (e.g., texts, images, software), and (c) vigorous discussions of the religious, ethical, and philosophical issues.

Notable Religious Online Firsts

Soon several other notable "religious online firsts," both online resources and related events, took place in various parts of the world. A select sample of these happenings is listed here in a short time line, starting with the experiments of the "CommuniTree" group and ending in 1996 with the first scholarly conference on religion and the Internet.

Early Online Religious Developments 1978–1996—a Time Line

1978 188 Internet hosts, a handful of BBS systems, and introduction of CompuServe dial-up services.

June: CommuniTree#2 BBS in Santa Cruz, California, forms the "Origins" group, an experimental "online community" focused on observance of a syncretic, "post-Enlightenment" spiritual practice. This pioneering project was abandoned in 1982 when it was trashed and vandalized by a sudden influx of aggressive outsiders (Allucquere, 1991).

1983 562 hosts, 10 million personal computers (PCs) in the United States.

February: First religion-orientated, general purpose Usenet discussion group, net.religion. In September 1986, following the Great Renaming of the Usenet, the original group was split into several groups forming the alt.philosophy, alt.religion, soc.culture, soc.religion, and talk .religion hierarchies.

May: First e-mail religious newsletter "United Methodist Information," http://www.ecunet.org/history. html. Initial use by 14 subscribers.

1984 1,040 hosts.

February: First Usenet discussion group dedicated to a specific religion, net.religion.jewish.

1986 5,089 hosts; 241 Usenet groups.

January: First online religious ritual: the memorial service for the late crew of NASA's *Challenger* spaceship.

March: First conference dealing with religion and computers, "CAMCON: Computer Applications for Ministry Conference," Los Angeles, California.

April: Establishment of the first religious network—Ecunet, now at http://www.ecunet.org. Initial online access to 15 Christian communication groups.

1989 80,000 hosts; 40 Internet chat-room (IRC) servers; 44 million PCs in the United States.

February: First scholarly mailing list for the study of a specific religion—Judaica/He'Asif, now H-Judaic at http://www2.h-net.msu.edu.

1990 313,000 hosts; 1,300 Usenet groups. Establishment of the first FTP archive with religious (Tibetan Buddhism) scriptures, "Asian Classics Input Project," now at acip.princeton.edu.

1991 376,000 hosts; 1,850 Usenet groups; Gopher, Wide Area Information Server (WAIS), and World Wide Web technologies introduced. Establishment of the first scholarly mailing list dedicated to the study of all religions, Religion@harvarda.harvard.edu.

1992 727,000 hosts; 20 Web servers; 4,300 Usenet groups. Establishment of BuddhaNet, the first BBS ever run by a monk, now at http://www.buddhanet.net.

June: Publication of the first guide to religion and spirituality online—"An Electric Mystic's Guide to the Internet" by M. Strangelove, now at wings.2buffalo.edu/sa/muslim/isl/dirs/elmystic.guide3.txt.

December: Establishment of the first virtual congregation, "The First Church of Cyberspace," a nondenominational, untraditional church with an emphasis on the virtual aspects of its belief system, http://www.godweb.org and www.afn.org/~fcoc/.

1993 1,313,000 hosts; more than 60,000 BBSs in the United States; more than 200 Web servers; 8,300 Usenet groups; "Mosaic" graphic Web browser introduced. Publication of the first electronic journal dedicated to the study of religions, *Discus—Religious Studies Journal.* The periodical was produced on stand-alone computers and distributed by hand and mail on computer disks.

May: First mention of a parody site, "The Church of the SubGenius," "a for-profit cult that believes in the pursuit of Slack," now at http://www.subgenius.com.

September: Establishment of "Nurel-L," an e-mail group discussing new religious movements, now at http://www.ucalgary.ca/~nurelweb/.

1994 2.2 million hosts; 850 Web servers; 750 WAIS servers; 10,700 Usenet groups; "Labyrinth" graphic three-dimensional virtual reality Web browser introduced. Establishment of the first Islamic Web site, The Mosque of the Internet, http://www.mosque.com. Establishment of a "comedic sin" site, EvilPeople, INC., http://www.gaijin.com/EvilPeople/. The site tests religious commitment of its visitors by explicitly inviting them to betray their core spiritual principles.

January: The Benedictine order establishes the first monastic mailing list, OSB-L@vm.marist.edu.

March: Web publication of electronic text files of the Bible and Qur'an, at http://www.hti.umich.edu.

April: Establishment of the first monastic Web site, the Monastery of Christ in the Desert, http://www.christdesert.org.

1995 5.8 million hosts; 23,500 Web servers; 16.5 million Usenet users; more than 950 FTP archive sites.

June: Publication of the first Islamic e-periodical, *Renaissance: A Monthly Islamic Journal,* http://www.renaissance.com.pk. Establishment of Virtual Memorial Garden, http://catless.ncl.ac.uk/vmg/, with tributes to people and pets.

September: Establishment of the first European network dedicated to online pastoral care, "Die Internet Seelsorge/Parrocchia Internet," http://www.seelsorge.net.

1996 14.3 million hosts; 100,000 Web servers; "Hotmail," free Web-based e-mail, introduced. Establishment of the first Zoroastrian cybertemple, http://www.zarathushtra.com. Islamic use of computers are officially approved by the Qom (Qum) Seminary in Iran.

April: First CyberSeder is held during Passover by Congregation Emanu-El of the City of New York (http://www.emanuelnyc.org).

September: An electronic magazine, *Mediamatic Magazine,* http://www.mediamatic.net, publishes a special issue dedicated to religion online.

November: First annual European Christian Internet Conference (ECIC), http://www.ecic.info, Frankfurt, Germany.

December: *Time* magazine publishes, in a special issue dedicated to religion online, an article "Finding God on the Web" by J. Chama (1996).

Following these pioneering developments (see also Ciolek, 2003), further religion-focused resources and communication channels have been established all over the world. The Internet, unlike the five more costly and more centralized traditional media (i.e., film, books, newspapers, radio, and television) has become freely used by both the mainstream and formalized traditions, as well as by their countless doctrinal and organizational offshoots. Simultaneously, there emerged a lively redundancy and replication of online efforts. Tens and dozens of sometimes parallel and sometimes successive e-mail and news groups, chat rooms, FTP sites, weblogs, and Web sites can be seen to cover exactly the same areas of interest.

STAGES IN THE DEVELOPMENT OF RELIGIOUS USES OF THE INTERNET

Networked religions in all parts of the world seem to follow a basic, four-part developmental sequence that parallels the technological milestones of the Internet itself as well as the skills acquired by members of religious communities:

Phase 1: Computerized Information

This phase predates the use of computer networks and involves first experimental and then subsequently more confident uses of standalone computers. This phase is driven by intellectual enthusiasm about possibilities of new technology. The religious uses of that stage are modest and revolve around (a) textual analyses of key religious documents and (b) the creation of detailed help and how-to files that are shared on diskettes. Throughout, (c) long-distance contacts with other people are maintained through phones, faxes, traditional mail, and exchange of computer disks. Limitations in this phase are many and relate to inadequate computer memory, processing speed, storage, fonts, resolution, graphics, as well as low interoperability of various computers. The target audience of these efforts is one's immediate circle of colleagues and friends. Information is created by independently operating individuals and their loose coalitions.

Phase 2: Networked Interchanges

This phase is marked by the use of long distance computer communications (e.g., e-mail, BBS, Usenet). New behaviors emerge, such as frequent, one-to-one and one-to-many exchanges that may be traceable or anonymous. Fresh challenges also arise, such as the never-ending influx of inexperienced users; the dire lack of netiquette; and the growth of malicious spamming and flaming. Simultaneously, (d) intellectual contacts among strangers flourish online and (e) virtual communities and groups of like-minded people start to coalesce, while (f) files with FAQs (frequently asked questions) for major topics are established and recommended to new participants. These ad hoc activities are directed to an anonymous, amorphous, global, and borderless audience. Information resources are developed by small collaborative groups.

Phase 3: Networked Documents

This phase is marked by the increased production of electronic publications. In the third phase, innovative religious uses of the Net include the following: (g) translation of earlier paper resources into digital online formats, (h) proliferation of online and printed directories and guides to the emerging religious cyberspace, (i) a flurry of papers (for instance, a 2000 paper titled "The Internet as a Metaphor for God," http://www.crosscurrents.org/henderson.htm) displaying an awareness of the planetwide cyberspace and its possible metaphysical implications. (j) Critical analyses of online religious developments appear in print, journals, and conference papers. Also, (k) official statements regarding the institutional attitude of a religion to the Internet are formulated and released. The gap between the moment a religious activity is introduced to the Net, and the time that such a development is noticed, accepted, and doctrinally justified by the respective religious authorities can be lengthy. It took the Catholic Church no less than 15 years, from 1983 to 1998, and Islam no less than 8 years, from 1988 to 1996, to assess and accept officially the new medium. At the same time, (l) ambitious e-publishing operations are launched by an interest group or a task force. These target a global audience of co-religionists and like-minded individuals.

Phase 4: Seamless Uses

In this phase technical innovations cater to the growing dependency on speedy creation and gathering of information. Texts and hypertexts can now be freely combined with images, videos, numbers, and sound. Documents and records of e-mail communications can be cross-linked; full-text indexing of Web documents becomes a norm. Massive, the planetwide cyberspace operates on a 24-hour-a-day, 7-days-a-week, 365-days-a-year basis. Weblogs proliferate. The Internet officially becomes accepted as the sixth branch of mass communication. Simultaneously, additional religious applications arise: (m) pastoral contacts flourish online, (n) electronic participation in live- and cyber-rituals becomes commonplace, (o) online hubs for specific religious communities are created and used, (p) portals offer customized access to multiple spiritual resources, (q) systematic sociological and demographic studies on the religious uses of the new medium are undertaken, (r) books with ethnographies as well as theoretical accounts of the online religion and religion online are published (e.g., Brasher, 2001; Hadden & Cowan, 2000a; Lochhead, 1997; Zaleski, 1997). Also, (s) religious and commercial organizations start underwriting costly, complex, and long-term online projects. These projects court attention of specific target audiences (e.g., youth only or individual lifestyle or language groups) from the pool of online users worldwide.

This sequence of phases describes past developments. With the arrival of additional technologies, content, and uses for the religious Internet, it will be possible to distinguish further phases in a few years time.

FACTORS INFLUENCING ONLINE PRESENCE

Certainly, not all religions make similarly intensive uses of the Net or have similar impacts on their online associates and neighbors, sometimes even when they operate within identical socioeconomic contexts. Also, not all religious communities are likely to progress at the same time or with the same speed through the four phases of development just described. It has been suggested (Ciolek, 2000) that the maturity and ease with which a given group of people uses the Internet is, in fact, an outcome of multiple, concurrent variables. The major factors are the following:

Number of Users

Bigger online populations create and evolve resources more readily than smaller ones. This is because they are able to draw on larger pools of knowledgeable electronic authors and publishers and interested readers. Also they are more likely to keep such online resources alive after the enthusiasm of their initial creators has dissipated.

Internet Infrastructure and Wealth

Not surprisingly, access to a reliable source of electricity, to computer hardware and software, to appropriate font sets, to modems, permanent or dial-up lines, reliable telephone and radio links, and connectivity providers varies widely between locations. For example, the Internet has been in use in North America since the late 1969, yet as late as June 2002 it still has not reached Haiti, Iraq, Somalia, Sudan, Syria, and Zaire (Internet Software Consortium [ISC], 2002). Money is also an important factor—not only because access to networked communications infrastructure must be purchased but also because wealth is positively correlated to use of the Net. Church ministers with higher incomes (and better education) were found to make more intensive use of the Internet than their less affluent and less educated brothers (Barna Research Online, 2001a).

Technical Skills and Training

Fluency in computer skills, Net navigation, online traffic rules (netiquette) and effective e-publishing are distributed unevenly across geographies, cultures and socioeconomic classes. Where these skills are not kept up to date, the online presence of a given congregation becomes meager and ineffectual.

Administrative Regulations

Administrative regulations also exert a powerful influence. Islamic Saudi Arabia hesitantly allowed the Internet into its country only in January 1999 (ISC, 2002), and as late as February 2002 the military government of Myanmar (Burma), a Buddhist country, continued to treat personal use of the Net as a criminal offense (U.S. Department of State, 2002). The existence of other government strategies can also be noted. For instance, in places such as the predominantly Christian nations of the United States, the United Kingdom, and Australia, all types of religious information (and misinformation) are freely accessible online. The agnostic Communist China, on the other hand, although it promotes the Internet in its country, relies heavily on police, censors, and automated content filtering (Zittrain & Edelman, 2002) to limit their citizens' exposure to any type of religious information, especially that associated with the Falun Gong/Dafa (http://www.falundafa.org) movement or Tibet's Dalai Lama (http://www.tibet.com).

Attitudes Toward Global Communications

In addition, the degree of online presence is influenced by doctrinal attitudes. Some traditions (e.g., Buddhism, Christianity, and Islam) have already declared their strong interest in and approval of educational and pastoral uses of online communication (Azzi, 1999; Fajardo, 2002; Kadous and Al-Wassiti, 1996; Pannyavaro, 1998, 2002; Pontifical Council for Social Communications, 2002a, 2002b). Moreover, new opportunities for the proactive uses of the Net for evangelism, mission work, and other ministry pursuits are also noted and welcomed.

At the same time, thoughtful reservations are also being voiced. Islam is clearly worried (Bunt, 2000a, p. 9) about the corrosive impact that misguided, heretical, or indecent electronic messages (and the processes of modernization and secularization as such) may have on the minds of the online and offline "umma" (i.e., community of Muslims). The Church of Scientology (2000) is apprehensive about unauthorized and malicious online circulation of its copyrighted religious texts and leaked-out internal documents. Finally, Christian authorities are chiefly concerned that "we [do not become] directed by the [electronic] tools themselves, become addicted to their use, or use them inappropriately [. . . and that they do not] become a divisive force in our communities or make us elitist, dividing us from those without the same access" (Benedictine Internet Commission, 1998).

ELEMENTARY STRUCTURES OF RELIGIOUS CYBERSPACE

The Offline and Online Contexts of Networked Religion

Individuals and groups conduct their online religious activities in the context of two concentric information ecologies. The first and broadest is the world of nondigital and nonnetworked information—both lay and religious—that is produced and distributed through the five traditional media, as well as face-to-face contacts, letters, written notes, telephones, and faxes. The nonnetworked world is immense. For example, the British Library (http://blpc.bl.uk) alone contains nearly 38,500 books on the subject of religion, an equivalent of 38.5 gigabytes of text, and this is but one library among many.

Within that physical world nests the realm of networked information. It forms massive archipelagos of thematic 'cyberspaces.' The religious cyberspace—the body of electronic documents and communication channels that directly and indirectly deal with religion as a whole, or its specific manifestations—is dominant among these. For instance, in February 2001 an online study (Dawson, 2001) of the contents of Web pages indexed by altavista.com found, in a sample of 46.6 million English-language documents, that 17% of them dealt with religion. In comparison, 50% dealt with cars, 25% with sex, and 8% with investments. Twenty months later, in October 2002, a similar query by this author showed that the topic of religion had maintained a strong presence in the altavista.com database (religion 15%, cars 40%, sex 40%, investments 5%) in a sample of 45.7 million English-language documents. At that time, it had a much greater presence in google.com database (religion 24%, sex 35%, cars 23%, investments 18%, a sample of 18.6 million). In other words, at the time of writing, the Net contains at least 200 gigabytes of electronic English-language materials that address religion, that is, more than 5 times the equivalent holdings of the British Library.

The Structure of the Religious Informational Archipelago

The religious informational archipelago is strongly fragmented along doctrinal, cultural, linguistic, and organizational lines. In addition, special interests (e.g., being gay, feminist, or vegetarian) also give rise to distinct clusters of documents and servers. Some of these resources stay resolutely aloof and rarely exchange hyperlinks or data with their spiritual cyberneighbors. Others form loosely structured constellations of sites and are more open to casual contacts, exchanges of information, or even exchanges of ideas. Their online visibility varies widely also. Nearly all of them, however, can be in principle, found on the Net through the judicious use of multiple key words applied to intelligent search engines such as google.com or wisenut.com. The overall number of nodes in religious cyberspace is several million strong, but the overall repertoire is surprisingly small. There appear to be only 10 major categories of online religious structures, as follows.

Community Hubs

These are electronic environments maintained by a particular congregation. An example is the Sri Ventakeswara Hindu Temple in Chicago (http://www.balaji.org), which offers information about the activities and facilities of the temple. Another is the Web site of the Order of Buddhist Contemplatives (http://www.obcon.org). These medium-sized environments typically offer local-interest information, including (a) a brief history of the community and its objectives; (b) a page with postal and e-mail addresses, contact details, and instructions on how to get there by train, bus, or car; (c) schedules of regular internal activities; (d) details of events, training courses, and spiritual retreats open to the public; and (e) information on schools and colleges that the community operates. In addition, community hubs often include (f) documents with biographies of their founders or spiritual leader, (g) links to or pages containing collections of their religious writings, and (h) pages providing general spiritual advice and religious outreach. Community hubs can also facilitate (i) local discussion groups. Sometimes they support (j) an online bookshop or the sale of religious artifacts and mementos or offer (k) online samples of the written and visual artistic expressions of congregation members. The inner organization of community Web sites, regardless their actual spiritual provenance, tends to be strikingly similar (Ciolek, 2000, pp. 652–653).

Establishment Hubs

These are online environments created by an organization formally claiming custodianship of a given spiritual tradition. Establishment hubs are typically aimed at the global community of the faithful. They specialize in the provision of general purpose factual information (often in several languages) such as official documents, news, and corporate details, as well as galleries of electronic images depicting the key artistic works influenced by a given religious tradition. An early example of such a hub is the 1995 "Vatican: The Holy See" (http://www.vatican.va) Web site. Establishment hubs are not only built by institutionalized religions. For instance, "The Celtic Connection" represents "a complete source for Wiccan, Witchcraft and Pagan knowledge & supplies" and is a place "where followers of Wicca, Witchcraft, Shamanism, Druidry and Pagan beliefs meet to celebrate the magickal life" (http://www.wicca.com).

Supporter Sites

These are electronic resources maintained by private individuals who sympathize with a given congregation or tradition (e.g., "Unification Home Page," http://www.unification.net, dedicated to the "presentation of the life, teachings, and public work of Rev. Sun Myung Moon and his wife Mrs. Hak Ja Han Moon"). Such quasi-official sites often provide additional communicational services, especially in the area of anonymous-use public access discussion groups. In addition, they offer congenial places for informal experiments with computer-mediated rituals and art. Also, despite their small sizes, they publish lists of annotated links to related congregations, and, importantly, to other, unrelated spiritual beliefs.

Commemorative Sites

These are electronic environments that tend to "close the gap between religion and popular culture" (Brasher, 2001, p. 20). Typically, they commence their operations without any previously existing constituency but, with the passage of time, they accrue a retinue of followers. For instance, there are "virtual shrines," such as "Flowers for People's Princess! A Virtual Memorial Site for the Everlasting Memory of Diana & Dodi" (http://www.interlog.com/~3mowchuk/kbasket/dflower02.html) or the "John Lennon Memorial"(http://rwbeatles.tripod.com/johnform.html). Virtual shrines idolize celebrities and organize scattered fans into lasting congregations of contributors and electronic pilgrims (Brasher, 2001, pp. 120–139). Other commemorative sites are more ordinary, more democratic and more pensive. For example,"Virtual Memorial Garden" (http://catless.ncl.ac.uk/vmg/) publishes tributes to those who passed away: people (e.g., a dedication to "Daniel Herbert J[. . .], 16 Oct 1913–14 Mar 1996, 'To my Grandpa who was my best friend, I miss you so much'") and pets (e.g., "Pallina, Jun 1994–Jul 1995, 'She was a little white kitten so sweet so happy so . . . Rest in peace!'"). Another and more recent example of a popular commemorative site is "A Virtual Memorial Remembering September 11, 2001" (http://www.xgmc.com/911/hall.htm), a site opened "to all who love and believe in Peace and Freedom."

Challenger Sites

These are electronic resources created by individuals and groups in opposition to the official beliefs and practices of one or more mainstream religious communities. This category is formed by at least four distinct subgroups: (a) heretic/dissident sites for example,"Partenia: Diocese Without Borders" (http://www.partenia.org) or "Ahmadiyya Muslim Community" (http:// www.alislam.org); (b) Opinion sites, for example, "Religious Fundamentalism" (http://rwendell.blogspot.com) with public reflections on excesses of spiritual and moral zealots or the communally run "Esoter@Life" (http://esotera_life.blig.ig.com.br) focused on Eastern Esoterism, Alchemy, Nudism, Perfumes, Dakshina Tantra Yoga, and Secret Societies; (c) Parody sites, for example, the thoughtfully irreverent "Virtual Church of the Blind Chihuahua" (http://www.dogchurch.org), the contentious "SuraLikeIt" (http://dspace.dial.pipex.com/suralikeit) that satirizes the Qur'an, or over-the-top "Landover Baptist Church" (http://www.landoverbaptist.org), which strives (with tongue-in-cheek) to keep "the temple of the living God a clean vessel, untarnished by even a hint of fellowship with the unrighteous"; (d) Comedic Sin sites (a term coined by Brasher, 2001, p. 109), for example, "EvilPeople, INC." (http://www.gaijin.com/EvilPeople/), where visitors can help to "bring about the apocalypse as forseen [sic] in the Book of Revelations," sacrifice virtual goats to the Evil One, and (a feature introduced October 3, 1995) sell their immortal souls. The latter opportunity, according to the site, is not too outlandish because you, the visitor "aren't really using it [i.e. your soul] and probably never will. It's such a tiny thing."

As a rule, all such sites address a global audience. Heretic sites earnestly question—as well as undermine—dominant doctrines and organizational structures and provide means for "alternative" communication, chiefly in the area of online ministry and interorganizational exchanges. Similarly, authors of commentary sites open logbooks of their religion-centered musings to public scrutiny. By availing themselves to responses by total strangers, they hope to engage and influence their visitors. The parody sites are far less earnest and far more manipulative. They know that the outrageous information they present will not fail to tease (and make think) members of the targeted tradition while entertaining visitors. Similar goals motivate the existence of comedic sin sites. There, again, reverse psychology is ruthlessly applied to shake visitors out of spiritual complacency.

Agorae

These are nondenominational environments that support online communications with a religious focus. For example, in October 2002 Yahoo (dir.groups.yahoo.com) reported the existence of 101,371 public access religious online discussion groups. The groups were formed around 49 wide-ranging topics, such as philosophy (3,668 groups), religious education (1,038), interfaith relations (695), atheism (536), meditation (524), and parody religions (514). It also included groups focusing on religion and youth (341), church–state issues (308), women (270), or the creation versus evolution dispute (130). The majority of discussion groups, however, dealt with intrareligious issues arising within Christianity (45,271 groups, 57.5%), Paganism (11,577, 14.7%), Islam (9,256, 11.8%), Judaism (3,101, 3.9%), Buddhism (1,580, 2.0%), New Age religions (1,575, 2.0%), Hinduism (1,467, 1.9%), occult practices (1,222, 1.6%), Satanism (882, 1.1%), Unitarian-Universalism (801, 1.%), Sikhism (386 groups), Shamanism (331), Baha'i (317), Jainism (205), Voodoo (201), Vedism (173), Taoism (117), Zoroastrianism (94), Scientology (68), Santeria (59), Cao Dai (56), and Shinto (17). In other words, at the time of writing 80% of religious communication channels listed by Yahoo revolved around no more than eight spiritual traditions.

Portals

These are large-scale online environments providing a multipurpose access point to specific parts of religious cyberspace. For instance, "Zlabia.com: le rendez-vous des Juifs d'Algerie" offers three chat rooms, three e-shops, announcements and classifieds, and a gallery of historical photographs of the rabbis and synagogues of Algeria. A much larger online venture is run by "Islamicity.com: Islam and the Global Muslim eCommunity." The portal has 10 major zones: (a) Bazaar & Shopping with Islamic products, Islamic books and artwork and Islamic audio and video; (b) Communications; (c) Community & Social; (d) Education Center; (e) Business & Finance; (f) News & Media Center; (g) Mosque & Religious (Scripture & Prophets, Islam Explained, The Five Pillars of Islam, The Islamic Society, References, Related Products and Links); (h) Multimedia Center; (i) Travel Center; and (j) World ePort Center. Similarly, the eponymous "CrossWalk.com," a Christian site, has several cross-linked sections: (a) Bible study tools: dictionaries, encyclopedias, ministry articles, (b) Devotionals: guided prayer, reading, and contemplation; (c) Ministry Audio: commentaries and opinions in the Internet's RealAudio format; (d) Worship Center, or an e-shop selling DVDs with images and music; (e) Newsletters; (f) Family, Fun & Community; (g) News; and (h) Shopping & Resources: singles' area, e-cards, and music. Portals are run as commercial ventures; they use religion as an attractant to large volumes of online visitors, and generate income from advertisements and e-commerce.

Propagator Sites

These are environments dedicated to a mix of specialist tasks: (a) strengthening the faith among believers (e.g., http://www.e-vangelism.com); (b) gaining converts from other faiths (or poaching members from another congregation); (c) managing public relations and providing rebuttals in online contacts with members of other, possibly unfriendly groups. "Jews for Jesus" (http://www.jewsforjesus.org), for instance, is published in two editions: one for believers and one for seekers. The latter offers several facilities, including Evidence for Jesus and Objections, Essays for the Thinking Jew, Downloadable E-books, Multimedia Gallery, Opinions, as well as Message Boards & Chat Room. The activities of that site are skillfully neutralized by another propagator site called "Outreach Judaism" (http://www.outreachjudaism.org), an international organization that offers countermissionary and countercult services and explores Judaism in contradistinction to fundamentalist Christianity. Sometimes, networked missionary activities try to be more subtle and disguise themselves as bona fide educational services. For instance, Bogdanyi (1997) suggested that for the Lutheran church in Hungary, the shortage of quality online material might present "a great opportunity for the church-owned schools . . . By creatively designing information services a 'gentle smuggling in' of Christian values may be accomplished." This pragmatic sentiment is shared by members of other religions as well (e.g., Azzi, 1999; Pannyavaro, 2002).

Bridge-Builders

These are environments dedicated to the promotion of understanding and good will between different traditions. "Islam and the Baha'i Faith," for example, hopes to promote "a better understanding of the relationship between the Baha'i Faith and Islam, and to dispel some of the misconceptions which may have led to feelings of mistrust and suspicion" (bci.org/islam-bahai). Similarly, "The Ontario Consultants on Religious Tolerance" (http://www.religioustolerance.org) insist that "to be tolerant does not require that we accept other beliefs as true. . . . [However, everyone] should be able to follow their own religious beliefs and practices freely . . . [if we do not want to head] for disaster." The bridge-builders amicably offer handy guides to the Internet, as well as multilingual documents, such as those published by the Pontifical Council for Interreligious Dialogue (Arinze, 2001). The council sends messages of good will to representatives of traditional (i.e., so-called primitive or native) religions as well as to Muslims (for the end of Ramadan), Hindus (on the Feast of Diwali), and Buddhists on the occasion of Vesakh (i.e., commemoration of Buddha's birthday). The spirit of these declarations is also complemented by the work of Monastic Interreligious Dialogue (http://www.monasticdialog.org), a task group established offline in 1978, which now uses the Internet to document emerging conversations, encounters, and mutual exchanges between monastics of different institutions and faiths.

Analysts

These are online environments built by universities, libraries, research institutes, and scholarly conferences. Such sites tend to provide materials and disciplined overviews and analyses not readily available elsewhere and address a global audience. Although they may offer electronic forums for public exchanges, they tend to focus primarily on the publication of factual information such as historical documents, research data and studies, as well as university course outlines (including that on "Computer applications that can be used in Ministry," http://www.ovc.edu/terry/sylcs290.htm) and topical guides to the Internet (e.g., "Matrics: A scholarly resource for the study of women's religious communities from 400 to 1600 CE," matrix.bc.edu; "Finding God in Cyberspace: A Guide to Religious Studies Resources on the Internet," facultyweb.fontbonne.edu/~jgresham/fgic/; or the "Cybertheology" pages, maintained online as an up-to-date resource for "those interested in the study of theology in cyberspace, theology of cyberspace and theology for cyberspace" (http://www.cybertheology.net). Analytical sites are also formed by electronic journals (e.g., "The Journal of Buddhist Ethics," jbe.gold.ac.uk, a pioneering peer-reviewed e-journal established in 1994) and special issues of online magazines, such as March 1997 "Computer-Mediated Communication Magazine" (http://www.december.com) and September 1999 "Cybersociology" (http://www.cybersociology.com), which discuss emerging connections between religion and the Internet.

TYPES OF ONLINE RELIGIOUS ACTIVITY

Several complementary ways to use the religious Net have evolved to date. The following sections provide an overview of these.

Information Gathering

The first and most important among online religious activities is searching for and consulting networked information. A recent American interview of people who use online religious resources (Pew, 2001) revealed that such "religious surfers" have a great hunger for factual information: 67% of the interviewed people looked for information about their own faith and 50% for information about another faith. Also, people went online to get ideas about the best ways to celebrate religious holidays (22%), to find a new church (14%), or to take a religious course online (3%).

There are several formats of online factual religious information:

1. *Summaries and Frequently Asked Questions* (FAQ) offered by discussion groups (e.g., Alt.Pagan FAQ, http://www.faqs.org/faqs/paganism-faq/, or Shamanism FAQ, http://www.faqs.org/faqs/shamanism/overview).
2. *Tools,* such as Orthodox Judaism's pages that "Find a Chabad Center near you" and "Check the Candle Lighting Times for any Friday or holiday eve—for any location in the world" (http://www.chabad.org); "CyberSalat," a multimedia computer program to teach Islamic Prayer (http://www.ummah.org.uk/software/cyber); "Moon Calculator" (http://www.starlight.demon.co.uk/mooncalc); and "Qibla Calculator" used to establish the correct direction of the holy city of Makkah (http://www.starlight.demon.co.uk/qibla). An example drawn from another tradition is the "Rapture Index," a fundamental Christian prophetic speedometer, which analyses and counts world developments to estimate the likelihood of Judgment Day (http://www.raptureready.com/rap2.html).
3. *Documents and canonical texts,* as well as a body of associated commentaries, sermons, and public statements. For instance, "The Internet Sacred Text Archive" (http://www.sacred-texts.com) covers 53 thematic groups, ranging from African Religions, Alchemy, Americana, through I Ching, Sacred Sexuality, Tolkien [sic], UFOs, Women and Religion, and Zoroastrianism.
4. *Online collections of music and images*. For instance, "The Cyber Hymnal—Dedicated To The Glory Of God" (http://www.cyberhymnal.org) boasts lyrics, scores, and sound files of more than 3,700 Christian hymns and Gospel songs from many denominations, and the "Huntington Archive of Buddhist and Related Art" (http://kaladarshan.arts.ohio-state.edu) has an online collection of tens of thousand of annotated high-resolution research photographs from South and East Asia.
5. *News, announcements, and press releases* (e.g., "Clergy Abuse Tracker," http://www.poynter.org/clergyabuse/ca.htm, or "Religion News Service (RNS)," http://www.religionnews.com).
6. *Corporate details of community and establishment hubs*.
7. *Research analyses and directories of online resources*.
8. *Popular education trivia* (e.g., religious crosswords, comics, and "Quote of The Day" at http://www.buddhanet.net).

Naturally, not all information is in equal demand: Pew (2001) finds, for example, that religious music is downloaded by more people (38%) than are sermons (25%).

Electronic Transactions

According to Pew's survey in 2001 34% of "religious surfers" bought religious items online, 5% played spiritual computer games, and 3% used a faith-oriented matchmaking services. Nevertheless, in the case of commercial activities, their overall frequency and volume seems to be strong enough to support operations of related portal sites.

Online Communication

This is a two-way exchange which can take complementary several forms.

Communication between the faithful is the most common activity. For example, since its establishment in April 1995 a Web site of the Order of St. Benedict in Collegeville, Minnesota (http://www.osb.org) has invited electronic visitors to "send or leave a message in the 'Pilgrims' Parlor'" or join the discussion about *lectio divina* at the "Tabula" (an electronic message board). This is not an isolated case. Throughout the Internet, countless discussions involving tens of thousands of people, co-religionists as well as people representing different traditions, are regularly conducted via e-mail lists, Usenet groups, and Internet relay chat channels. Regular communication among members of online religious communities and groups generates a significant amount of online traffic. The Pew study (2001) finds that no less than 27% of respondents subscribed to a religious listserv and 10% participated in religious chat rooms.

Online ministry (e.g., the multilingual "Pastoral Care on Internet," http://www.seelsorge.net, aimed at the Danish, Dutch, English, Hungarian, Italian, French, and German faithful). The exchanges of this type are sometimes conducted by means of a public-access bulletin board and sometimes via one-to-one e-mail. Here the spiritual leaders help members of their congregation with matters of spiritual and personal well-being. There is also scope for lay people to offer each other counsel and solace. Provision of advice seems to be more common than its receipt. The Pew report (2001) notes that 21% of the respondents sought spiritual guidance via e-mail, although 37% of them had tried to give such advice.

There is a great variety of styles for online ministry. Sometimes the pastoral care is firmly focused on the needs of the faithful, and conducted in an atmosphere of strict privacy (e.g., Chabad-Lubavitch's "Ask the Rabbi" service at http://www.chabad.org or the Zen Mountain Monastery's "Cybermonk" service, http://www.zen-mtn.org, where a senior monastic uses private e-mail to answer privately posed Dharma questions). It can also be conducted in full public view and serve as a vehicle for advertising the depth of religious insight (and compassion) on the part of the counselor (e.g., Question: "How to experiment zazen with a cancer?" Answer: "The nose vertical and the eyes horizontal," http://www.zen-deshimaru.com/EN/teaching/online/online-tous.html).

Other type of contacts also flourish. Miller and Slater (2000) and Bogdanyi (1997) also noted examples of online exchanges between church and society in form of official statements and public relations releases, business and technical contacts, and communication among churches and church-sponsored communication between diasporas and their homeland. To quote a Lutheran minister, "Churches play traditionally a great role in preserving the national identity of minorities. Internet contacts between minority and mother country may be initiated by the churches of both countries which have contacts with one another" (Bogdanyi, 1997).

Online Participation

Participation in computer-mediated worship and ritual is another important type of online activity. This is not in any sense a new development because it simply elaborates on earlier technical arrangements. These include real-time radio and TV broadcasts of religious ceremonies and teachings or telephone-based prayer services (now also on the Web, e.g., http://www.dialaprayer.com). In most of these situations, the Net has became a medium that is fully "transparent" to its skilled users. For example, The Pew study suggests (2001) that those who send and receive e-mail prayers, consider their devotions no different from those who have been aided in their prayers by means of books and leaflets. Networked environments are used to emulate, with varying degrees of realism, a wide range of activities such as:

Live Services

For example, the Monks of Adoration, also known as The Order of St. Augustine (http://www.monksofadoration.org) use RealAudio narrowcasting software to publish an online edition of Chanted Rosary and use webcams to provide a real-time view of their chapel. Similarly, Islami City.com invites visitors to take part in "Ramadan Special! Watch Live Taraweeh and all the other daily prayers from Makkah . . . and much more for only 68 cents a day!" The numbers of virtual participants can be large. In 1996, the first CyberSeder organized by Emanu-El synagogue in New York (http://www.emanuelnyc.org) was joined online by more than 32,000 people (both Jews and non-Jews) in 21 countries. Two years later, in 1998, the religious event drew more than 1 million online participants from 71 countries;

Pilgrimages

HolyLand Network, Christian Information Center, operates "A Virtual Pilgrimage to the Land of Jesus" (http://www.holylandnetwork.com), where readers can "visit" Bethlehem, Nazareth, the Jordan River, Tiberias, and Jerusalem and explore their respective histories, videos, sounds, churches, sites (including Via Dolorosa and the Temple) and, finally, photo galleries (followed by a quick visit to an online shop with "Holy Land Christian and Jewish gifts, and international Christian gifts"). Virtual Islamic and Hindu pilgrimages are also available. For their specific spiritual benefits, see respectively "The Holy Shrine of Imam Rida" (http://www.aqrazavi.org/vrml/english/v001.htm) and "Virtual Puja to Ahmedabad for Lord Ganesha" (http://www.ahmedabadcity.com/Ganeshchaturthi/);

Prayers

Praying is conducted in the form of data packets routed by a computer around a long-distance international network (Glasner, 1999), regular weekly meetings of electronic avatars near the "altar" of a "church" in a three-dimensional virtual reality environment (Schroeder, Heather, & Lee, 1998), or, more commonly, in form of texts posted by readers. For example, in July 2002 http://www.explorefaith.org presented a plea: "Please pray for my family. My husband lost his job 6 months ago when the company he was working for went out of business. He has been unable to find a job in his field and even a lesser paying job. We will lose our house and everything if a door is not opened for a new job. He has been a Network Technician for 10 years. Thanks!"

Confessions

One site, http://www.managingdesire.org, invites readers to "Unburden yourself to our virtual priest. . . . Your confession is completely anonymous, unlike [some of the] test sites in California."

Funerals

The site http://www.online-funeral.com invites family members and friends to "Buy flowers, tell a friend" and to view a real-time live (webcam) Internet recording of a visit to the funeral home. It also provides access to stored video recording of the funeral service at the cemetery, a memorial photo album, and a condolence-message center. In addition to being an efficient conduit for dissemination of information about the death of a friend, the Internet can also be used a special place where the deceased himself is honored. For example, according to Slashdot.com, in early September 2002 a group of online wargamers staged a virtual funeral in honor of a player, nicknamed "Warsinger," who had died of a heart attack. All hostilities were briefly suspended while avatars of his peers from the

"Dark Age of Camelot" wargame "formed a heart on one of the game screens."

Weddings and Ordinations

These and other emotionally and socially potent, although—depending on circumstances—not always legally binding ceremonies take place on the Internet. Several activities of this type are conducted in appropriate multiuser and avatar-populated graphical virtual worlds (for example, http://www.ccon.org/events/wedding.html), whereas others (for example, http://www.spiritualhumanism.org) are more simple, depending on text typed into Web input pages and updates to a participants' database.

Again, the popularity of such facilities varies. The Pew study (2001), for example, found that whereas 38% of respondents had e-mailed prayer requests, only 4% had taken part in online worship. In the next few years, this ratio may dramatically change. Barna Research Online (2001b) suggested that by 2010 approximately 50% of American religious surfers "will seek to have their spiritual experience solely through the Internet, rather than at a church."

Expression of Religious Feelings and Affiliations

Religious expression, the last category of online behaviors, ranges widely. Mostly, it takes the form of collections of personal drawings, photographs, digital collages (e.g., http://www.buddhanet.net/mag_art.htm), computer clipart (e.g., Jewish Clipart Database, http://www.j.co.il), and recordings of original, religiously inspired music and poetry. On the other hand, it can also be succinct, consisting of brief invocations or declarations placed on an opening page of a Web site, for example, a line exclaiming, "In the Name of God, the Most Compassionate, the Most Merciful" (wings.buffalo.edu/sa/muslim) or Zen inspired e-mail signatures (e.g., "Before Enlightenment: chop wood and carry water. After Enlightenment: chop wood and carry water").

THE IMPACT OF THE INTERNET ON ORGANIZED RELIGION

Like other technologies before it, the Internet is far from being a neutral medium. Since the emergence of religious cyberspace, several long-term effects have already become noticeable, especially for institutionalized religions:

Grassroots Rapprochement Among Faiths

The flow of information about religions across tens of thousands of informal publications and venues tends to encourage grassroots rapprochement between members of different denominations and traditions. Such exchanges happen spontaneously and bypass the formal channels of communication between institutions. In consequence, as Brasher (2001, pp. 6–7) noted,

> for all the risks entailed, the wisdom pages and holy hyperlinks that are the stuff of online religion possess the potential to make a unique contribution to global fellowship. . . . As it widens the social foundations of religious life, cyberspace erodes the basis from which religion contributes to the destructive dynamics of xenophobia. In the process, it lessens potential interreligious hatred.

Fragmentation of Locality-Based Communities

Thumma (2002) signals the emergence of the "digital religious divide" between, as well as within, congregations. This is so, because even within well-wired, culturally open, and comparatively affluent North America, the Internet has been adopted more widely in those groups that are (a) large and wealthy (up to 85% uptake), (b) whose leader is better educated, and (c) who represent mainstream nonfundamentalist Christian groups or non-Christian traditions, such as Tibetan or Zen Buddhism. Moreover, even within such technically savvy congregations, 30 to 40% of members may still be unable to get online because they are older, less affluent, or less educated than their "techie" coreligionists. This would suggest that contrary to the hopes of the Benedictine Internet Commission (1998), the Net could, in fact, turn into an additional source of divergence and tensions within faiths.

Shifts in Institutional Power Structures

Although Net-based communications have in general strengthened the triangular web of contacts between religious establishments, community spiritual leaders, and the faithful, such communications are also responsible for subtle changes in the nature and substance of those contacts. Miller and Slater (2000, p. 180) noted, for instance, that the borderless nature of the Net means that high-ranking religious officials may now easily bypass all intermediate echelons and communicate directly with the faithful. By doing so, they may render lower ranking officials redundant or irrelevant. If this is the case, then the centuries-long tradition of careful harmonization of authority and jurisdiction could be disrupted.

In addition, localized microchanges in the distribution of power within religious congregations also start to take place. Some religious leaders have had to compete for the time, attention, and affection of those who stay in electronic touch with teachers who have moved elsewhere, retired, or been dismissed from their religious duties. It has also been observed that people with good information technology skills and the time to work on the community's Web hub tend to get higher social ranking that those without those assets (Thumma, 2002).

Challenges to Orthodoxy and Authority

Unless they are firmly moderated, online religious discussions tend to be undisciplined, impatient, inadequately informed, and repetitive. As Hayes (1999, p. 169) observed, however, all such discussions share a common semantic thread. Although they raise many topics and use many styles, they tend to revolve around key questions of *who* exercises spiritual or moral authority over followers of a given tradition, and *why* this should be so. Simultaneously, online agorae also provide convenient forums for posing anonymous yet passionate questions (see Bunt, 2000a, pp. 115–122) concerning personal religious and ethical issues that are in conflict with the prevailing doctrine (e.g., views on contraception, abortion, suicide, "immoral" sexual practices, "experiments" with other faiths, or consuming "prohibited" foods and substances). In other words, the Internet provides a unique and revolutionary vehicle for the public and ample exploration of themes that traditionally were never voiced nor discussed publicly because of the fear of ridicule or the wrath of enforcers of orthodox thought. The advent of the Internet has changed the rules. As Miller and Slater (2000) observed in the context of Trinidad's online Christian communities "no longer can a church be held back by merely a leader. . . . now anybody who could read, and has semipermanent access to a computer can check and ask questions about doctrine" (2000, p. 181).

The Emergence of Two-Way Communication in Churches

Historically, information in the institutionalized religious traditions flowed from the top down. The Internet and especially its potential for anonymous communication provides hitherto unavailable means for a publicly visible flow of messages from the grassroots level up.

Altered Styles of Spiritual Interaction

While bulletin boards and e-mail bridge physical distance, their use also reduces the number of actual physical meetings. This is a boon in technical contexts; in several religious traditions, however, such as mystical Islam and contemplative Buddhism, frequent one-to-one personal encounters between teacher and disciple are indispensable to the spiritual growth of both parties.

In this sense, the reduction in face-to-face contact is an unwelcome consequence of new communication media. For example, it has been found that the streamlined and often greatly abstracted electronic exchanges between a leader and a person in his or her pastoral care may lead to feelings of ill ease and disenchantment when the two interact face-to-face (Miller & Slater, 2000, p. 183).

In addition, styles of pastoral interaction are also altered by the awareness of the risks of electronic messages being intercepted by, or accidentally forwarded to, the wrong parties or being subpoenaed by courts. These factors make online pastoral communications more circumspect, more deliberate, and more ritualized than they would be in the context of a more private, because undocumented, verbal encounter.

Changes to Institutional Self-Image

As Thumma (2002) observed, posting sermons, newsletters, bulletins, and prayer requests on Web sites or e-mail lists inevitably exposes the inner workings of a congregation to outside scrutiny and comments. As a result, religious organizations with an online presence inevitably develop a particular "digital persona," a stylish self-image, which is then energetically defended and protected for the sake of consistent public relations, even if it is costly or disruptive to the group's internal dynamics. Also, increasingly often, faithful get tempted "to adapt offline religious practices to the tastes and assumptions cultivated by online experience" (Brasher, 2001, p. 23).

CONCLUSION

Gradual Progress and Adjustments

Since its initially hesitant start in the late 1970s, religion online has undergone a succession of cumulative experiments and transformations. As a result, for many people the Internet—an initially confusing and exotic medium—has become a thoroughly familiar and transparent feature of their lives, including their spiritual life.

The focus of one's beliefs is not necessarily realigned, nor are its sincerity and strength are altered through exposure to the Net. The range of opportunities for religious experience, however, is greatly widened, for the Internet provides additional and customized means of access to faith and for its expression. To most of those involved this happens step-by-step and almost imperceptibly.

This is clearly affirmed by a middle-age Buddhist monk:

> after some years of intensive meditation practice and study in Thailand, Burma and Sri Lanka [I] returned home to Australia and established a meditation centre in Sydney about six years ago, without the traditional support. I started to use computers for word processing and simple desktop publishing, then, acquiring a modem started the first bulletin board service (BBS) ever run by a monk, called BuddhaNet. Naturally as the technology developed I moved with it. I progressed to the net. . . , hand-cutting simple HTML code . . . , growing naturally with the new medium. I must confess I was [also] a betatester for Windows 95. . . . [Nowdays] as a teaching monk, I give regular meditation classes and talks during the week—perhaps up to sixty people or more. But on BuddhaNet's web site [now at www.buddhanet.net], there are over 50,000 visitors per day, and a plethora of e-mail inquires on Buddhism as well. (Pannyavaro, 1998)

There is a reason to believe, therefore, that each of the forthcoming technological advances (such as the provision of a high-speed, wireless and completely ubiquitous Internet or of complex immersive cyberarchitectures) will spawn a similar sequence of personal experiments and adjustments, before the new tools become "tamed" and "domesticated." Ultimately, these tools and environments will be made to work completely in the background, exactly the same as electricity, motorcars, and the telephone do today.

Into the Near Future

Designers of postmodern ministry (Couchenour, 2002) already invite us to imagine "High Tech, High Touch, Hi Jesus" religious experience centers. These are hybrid environments, which skillfully combine the most potent elements of real life, virtual reality, and cyberspace.

In such places computer kiosks and wireless hookups will provide access to the best of the Christian Net and online corporate and reference materials. Communication centers, touch-screen art galleries, and multimedia verses from the Scriptures will be tastefully scattered throughout lush indoor gardens. Moreover, strategically placed plasma screens will give "the opportunity for friends to meet and experience a church service or seminar while seated at a table drinking coffee from the Higher Ground coffee shop." Also, The faithful will also have a personal choice of inspirational music ranging "from a single acoustic guitar player to an alternative Christian rock band" while they listen to larger-than-life preachers and life-size holograms of prophets and have their offerings collected via a debit card and validated through biometric technologies.

There is no doubt that from a technical point of view, these are already perfectly feasible projects ("Mega-Church Goes High Tech," 2000). There is also no doubt that religiously minded people, after a period of initial bewilderment and playful experimentation, will make everyday use of them—and that they will do so with social and organizational consequences that will be more lasting, and more pervasive, than the specific technical inventions that triggered them.

GLOSSARY

Agora In ancient Greece, an outdoor place used for public meetings. In the context of the Internet, an online place used for electronic discussion forums.

Avatar In Hindu mythology, the incarnation of a god. Within the language of cyberspace, a schematic yet unique visual representation (e.g., an animal, a human being, a mythological character) of a user operating in a shared virtual environment.

Bulletin board system (BBS) An early form of computerized "meeting" and announcement system that bypassed the Internet by using modems to dial-up select computers directly to upload and download messages and files.

File transfer protocol (FTP) An early, pre–World Wide Web method of moving electronic files between two Internet sites.

Flaming An online exchange of deliberately inflammatory arguments and opinions. "Flame wars," as well as spams and viruses, are commonly used by vandals to wreck normal operation of Usenet discussion groups and unmoderated mailing lists.

Frequently asked questions (FAQ) An electronic document summarizing the main points of a body of knowledge shared by a group of people, such as members of a discussion group.

Internet relay chat (IRC) A type of an online channel of communication that is a large-scale multiuser live chat facility.

Usenet A decentralized worldwide archipelago of electronic discussion areas, called newsgroups. Not all Usenet servers reside on the Internet.

Weblogs ("blogs") Personal Web sites that, similar to journals and diaries, use a time and date format to organize and publish their authors' (bloggers') private reflections, commentaries on current issues, polemics with readers, and other forms of written self-expression.

CROSS REFERENCES

See *Internet Etiquette (Netiquette); Internet Navigation (Basics, Services, and Portals); Internet Relay Chat (IRC); Online Communities.*

REFERENCES

Allucquere, R. S. (1991). Will the real body please stand up? In M. Benedikt (Ed.), *Cyberspace: First Steps* (pp. 81–118). Cambridge, MA: MIT Press. Retrieved January 11, 2003, from http://www.rochester.edu/College/FS/Publications/StoneBody.html

Arinze, F. (2001). *Promoting human values in an era of technology. Message for the end of Ramadan Id Al-Fitr 1422 A.H./2001 A.D.* Vatican: Pontifical Council for Interreligious Dialogue. Retrieved September 30, 2002, from http://www.vatican.va/roman_curia/pontifical_councils/interelg/

Azzi, A. R. (1999). Islam in cyberspace: Muslim presence on the Internet. *Renaissance: A Monthly Islamic Journal, 9*. Retrieved October 7, 2002, from http://www.renaissance.com.pk/julnevi99.html

Barna, G. (2001). *The Cyberchurch: A study by the Barna Institute*. Ventura, CA: Barna Research Group.

Barna Research Online (1998, March). The cyberchurch is coming: National Survey of Teenagers shows expectation of substituting Internet for corner church [press release]. Oxnard, CA: Barna Research Group. Retrieved September 16, 2002, from http://web.archive.org/web/*/www.barna.org/PressCyberChurch.htm

Barna Research Online (2001a, March 29). Pastors paid better, but attendance unchanged [press release]. Ventura, CA: Barna Research Group. Retrieved September 16, 2002, from http://www.barna.org

Barna Research Online (2001b, December 17). The year's most intriguing findings, from Barna Research Studies [press release]. Ventura, CA: Barna Research Group. Retrieved September 16, 2002, from http://www.barna.org

Barna Research Online (2002, January 29). American faith is diverse, as shown among five faith-based segments [press release]. Ventura, CA: Barna Research Group. Retrieved September 16, 2002, from http://www.barna.org/

Bedell, K. (2000). Dispatches from the electronic frontier: Explorations of mainline Protestant uses of the Internet. In J. K. Hadden & D. E. Cowan (Eds.), *Religion on the Internet: Research prospects and promises* (pp. 183–203). New York: JAI Press.

Benedictine Internet Commission (1998, January 15). *Report of the Benedictine Internet Commission, final report*. Retrieved September 17, 2002, from http://www.osb.org/bic/report.html

Bogdanyi, G. (1997, July). The role of Internet in the East European churches today. Presentation for the

2nd European Christian Internet Conference (ECIC 2), Lancaster. Retrieved October 20, 2002, from http://www.lutheran.hu/infochurch/ecic2pap.htm

Brasher, B. E. (2001). *Give me that online religion*. San Francisco: Jossey-Bass.

Bunt, G. (2000a). *Virtually Islamic: Computer-mediated communication and cyber Islamic environments*. Cardiff: University of Wales Press.

Bunt, G. (2000b). Surfing Islam: Ayatollahs, shayks and hajjis on the superhighway. In J. K. Hadden & D. E. Cowan (Eds.), *Religion on the Internet: Research prospects and promises* (pp. 127–151). New York: JAI Press.

Chama, J. C. R. (1996, December 16). Finding God on the Web: Across the Internet, believers are re-examining their ideas of faith, religion and spirituality. *Time, 149,* 52–59. Retrieved September 17, 2002, from http://www.nextscribe.org/background/time/timecover.htm

Church of Scientology International (2000, June 30). Briefing re: The Church of Scientology and the Internet. Retrieved January 3, 2002, from http://religiousmovements.lib.virginia.edu/nrms/scientology_briefing.html

Ciolek, T. M. (2000). Internet, Buddhist and Christian. In W. M. Johnston (Ed.), *Encyclopedia of Monasticism, Volume 1* (pp. 650–655). Chicago: Fitzroy Dearborn Publishers. Retrieved October 6, 2002, from http://www.ciolek.com/PAPERS/fitzroydearborn2000.html

Ciolek, T. M. (2003). Global networking: A timeline. Asia-Pacific Research Online, Canberra. Retrieved January 9, 2003, from http://www.ciolek.com/PAPERS/milestones.html

Couchenour, J. (2002). Practical Ministry Innovations home page. Aristotle Institute. Retrieved October 12, 2002, from http://www.affinitycommerce.com/Aristotle/upload/issue_091002/article_0274.html

Dawson, L. L. (2001, April). Cyberspace and religious life: Conceptualizing the concerns and consequences. Paper presented at the 2001 INFORM, CESNUR, and ISAR Conference "The Spiritual Supermarket: Religious Pluralism in the 21st Century," London. Retrieved October 3, 2002, from http://www.cesnur.org/2001/london2001/dawson.htm

Encyclopaedia Britannica (1987). Religion. In *The New Encyclopaedia Britannica, Volume 9* (15th ed.). Chicago: Author.

Fajardo, A. G. (2002, March). Internet in the documents of the Church. Vatican, Rome. Retrieved January 8, 2003, from http://nuntia.cs.depaul.edu/webmissiology/church%20and%20internet.htm

Glasner, J. (1999, June 21). Spamming God. *Wired News*. Retrieved October 29, 2002, from http://www.wired.com/news/culture/0,1284,20313,00.html

Hadden, J. K., & Cowan, D. E. (Eds.). (2000a). *Religion and the Internet: Research prospects and promises*. New York: JAI Press.

Hadden, J. K., & Cowan, D. E. (2000b). The promised land or electronic chaos? Toward understanding religion on the Internet. In Hadden, J. K., & Cowan, D. E. (Eds.), *Religion on the Internet: Research prospects and promises* (pp. 3–21). New York: JAI Press.

Hayes, R. P. (1999). The Internet as window onto American Buddhism. In D. R. Williams & C. S. Queen (Eds.), *American Buddhism: Methods and findings in recent scholarship* (pp. 168–179). Richmond, UK: Curzon Press.

Helland, C. (2000). Online-religion/religion-online and virtual communitas. In J. K. Hadden & D. E. Cowan (Eds.), *Religion on the Internet: Research prospects and promises* (pp. 205–223). New York: JAI Press.

Herring, D. (n.d.). Cybertheology: Theology in, of and for cyberspace. Retrieved October 20, 2002, from http://www.cybertheology.net/

Internet Software Consortium (2002). Internet domain survey. Retrieved October 24, 2002, from http://www.isc.org/ds/

Kadous, M. W, Br., & Al-Wassiti, A. Br. (1996, October). Islam on the Internet—Internet: Gimmick or tool? Islamic resources & activities on the Internet. *Nida'ul Islam magazine, 15* (Sydney, Australia). Retrieved September 4, 2002, from http://www.islam.org.au/articles/15/INTERNET.HTM

Lochhead, D. (1997). *Shifting realities: Information technology and the Church*. Geneva: World Council of Churches (Risk Book Series).

Mega-church goes high tech to aid worshipers. (2002, October 28). *The Holland Sentinel.* Retrieved October 29, 2002, from http://www.thehollandsentinel.net/stories/102800/rel_32.html

Miller, D., & Slater, D. (2000). *The Internet: An ethnographic approach*. Oxford: Berg.

Pannyavaro, Ven. (1998?). Should monks surf the Internet? *Buddhazine—BuddhaNet's online magazine*. Buddha Dharma Education Association. Retrieved September 12, 2002, from http://www.buddhanet.net/mag_surf.htm

Pannyavaro, Ven. (2002). E-learning Buddhism on the Internet. *Buddhazine—BuddhaNet's online magazine*. Buddha Dharma Education Association. Retrieved September 12, 2002, from http://www.buddhanet.net/gds-speech.htm

Pew Internet and American Life Project (2001, December 23). CyberFaith: How Americans pursue religion online. Retrieved September 23, 2002, from http://www.pewinternet.org/reports/toc.asp?Report=53

Pontifical Council for Social Communications (2002a, May 12). Message of the Holy Father for the 36th World Communications Day Theme: "Internet: A New Forum for Proclaiming the Gospel." Retrieved October 20, 2002, from http://www.vatican.va/holy_father/john_paul_ii/messages/communications/documents/hf_jp-ii_mes_20020122_world-communications-day_en.html

Pontifical Council for Social Communications (2002b, February 22). *The Church and Internet*. Retrieved October 20, 2002, from http://www.vatican.va/roman_curia/pontifical_councils/pccs/documents/rc_pc_pccs_doc_20020228_church-internet_en.html

Rheingold, H. (1994). *The virtual community: Finding connection in a computerized world*. London: Secker & Warburg.

Schroeder, R., Heather, N., & Lee, R. M. (1998). The sacred and the virtual: Religion in multi-user virtual reality.

Journal of Computer Mediated Communication, JCMC 4(2). Retrieved January 10, 2003, from http://www.ascusc.org/jcmc/vol4/issue2/schroeder.html

Thumma, S. (2002, April 18). Religion and the Internet. Communications forum lecture at the Massachusetts Institute of Technology. Retrieved September 19, 2002, from http://hirr.hartsem.edu/bookshelf/thumma_article6.html

U.S. Department of State (2002, May 6). *Burma (Myanmar)*. Consular information sheet. Retrieved October 12, 2002, from http://travel.state.gov/burma.html

Zaleski, J. P. (1997). *The soul of cyberspace: How new technology is changing our spiritual lives*. New York: HarperCollins.

Zittrain, J., & Edelman, B. (2002). Documentation of Internet filtering worldwide. Berkman Center for Internet & Society, Harvard Law School. Retrieved October 20, 2002, from http://cyber.law.harvard.edu/filtering/

Online Stalking

David J. Loundy, *DePaul University*

WHAT IS ONLINE STALKING?

Definition

What is online stalking, often referred to as "cyberstalking"? The answer depends on whom you ask—there is no definitive definition. The term is usually defined based on an analogy to the crime of traditional stalking. Traditional stalking involves a form of repeated harassment that generally involves following the victim. It is harassment that leaves the victim with the fear that he or she will be physically harmed. Because "stalking" is an emotionally charged term, and one that often has specific legal implications, it is important to define the term properly.

"Cyberstalking" does not refer to annoying e-mail. It does not apply to irritating instant messages. It does not refer to defamatory message-board posts. It does not refer to identity theft. All of these types of behavior, however, may fit into an overall pattern of conduct exhibited by stalkers. Online stalking is conduct similar to traditional stalking but carried out online. For instance, sending e-mail messages detailing the recipient's day-to-day activities and implying or threatening harm to the recipient may be a case of online stalking.

Various countries and states may have laws against stalking specifically or against various actions that may be part of a stalker's course of conduct. Service providers also may be willing to help customers who are being stalked online, or they may be willing to terminate service to their own customers who are stalking others. Conduct that is merely irritating, however, may simply be seen as a natural part of life, and complaints about such conduct often will not be taken seriously by the law or a service provider.

What Is "Traditional" Stalking?

"Stalking" as a distinctly defined crime is fairly new, first appearing in the statute books in the early 1990s. It generally involves repeated contact with the victim—contact that makes the victim fear for his or her physical safety. An example of a stalking statute (California Civil Code, 2001) provides the following:

> (a) A person is liable for the tort of stalking when the plaintiff proves all of the following elements of the tort:
>
> (1) The defendant engaged in a pattern of conduct the intent of which was to follow, alarm, or harass the plaintiff. In order to establish this element, the plaintiff shall be required to support his or her allegations with independent corroborating evidence.
>
> (2) As a result of that pattern of conduct, the plaintiff reasonably feared for his or her safety, or the safety of an immediate family member. For purposes of this paragraph, "immediate family" means a spouse, parent, child, any person related by consanguinity or affinity within the second degree, or any person who regularly resides, or, within the six months preceding any portion of the pattern of conduct, regularly resided, in the plaintiff's household.
>
> (3) One of the following:
>
> (A) The defendant, as a part of the pattern of conduct specified in paragraph (1), made a credible threat with the intent to place the plaintiff in reasonable fear for his or her safety, or the safety of an immediate family member and, on at least one occasion, the plaintiff clearly and definitively demanded that the defendant cease and abate his or her pattern of conduct and the defendant persisted in his or her pattern of conduct.
>
> (B) The defendant violated a restraining order, including, but not limited to, any order issued pursuant to Section 527.6 of the Code of Civil Procedure, prohibiting any act described in subdivision (a).

Stalkers are generally motivated by a desire to control their victims. Stalking is seen as a problem because it implies an intentional and concerted effort to place the victim in fear of bodily harm, rather than mere casual but unwelcome contact. Thus, accidentally running into a former boyfriend at a restaurant may be an annoyance but such an encounter by itself would not constitute stalking.

On the other hand, running into a former boyfriend every night in the parking lot of a different restaurant, as if he is following the woman around waiting in the dark, could indicate conduct venturing into the realm of stalking. Stalking laws generally will not provide a remedy for the accidental encounter, but they will assist where there is a pattern of activity intended to instill a sense of fear in the victim. (Check with local legal authorities to determine what conduct is prohibited and what remedies are available in your local area.)

How Does Online Stalking Differ from Traditional Stalking?

Online stalking is a form of harassment, but it is more pointed, as is traditional stalking. It occurs through the use of a computer, and it may have more specific motivations than other forms of harassment. One way in which traditional stalking differs from online stalking is in some jurisdictions' specific requirement that the stalker "follow" the victim. How do you follow someone around in cyberspace? Such a spatial requirement may preclude the application of a specific stalking statute to online-only conduct. Second, because stalking statutes generally require a fear of bodily harm, it is less likely that an e-mail message will convey the same fear that a personal encounter would create. With an e-mail message, there is no way to know if the sender is even on the same continent as the recipient, much less a legitimate threat. Online stalking may lack the immediacy that is present with "real-world" encounters. For electronically received communications, a threat of harm may be more remote or less reasonable than, for example, a personal encounter. Additionally, online stalking adds a level of anonymity that may be missing in a real-life encounter. This anonymity also may have the side effect of emboldening stalkers who are more willing to attack or harass than they would be if their conduct was more readily traceable.

Why Does the Definition Matter?

Harassing conduct may be annoying, but it is conduct for which there are only limited remedies available under the law. Although the extent or availability of any remedy varies depending on applicable laws in your country or state of residence, as a general rule you have no right to be protected from being annoyed—especially when the source of your annoyance is someone else's protected speech. Stalking, on the other hand, may be a criminal offense in many jurisdictions for which there are specific legal remedies. The law does not punish the speech per se, but punishes the use of speech as a weapon intended to cause harm in much the same way as hate speech or defamation may be outlawed because of the harm it causes. Stalking is essentially seen as a form of assault rather than an attempt to communicate.

The law is slow to keep up in many areas with changing technology, including in the area of addressing aspects of online harassment. Traditional stalking law may not help a victim if the online conduct does not fit within a particular statute's definition of stalking. Thus a statutory definition of stalking affects whether a particular stalking law applies to the unwelcome conduct. An example of an online-specific law is Illinois' Cyberstalking (2001) law, which reads as follows:

> Sec. 12–7.5. Cyberstalking.
>
> (a) A person commits cyberstalking when he or she, knowingly and without lawful justification, on at least 2 separate occasions, harasses another person through the use of electronic communication and:
>
> (1) at any time transmits a threat of immediate or future bodily harm, sexual assault, confinement, or restraint and the threat is directed towards that person or a family member of that person, or
>
> (2) places that person or a family member of that person in reasonable apprehension of immediate or future bodily harm, sexual assault, confinement, or restraint.
>
> (b) As used in this Section: "Harass" means to engage in a knowing and willful course of conduct directed at a specific person that alarms, torments, or terrorizes that person. "Electronic communication" means any transfer of signs, signals, writings, sounds, data, or intelligence of any nature transmitted in whole or in part by a wire, radio, electromagnetic, photoelectric, or photo-optical system. "Electronic communication" includes transmissions by a computer through the Internet to another computer.
>
> (c) Sentence. Cyberstalking is a Class 4 felony. A second or subsequent conviction for cyberstalking is a Class 3 felony.

However, even in the absence of a specific online stalking law, there may be other less obvious laws that provide a remedy, as well as technical options that also provide an adequate remedy to online harassment.

Defining conduct as stalking is also important because it provides a means to frame the discussion. Stalking is a concept that many people have encountered, at least via the media, and can therefore understand. By framing the online conduct in terms of traditional stalking, it may produce greater sensitivity to the effects of the unwelcome conduct. Merely calling the police and telling them that you are "receiving unwelcome instant messages" may be greeted with the suggestion that you simply turn off your computer. Obviously, such a dismissive reaction will not be of much comfort to a victim, because it shows no acknowledgment of the real harm that can result from a stalker's interests. If the police do not understand what an instant message is, at least they can be educated to the point that they understand the message sender could be a legitimate stalker intent on causing the same harm that any other stalker sending the same message via a more traditional medium could intend.

Who Is a Stalker?

A 2001 unscientific study done by Working to Halt Online Abuse (WHOA), an organization started by a former online stalking victim, describes the profile of an online stalker as follows:

- 83.58% of cases reported to WHOA involved female victims, although there is no indication if women are harassed more frequently or are instead more likely to seek help with harassment.
- 44.53% of people who contacted WHOA for help reported their ages as between 18 and 30.
- 63.88% of cases reported to WHOA involved harassment by a male, 29.23% of the harassers were female, and 6.9% were of unknown gender.
- 46.31% of those reporting to WHOA had some previous contact with their harassers, 49.26% of victims reported no prior relationship or contact with their harassers, 4.43% were unsure of the identity of their harassers.
- 39.57% of all online harassment (as distinguished from other contact) began via e-mail.
- 26.44% percent of online stalking cases reported to WHOA included some form of offline harassment as well.

The study is unscientific because the responses are self-selected responses of those who decided to contact the organization. In the absence of scientifically valid survey data, however, it provides useful anecdotal information that helps define the problem.

What motivates a stalker? As discussed by Howard (1999) in *Cyber-Stalking: Obsessional Pursuit and the Digital Criminal,* there are four basic types of cyberstalkers:

Simple Obsessional: The largest category; typically involves a victim and a perpetrator who have a prior relationship. This group also poses the biggest threat to the victim. The motivation behind the stalking is often to restart a relationship—or seek revenge for the ending of a relationship—through the inducement of fear.

Love Obsessional: Such stalkers generally have no prior relationship with the victim. Victims are often encountered through the media or the Internet. A large percentage of such stalkers may be suffering from a mental disorder. A typical example of a love obsessional stalker is the "obsessed fan" of a celebrity.

Erotomanic: Similar to the love obesessional category, these stalkers go a step further and possess the delusion that their target is in love with them.

False Victimization Syndrome: This group accuses another person of stalking, either real or imaginary, to foster sympathy and support from those around them.

HOW MUCH OF A PROBLEM IS ONLINE STALKING?

Online stalking earns media attention, but how serious is the problem? This question is open to debate. A 1999 report prepared by the U.S. Department of Justice (1999), and later reports that merely respun the same content, stated that the problem is pervasive and, using "back of the envelope" calculations, theorized that online stalking could be a crime with tens or hundreds of thousands of victims. The report, however, is sparse in actual support for these claims. In fact, it even cites to a study conducted at the University of Cincinnati, of which the authors have stated does not measure the statistics that the Department of Justice cites the study to support (Koch, 2000). Essentially, there are no empirical scientific studies or data as to the scope of the problem.

Although the media, legislatures, and other interested groups may provide more than a fair amount of hype for the topic, online stalking and other forms of online harassment are a legitimate concern, especially for victims. As more and more daily activities move online, a statistical analysis isn't required to see that more harassing conduct likely will move online as well—an assumption supported at least by anecdotal evidence.

Examples of Online Stalking

Andrew Archambeau has the distinction of being the first person convicted of "cyberstalking," although his actions extended into the "real world" as well (Eckenweiler, 1996). The real-world aspects of his conduct are rarely mentioned in discussions of online stalking, however; only the fact that he harassed his victim by e-mail is discussed at any length. Archambeau had met a woman through a video dating service in early 1994, and they went out on two dates. Apparently he thought more of the relationship than she did, because she sent him an e-mail message saying that she did not want to see him again. Over the next few weeks, he sent her approximately 20 e-mail messages and also left her telephone messages (including one saying, "I stalked you for the first time today"). Requests to leave her alone were ignored. Unfortunately for him, he lived across the street from the school where she worked, thus making it appear that he was waiting outside her workplace. After he ignored a police warning to leave her alone, criminal charges were filed against him for violating Michigan's stalking law. After mounting an unsuccessful challenge to the statute's constitutionality, he finally pleaded no contest (Ellison & Akdeniz, 1998). Archambeau's conduct was obviously objectionable, and even absent the e-mail messages sent to his victim, he may have run afoul of the stalking law because of his real-world actions.

Another incident that had a much stronger online component, yet still included a variety of real-world contact, involved two law students at the University of Dayton who had begun dating. The woman decided to end the relationship, but the man wasn't willing to accept this proposition. Over winter break, he began sending his former love interest notes about how he was starving himself to death, including the details of the pain he was to suffer in the process. The two reconciled—briefly. After recovering from a suicide attempt provoked when the woman broke off the relationship a second time, Mr. Davis, the former boyfriend, sent the woman numerous e-mail messages stating that he had been researching her hometown and regularly spending time in a park near her apartment. The notes did not contain explicit threats of harm, but Davis's tone "fluctuated between despair over the break-up, anger, threats to commit suicide, a desire to see [the woman] in pain, and blaming [her] for ruining Davis's life" (*Dayton v. Davis,* 1999). Davis also included in his e-mail messages information that lead to the belief that he had been watching the woman, such as knowing what she had watched on television, as well as a link to the Web site he had created. Davis's Web site "portrayed, among other things, the image of [her] head transforming into a skull amidst flames, dripping blood, and charging horses ridden by robed skeletons. Interspersed with these images were quotations from the Bible and other sources in which the common theme was love, death, and destruction. On another Web page, Davis had posted pictures of [her] home town . . . although when questioned . . . Davis denied ever having been to the town" (*Dayton v. Davis*). Davis was charged by the authorities with violating the Ohio State law against menacing by stalking (R.C. 2903.211) and aggravated menacing [R.C.G.O. 135.05(A)]. He defended himself by arguing that he had never actually threatened the woman. The court found that he had still succeeded in knowingly placing her in fear of serious physical harm, as evidenced in part by her moving out of her apartment and eventually transferring schools. Intentionally creating such fear, rather than making threats, is what the statutes prohibited. Davis was convicted and sentenced to 180 days in detention.

The first person convicted under the California state online stalking statute was Gary S. Dellapenta (Miller & Maharaj, 1999), a 50-year-old security guard who decided to generate some unwanted notoriety for his former girlfriend. In this case, the stalker's conduct was purely online, although it generated real-world consequences. Dellapenta posted messages on America Online and sent e-mail messages purportedly from the woman stating that she had "rape fantasies"—and soliciting assistance in living out those fantasies. The messages included her home telephone number, address, and instructions on how to disable her home security system. Six people decided to take up "her" offer and dropped by her apartment (she was not physically harmed by these visitors). Criminal charges were brought against Dellapenta for stalking, computer fraud, and solicitation of sexual assault, and he was sentenced to six years of incarceration (Brenner, 2001). A similar false impersonation case in Korea that involved the culprit forging offers of sexual services in the name of the victim resulted in a warrant for the arrest of a man on criminal slander charges (Soh-Jung, 2001).

People v. Kochanowski (2000), involved a man who asked a coworker to help set up a Web site. The site contained suggestive photographs of his former girlfriend, along with her address and telephone numbers. The page also contained what the court described as "express references to intimate body parts and attributed to [the girlfriend] an infatuation with sex" (*People v. Kochanowski,* p. 462). The woman then began receiving disturbing calls at work. The court found that the former boyfriend was guilty of violating the New York statute prohibiting "aggravated harassment" (New York State Penal Law §240.30(1)) that prohibited certain intentionally annoying or alarming communications. Kochanowski argued in his defense that he had not made the alarming calls, which would have been a violation of a protective order the woman had obtained prohibiting the man from communicating with her. The court found it sufficient under the aggravated harassment statute, however, that he had intentionally caused such alarming communications to be made by others (although the court did not find the specific terms of the protection order violated by Kochanowski's Web site).

Other examples demonstrate the international reach of online harassment—it may be just as easy to stalk someone on the other side of the planet as it is to stalk someone up the street. In the case of Rhonda Bartle, international online harassment did lead to an attempted real-world encounter (Hubbard, 2001). Bartle, a New Zealand author, started exchanging e-mail messages with Peggy Phillips, an American writer living in California. When the 84-year-old Phillips showed that she was interested in a deeper relationship than Ms. Bartle wanted, Ms. Bartle decided to end the relationship. Ms. Phillips commented about feeling suicidal and then took a plane to New Zealand. Bartle told the local taxi company not to take anyone to her house. When Phillips attempted to order a cab, the taxi company called the police, who then served Phillips with a trespass notice. After returning home, Phillips continued sending her unwanted e-mail messages (which, for the most part, were automatically filtered out of Bartle's e-mail). Because New Zealand did not have online stalking laws at the time, Bartle contacted the Orange County, California, police in an effort to have them enforce the California law against Phillips. (Unfortunately, news accounts did not state whether this proved productive.) Australian courts have also been faced with a jurisdictional fight over where an online stalking case can be heard against an Australian man who stalked a Canadian actress by e-mail (Cant, 2001).

WHAT CAN YOU DO IF YOU ARE A VICTIM?

According to information provided to members of the CyberAngels, who may assist victims of online stalking,

> The victim is often embarrassed and does not seek help till the situation becomes out of hand. Why do you think this happens? They often know the stalker in some way. May even have had a relationship with them to some degree. Maybe online boyfriend and girlfriend. Maybe the victim had sought out the person first. Maybe they have given personal identifying information that has come back to haunt them. So they are embarrassed and ashamed and don't know what to do. The victim naturally will try to reason with the stalker, to get them to back off. It rarely works. Any attention given to a stalker is still attention and empowers them. If they can't have your love they will take your hate, anger and fear. (CyberAngels, n.d.)

In other words, asking a stalker to stop is not likely to be effective—although it is the first place to start. One of the most important things a victim can do is to document the situation. Save e-mail messages. Capture chat sessions. Document everything. Although the details on how to capture this information are beyond the scope of this article, the need is clear. This information is what allows you to make a case—be it with a service provider, law enforcement, or in court. Beyond this basic record keeping, there are a few places to seek help.

Service Providers and Technical Fixes

One remedy that may be available is to have the Internet access of a stalker eliminated. According to WHOA's statistics, 35.87% of all of the cases reported to the organization were resolved after complaints to the sender's Internet service provider (ISP; WHOA, 2001). The vast majority of ISPs have some sort of "terms of service" agreement or "acceptable use policy." These are generally contracts with the service providers and their users that restrict the use of the service providers' systems. Although these policies generally do not create a legal remedy on behalf of a victim, they do provide a means for the service provider to terminate someone who is abusing people via the service provider's system. A service provider is often willing to terminate "problem users" because it does not want to be seen as contributing to the continuation of abusive activities, it actively wants such conduct removed from the Internet, or it simply does not want the hassle of dealing with a user who generates complaints and the possibility of (expensive) legal hassles. As a result, if a victim can trace a stalker back to his or her Internet service provider, it is always sensible to look at the service provider's Web site to see if the provider has an acceptable use policy or terms of service agreement that is being violated by the stalker's actions. This policy then can be brought to the service provider's attention when describing the actions of the provider's user. In the course of dealing with a stalker, however, although this line of action may be effective, having his or her Internet access shut down is likely to provoke anger, possibly producing more harmful behavior in the end (either through a different service provider or through real-world contacts) rather than eliminating the threat altogether.

Tracing the source of online harassment may be straightforward, or it may be almost impossible. A harasser may make no effort to hide his or her identity or the service provider being used to originate harassing messages. In such a case, contacting a service provider is a fairly simple process. In addition to looking to see if the provider has a Web site with contact information, many providers maintain an e-mail account intended for abuse complaints—generally in the form of "abuse@[serviceprovider.com]" or the like. In addition, the domain name registration for a service provider will usually have contact information. Domain name registration information can be checked from a Web site such as http://www.allwhois.com/.

It may not be easy to track the source of harassing communications because of the ease of anonymously communicating on the Internet. In addition to services that specifically provide for anonymous communication by stripping identifying information from e-mail messages and other public posts, it is also possible to forge routing information. Although tracing such forgeries is beyond the scope of this article and techniques change with advances in the technology involved, good resources can be found online. (For e-mail, see the SPAM-L FAQ in the Further Reading section.) Of course, even if you can trace a message back to its source, it may be the case that the originating account was opened with false information.

If one can identify a service provider but not the real name of the user, the provider is not likely to be helpful in turning over its user's identity. Often a service provider will insist on some sort of subpoena or court order before turning over identity or contact information about one of its users, even in a case where it is willing to cancel that same user's account. (Some jurisdictions make it relatively easy to file a suit or otherwise compel the release of information possessed by an ISP that identifies a harasser. Some jurisdictions will allow the filing of a "John Doe" lawsuit for the purpose of discovering who the true target of the suit should be.)

Without resorting to help from a service provider, there are some steps that victims can take for themselves. The most obvious is to avoid unwanted contact. Delete or "kill file" messages from the stalker, either manually or through an automatic filter that most popular e-mail programs and newsreaders allow users to establish. Avoid logging on to chat rooms at times when the stalker is likely to be logged on as well.

The Law

If you are being stalked, a typical response is to call the police. For traditional stalking, this may result in gaining the assistance one needs to stop the stalker's conduct. According to WHOA's (2001) statistics, 14.91% of all of the cases reported to the organization were addressed by referring the matter to law enforcement. Unfortunately, in the case of online stalking, the response from the police may not be as useful. Many people who call the police are unable to obtain redress for their problem. As the Department of Justice (1999) report found, training of law enforcement officers to handle online stalking is erratic and, in large measure, insufficient. Depending on location, the type of law enforcement to contact may vary—for instance, are state or federal authorities the best resource? In the United States, a victim could try contacting the local or state police, the Federal Bureau of Investigation, or the Secret Service—all of which may handle some types of online crimes. The amount of coordination and awareness between law enforcement groups appears to be growing. The level of technical sophistication and resources devoted to computer crime–related issues is also improving. In some cases, such improvements are required by legislation. Often the agency contacted may not know how to help but will know to whom the victim should be referred. Victims who do not get a knowledgeable response may want to try a different law enforcement entity.

Another concern with contacting law enforcement is a matter of resource allocation. Simply put, law enforcement resources are limited, and there may not be anyone capable of taking time to help a victim of online harassment if other matters that affect more than a single victim or that appear likely to result in more immediate harm are consuming available resources. There may also be a lack of investigative tools with which to pursue a matter even if there is sufficient interest in the case.

Finally, contacting law enforcement may result in either a very public or a very private investigation. The fact that a police report has been filed, for instance, is a matter of public record in many jurisdictions and may attract media attention. On the other hand, the details of an investigation undertaken by law enforcement may be confidential, even from the victim who reported the incident and stands to benefit from the results of any investigation. An investigation by law enforcement is generally out of the hands of the victim because it becomes the government's rather than the victim's case.

Another means of using the law to provide a remedy is through the hiring of a private attorney. The ability of an attorney to provide assistance will be dictated by where the victim and the stalker are located. This will determine, in part, the applicable laws, which will determine if a "private cause of action" exists. In other words, an individual generally cannot sue someone for violating a criminal law. Generally, only the government can sue someone to enforce the criminal law. Some statutes do, however, provide a private cause of action or some other civil remedy that will allow a victim to go after a stalker directly, without assistance from the government. The results of such a suit could include injunctive relief which, for example, would require the stalker to leave the victim alone, or provide for recovery of monetary damages.

What kinds of laws are there? As mentioned, traditional stalking laws may apply to cyberstalking cases. These are generally criminal statutes that require law enforcement assistance. These statutes generally require multiple incidents of harassment that cause the victim to fear for his or her physical safety. Sometimes the safety of family members is also covered under these statutes. The statute may also dictate a specific behavioral requirement, such as the stalker's having physically followed the victim. Some jurisdictions have expanded or clarified their laws in recent years to provide a remedy specific to online stalking. These laws may acknowledge certain harmful uses of technology in defining the offense or merely serve to remove obstacles contained in traditional stalking statutes containing physical requirements that do not apply in the online world.

Some jurisdictions address stalking-like behavior with traditional harassment law or laws prohibiting intimidation. Harassment and intimidation laws may cover a broader range of conduct than traditional stalking laws, but they may have other sorts of obstacles to overcome to ensure that only egregious conduct is prohibited as unlawful harassment. Hurdles, for example, may include requirements that a harasser has the intent to cause harm, or they may require a certain level of damage before providing a remedy. These statutes may be technology-dependent, such as harassing telephone call statutes, although use of modems to connect to the Internet may fit within the statutes' coverage.

If messages from a stalker contain actual threats, many jurisdictions have statutes that provide a remedy for the transmission of credible threats. One famous online "stalking" case, the "Jake Baker" case (*U.S. v. Alkhabaz,* 1997), involved a student at the University of Michigan who posted to Usenet news a piece of "erotic fiction" describing the sexual torture of a woman, a woman who was given the name of one of his classmates. The man was arrested and initially held in jail as a threat to society pending a psychiatric evaluation. The court held that the man's actions were not criminal, because no evidence was presented that his actions were anything more than a sick fantasy. Because there was no evidence that he would really act out his fantasy, there was no credible threat, and thus the federal statute at issue was not violated. Some jurisdictions may have different standards for threats made to certain types of people, such as the U.S. prohibition of threats made against the president (18 U.S.C. §871).

Statutes that address specific conduct such as identity theft, false attribution of origin of the messages, or eavesdropping may also come to bear in a cyberstalking case. Some "common law" concepts—traditional legal concepts that have evolved through court decisions and for which there may be no statute—may also provide a mechanism for legally attacking a stalker. For instance, assault (where a victim is placed in fear of bodily injury); intentional infliction of emotional distress; or, depending on the particular actions, defamation, trespass, or fraud-type arguments may also allow for a remedy. Depending on the law of a particular jurisdiction, "family law" remedies such as restraining orders and orders of protection, often aimed at keeping away ex-spouses or love interests, may be relevant.

CONCLUSION

Online stalking is a problem of unknown proportions. Just as traditional stalking is a concern for its victims, however, so, too is the online equivalent. As more people live more of their lives online, all forms of online crime are likely to increase. Although this will produce more claims of online stalking and other forms of online harassment, it will also force law enforcement to be more prepared to address the needs of victims. In the case of legislators, they will also be required to ensure that the laws have evolved to address the concerns of victims—without having undue impact on legitimate, but perhaps heated, interactions. Because of the international nature of the Internet, cooperation between governments will be essential to address foreign harassers. Efforts are already being put into place to aid in the international enforcement of criminal laws, as evidenced by the Council of Europe's Cybercrime Convention (Howard, 1999) that requires all signing countries to outlaw certain types of objectionable conduct and to provide international assistance in enforcing other countries' laws that prohibit this minimum level of criminal conduct. It is reasonable to expect that the more egregious forms of online stalking are likely to be sufficient to result in some degree of international cooperation.

GLOSSARY

Acceptable use policy An Internet service provider's rules describing what one of its customers may and may not do with or through the provider's computer system.

Criminal lawsuit A lawsuit brought by the government against an individual. Unlike a private or civil lawsuit, a criminal lawsuit could involve jail time or other more serious sanctions depending on what remedies are contained in the applicable law.

Cyberstalking Another name for online stalking (see below).

Harass To engage in an intentional course of conduct directed at a specific person that alarms, torments, or terrorizes that person.

Injunctive relief Relief granted by a court to a victim that generally orders another person not to do something, such as ordering a stalker to stay away from a victim, or to do something, such as remove objectionable material from a Web site.

John Doe lawsuits Lawsuits filed against an unknown person. Once the lawsuit is filed, the party filing the suit can try to obtain more information on the person being sued.

Kill files Filters in many e-mail programs that are used to block messages sent from people or addresses listed in the kill file (also called bozo filters). These are used so that e-mail or other electronic communications from a harasser can be deleted by the victim before they appear in the inbox.

Online stalking A form of computer-mediated harassment analogous to traditional stalking.

Private cause of action or **private right of action** The ability to sue a person engaging in objectionable conduct without the need for government involvement to enforce the law.

Restraining orders Orders issued by a judge that may prohibit someone from contacting another person. What these orders may cover varies by jurisdiction.

Stalking A form of harassment generally targeted at a specific individual that causes fear of physical harm. As a legal term, this definition varies by jurisdiction.

CROSS REFERENCES

See *Cybercrime and Cyberfraud; Cyberlaw: The Major Areas, Development, and Provisions; Legal, Social and Ethical Issues; Privacy Law.*

REFERENCES

Brenner, S. (2001, June). "Cybercrime investigation and prosecution: The role of penal and procedural law. *E Law—Murdoch University Electronic Journal of Law, 8.* Retrieved March 23, 2003, from http://www.murdoch.edu.au/elaw/issues/v8n2/brenner82text.html California Civil Code §1708.7—Stalking (2001).

Cant, S. (2001, March 27). Courts wrangle over cyberstalking. *The Age,* p. 3.

Council of Europe Convention on Cybercrime (2001, November 23). Retrieved March 23, 2003, from http://conventions.coe.int/Treaty/en/Treaties/Html/185.htm

Cyberstalking [Illinois Criminal Statute] 720 ILCS 5/12–7.5 (2001).

CyberAngels (n.d.). Retrieved prior to June 1, 2002, from the CyberAngels Web site (no longer available).

Dayton v. Davis, 735 N.E.2d 939 (Ohio App. 2 Dist., Nov. 24, 1999).

Eckenweiler, M. (1996, February 1). *NetGuide,* p. 35.

Ellison, L., & Akdeniz, Y. (1998, December). Cyberstalking: The regulation of harassment on the internet. *Criminal Law Review* [Special Edition: *Crime, criminal justice and the Internet*], 29–48.

Howard, C. (1999). *Cyber-stalking: Obsessional pursuit and the digital criminal. Stalking typologies and pathologies*. Retrieved March 23, 2003, from http://www.crimelibrary.com/criminal_mind/psychology/cyberstalking/3.html?sect=19

Hubbard, A. (2001, June 24). When truth is stranger than fiction. *Sunday Star-Times,* p. A7.

Humphreys, L. (2001, June 20). NP author won't give in to stalker. *Daily News* (New Zealand).

Koch, L. (2000, May 25). Cyberstalking hype. *Interactive Week.*

Miller, G., & Maharaj, D. (1999, January 22). N. Hollywood man charged in 1st cyberstalking case. *Los Angeles Times.*

People v. Kochanowski, 719 N.Y.S.2d 461 (Sup. Ct. App. Term, NY, Oct. 18, 2000).

Soh-Jung, Y. (2001, July 6). Internet stalker punishable for libel. *The Korea Herald*.

U.S. v. Alkhabaz, 104 F.3d 1492 (6th Cir. 1997).

U.S. Department of Justice. (1999, August). *Report on cyberstalking: A new challenge for law enforcement and industry.* A report from the Attorney General to the Vice President. Retrieved March 23, 2003, from http://www.usdoj.gov/criminal/cybercrime/cyberstalking.htm

Working to Halt Online Abuse. (2001). http://www.haltabuse.org/resources/index.shtml

FURTHER READING

CyberAngels (Initially a project of the Guardian Angels neighborhood watch organization, on June 1, 2002, the group splintered and divided into the CyberAngels and WiredPatrol.org): http://www.CyberAngels.org/

1999 Report on Cyberstalking: A New Challenge for Law Enforcement and Industry: http://www.usdoj.gov/criminal/cybercrime/cyberstalking.htm

Findlaw Cyberstalking and E-Mail Threats Resource Page: http://cyber.findlaw.com/criminal/cyberstalk.html

Working to Halt Online Abuse: http://www.haltabuse.org/

Lucke, K., Reading E-Mail Headers: http://www.stopspam.org/email/headers/headers.html

The Stalking Assistance Site: http://www.stalkingassistance.com/index.htm

The SPAM-L FAQ: http://www.claws-and-paws.com/spam-l/tracking.html

WiredPatrol.org Cyberstalking Index: http://www.wiredpatrol.org/stalking/index.html

Open Source Development and Licensing

Steven J. Henry, *Wolf, Greenfield & Sacks, P.C.*

OPEN SOURCE: WHAT IS IT? WHAT IS IT NOT?

Computer programs are now accepted in most countries as copyrightable works and are explicitly so recognized in Section 106 of the U.S. Copyright Act, Title 17, U.S. Code, Sections 101-810. Actually, computer programs have been recognized as "literary works" under the Copyright Act since at least the 1970s. In 1978, it was decided by a Congressional study (the report by CONTU, the Commission on New Technological Uses of Copyrighted Works) that no amendment of the Act was needed to cover programs per se (CONTU, 1978). In most other countries, copyright is considered a "natural" right of the author and (unlike the United States) no registration is required, but statutes and treaties do exist to circumscribe the nature of copyrightable works (some expressly naming programs, others accepting them under a broader category, as in the United States), the rights encompassed by copyright, the term of protection, and so on. In many important respects, these laws are similar to those in the United States, and no distinction is necessary within the context of this chapter. As such, the law provides that the "author" of a program (be that the programmer or his or her employer) has a certain bundle of rights. Among these are the exclusive right to copy, distribute copies of, and prepare derivative works from the copyrighted program code. In addition to the rights that copyright law provides, software authors who commercialize their work product often wish to condition their grant of rights in the distributed copies and to establish contractual terms with their customers, such as warranty terms, covenants to keep the software secret, support and maintenance terms, and so forth. The grant of rights can be established in a unilateral declaration wherein the owner (author) states the scope of permission it is granting. This is called a "license." A license gives the receiver (grantee) permission to do something that otherwise would subject it to liability as an infringer. The practice most commonly employed is to include the grant within a contract, usually called a license agreement, which spells out the grant and additional rights, promises, and obligations between the customer (licensee) and author (licensor). Thus, when one acquires a copy of a program, the transaction may appear to be a sale (i.e., one party makes a payment and the other party delivers a CD in a box), but it normally is (a) a license to use the program subject to certain promises and conditions and (b) sale of the medium (i.e., the CD). The "buyer" gets only the rights spelled out in the license grant (which is displayed on the package or provided on screen as part of the installation process). Most non–open source software licenses prohibit copying and modification of the software.

Keeping in mind that the distinction between a license grant (unilateral from the grantor) and an agreement (to which two or more parties bind themselves) is important. A licensee cannot give to a downstream party more rights than it acquired in the license grant. Consumer software is normally licensed with a restrictive grant that permits use but not copying or redistribution, so even without an enforceable contract with the licensee, the licensor expects to be able to enforce the grant limitations. The case law is mixed as to whether the conventionally used "shrinkwrap" and "clickwrap" documents styled as license agreements between the copyright owner and the customer are, indeed, enforceable contracts. Nonetheless, the licensor may maintain that it can impose unilateral restrictions on the license granted even without securing express "agreement" from the customer. If, however, the transaction is considered a sale, rather than a license, the copyright owner cannot, at least in the United States under the "first sale" doctrine of copyright law, control subsequent use or resale of the copy sold (although copying of the copy is prohibited as copyright infringement). Care is needed in drafting and implementing license terms.

"Open source" licensing (OSL) is a distinctly contrasting set of licensing policies or terms under which some parties have, it has been said, created a revolution in software distribution (i.e., license agreement terms and practices). With apologies to the reader, it will take a few paragraphs to describe and define these policies and terms so as to answer the opening question. The short answer, however, is that open source software is not given away but rather is made available in accordance with an open source license characterized by the following:

1. Delivery of source code and granting the licensee not only the right to use the software, but also the right to access and study the source code;
2. Granting the licensee the right to modify the software;
3. Granting the licensee the right to redistribute the software; and
4. Obligating the licensee to distribute its modifications under the same terms 1–3.

In part, one must understand OSL by comparison to what it is not: a limited grant of rights to use but not examine, probe, or modify, as in conventional software licenses. OSL is a model for software development and distribution that defies the historical practice of most software developers to (a) develop proprietary products; (b) only distribute object code or machine code versions of their software products; (c) prohibit, in their software license agreements, any copying or other reproduction, reverse engineering, decompilation, disassembly, or modification of the proprietary software product; and (d) allow only as-is use of a program on a single computer.

Neither nirvana nor the end of the world, although proponents and opponents, respectively, might disagree, OSL has been called many things, including disruptive (in the sense economists use that word) and communistic, as well as the only socially responsible way for software developers to behave. One point on which virtually all agree, however, is that OSL has been controversial and thought provoking and now is entering mainstream usage.

WHAT ARE THE ORIGINS OF OPEN SOURCE?

Although often thought of in the context of the licensing of Internet software, the open source movement actually arose out of work on operating systems, not Internet transactions. Moreover, "open source" embraces not just a set of *license* terms but also, importantly, an approach to software *development*, inherently affecting software distribution, that is partly manifested in those license terms (Pavlicek, 2002, pp. 7–9). As discussed below, conventionally developed products also sometimes are distributed under open source licensing terms, so it is important to distinguish open source development and open source licensing.

OSL started as somewhat of a grassroots movement among a small group of software engineers, but open source development and licensing have now been embraced by some major software developers, including IBM, to at least some degree. More important, corporate information technology (IT) departments that a few years ago treated open source products as unworthy of their attention now regularly acquire and use open source products, for their low cost, for their quality, and for the availability of support. An open source Web server package, Apache, is among the most popular products—and is the most popular product in its class (Pavlicek, 2002, pp. 23–24).

Before the open source movement and its immediate predecessor, the free software movement, most commercial software developers (and many "freelance" developers, as well) sought to control their products and their code as tightly as possible. The developers' goals included preventing competitors from freely copying their products and forcing customers who desired modifications to commission the developer (for a sizeable fee) to make the desired modifications or enhancements. The (higher level) source code, in which the program originally was written, became a closely guarded trade secret of the company; only machine-readable object code was released to customers. Putting aside issues of copyright infringement and the software licensing agreement prohibiting such activity, a competitor or a customer desiring to fix a bug or enhance a software product obtained from a third party was at a serious disadvantage. Without access to the source code, it was a formidable task for a software engineer to determine precisely how a program operated and figure out how to fix or change it. This certainly had (most of the time) the desired impact of forcing competitors to write their own competitive programs from scratch, rather than trying to copy or build on the original developer's program. The new code would have to be developed under the same labor-intensive process as produced the first code, and it would have to go through the same sort of testing and debugging process. The objective of the first company was achieved: the second company did not get a running start by capitalizing on the first company's work (at least not to much extent). Customers needing modifications, however, often were (and still are) stymied by the costs and delays in obtaining corrections or modifications. Even worse, there was (and is, in current closed source, object-only distribution) no guarantee the developer would be willing to address the customer's needs, at all; sometimes, the developer has decided not to support a customer's request, and the customer has had no recourse. For example, a developer may limit the time it will support a particular version of its code. Thereafter, the customer has to pay for upgrades to receive support. That may not only be expensive, but there also are instances in which the upgraded product is not satisfactory for a particular user (who then must face the unenviable task of migrating to a different product instead of simply tweaking the version in use to deal with a particular issue).

Note, however, that in the early microcomputer era, some programmers, both commercial and hobbyist, shared much of their code, pursuing success through support and documentation. The open source movement rejected the concept of developer control, re-birthed the lost approach of the early microcomputer era, and established itself on the joint premises that (a) software development should be collaborative, rather than competitive, and (b) the user should be free to address its problems and needs. Because collaboration is difficult to legislate, open source *agreements* (as opposed to the development projects) typically focus on the user freedom aspect.

In outline, the idea behind the open source movement is that Company A should distribute not only its object code but also, more importantly, its source code, and that it should do so while giving permission to its customer (licensee), Company B (or anyone else who acquires a lawful copy of the source code through Company B), the right to modify and enhance the source code. However, Company A may—indeed, should—impose a condition that Company B accept an obligation to distribute its resulting products on the same basis. In other words, everything derived from an open source starting point is "tainted" and becomes an open source product, itself. (This approach is taken in the GNU General Public License [GPL] discussed later. However, the open source definition used by Open Source Initiative, also discussed later, has backed off from the "Midas touch" requirement which makes the use of the GPL "viral".) In this way, say the proponents, the number of individuals developing and improving a software product can be greatly enlarged and all developers can build on the shoulders of those before them. As both a goal and a consequence, a "community" of users can develop. The members of that community share the benefit of all of the bug fixes to and enhancements of the base code, as they mutate. Through online bulletin boards, listservs, and the like, developers and users interact with respect to ideas, problems, improvements, and so on. Even if the original developer or distributor were to go out of business, the community would still be there to support one another.

Notice that not once has a licensee fee been mentioned. Nonetheless, OSL does not mean the code must be distributed without charge; rather, the exchange of funds is irrelevant to the licensing so long as it does not thwart the previously stated objectives. Contrary to the belief of some, the "open" in open source licensing does not mean "free." It merely means the opposite of a "closed," secret source code. (In fact, some of those who chose the "open source" nomenclature did so expressly to avoid the confusion created by the term "free software" [*History of the OSI,* n.d.]). An open source developer may charge for its product. (It may not impose fees further downstream than its immediate customer, however. Otherwise, such "tolls" might inhibit expansion of the open source community.)

Where did this new OSL creature come from? Did someone go insane? Give away the fruits of creative effort and ingenuity and the investments of time and money? It reminds some of communism, not capitalism, and communism does not work, right? How could anyone decide to "give away" control over the fruits of his or her creative work? Isn't this like the Coca-Cola Company publishing the formula for Coke on the front page of the *New York Times*?

Mistakenly, many people believe that the open source movement is an outgrowth of the shareware and freeware movements. In the latter, software developers allow others to copy, use, and redistribute their products without paying for them. In the former, developers put personal computer users on their honor to pay a fee to the programmer (developer) if, after trying out the program, they find they like it. It is true that on a time line, OSL was born after the advent of both shareware and freeware, but it was not born "of" them. Typically, neither shareware nor freeware is distributed in source code form, and most frequently they are distributed with full claims of copyright protection. Moreover, the shareware and free software models were developed in the personal computer industry whereas the open source model and its direct antecedent, the "free source" movement, were born in the Unix industry and academic world. In many ways, however, OSL benefited from observation of the problems inherent in shareware and free software distribution models.

Turning the clock back to the 1960s, computer platforms—both hardware and operating system—were pretty much proprietary creations. "Standardization" occurred only at the level of programming languages. This meant that a platform-specific compiler (or interpreter) had to be written for each proprietary operating system to be run on that platform. Often, not only the compiler but also individual application software had to be rewritten to work on a new machine. This resulted in lengthy development times and high development expenses. Against this background, computer scientists and engineers at AT&T's Bell Laboratories in New Jersey decided in about 1969–1970 to create an operating system that would be portable between hardware architectures (Raymond, 1999). That would facilitate movement of computer programs—software—from one hardware platform to another and minimize the time and expense required to reengineer application software each time a new computer was built. The Bell Labs team decided to build its operating system in a modular fashion to facilitate portability. This was a distinct departure from prior OS development projects, prior operating systems being created as unitary products. The major motivator was to allow users to install only those modules they required, so that unnecessary software would not have to be installed and the required computer resources, particularly memory, would be reduced, thereby speeding up the operating system. The resulting product was the first Unix operating system. It was designed for internal use at Bell Labs, not to be a commercial product.

Soon, AT&T began to license Unix inexpensively, and in doing so, AT&T provided users its source code, which they could then use to analyze the operating system and fix problems as they occurred (that way, AT&T did not have to support them). As time went on, however, AT&T's licensing practices changed. As with the rest of the industry, their license fee became much more expensive, and they stopped providing source code. For many years, this left a void for people who desired an operating system that was robust, documented, and accessible to them in source code. But note that limits on the use of creative works are commonplace in other fields such as art, music, theater, and literature, so this is not to suggest that AT&T was doing anything improper or unusual. It was merely exercising its rights and options.

In the early 1980s, Unix versions proliferated, partly as a result of AT&T's changing position and partly as a result of a striving for improvement. The Berkeley Standard Distribution (BSD) of Unix was developed at the University of California at Berkeley and made popular at this time. Aside from its technical merits, BSD was noteworthy for being published under a license permitting use, copying and redistribution of the software. The license, however, was silent as to imposing terms on subsequent publication of derivative works. Consequently, numerous derivative versions proliferated, some open source and some not, some free and some not. Much was learned about the impact of unconstrained licensing of derivative works. In a real sense, one result of the unconstrained licensing of Unix derivatives was the loss of interoperability. That spawned a desire for a return to interoperability and spawned, in part, Linux. Space does not permit a detailed elucidation of the BSD and other licenses of the era. Several sources cited herein provide useful detail. Also see Wayner, P. (n.d.).

The Free Software Movement and the GNU Operating System

In the 1980s, Richard Stallman, a software engineer in MIT's Artificial Intelligence Lab, took the next steps.

In 1971, when Stallman joined the AI Lab, he used an MIT-developed time-sharing operating system called ITS (the Incompatible Time-Sharing System) written in assembler language for the Digital Equipment Corporation PDP-10 computer. The source code was available to Stallman and others, and he modified or enhanced the source code from time to time. The same was true with respect to a printer in the lab and other resources.

A critical juncture occurred in the early 1980s. DEC discontinued the PDP-10 computer family. This immediately obsolesced nearly all the programs composing the ITS operating system. New replacement computers would use their own proprietary operating systems which, by the way, would carry significant price tags. Worse, those operating systems were protected by copyright law and restrictive licenses. Now he could not develop bug fixes or enhancements to the operating system and could not share them with others or receive such work from others. Stallman's view was that such licenses had the effect of forbidding the user community from helping one another. He concluded that this was antisocial and unethical. This presented him with a choice. He would either have to join the ranks of those developing proprietary software released under restrictive agreements, leave the computer field, or figure out a way that he could do something "for the good" (Stallman, 2002).

Frustrated, Stallman came to the conclusion that restrictive licensing was improper and that all software should be "free." By that, he meant that software should be treated as an expression of free speech and not as a proprietary engineering development or product. Stallman's concept was that source code is merely knowledge and knowledge should be free so that it cannot be dominated by a few powerful people. ("Free as in speech, not as in beer.") No monetary connotation was intended to be attached to the word "free," just the suggestion of freedom. All Stallman wanted was for everyone to be able, without restriction, to use, copy, share, and modify software. Free software was to have these attributes, which obviously means that the resulting modifications of free software also had to be treated as free software, no matter how many hands it passed through.

Stallman was vocal and his beliefs attracted a following. He advocated that software be distributed under a "copyleft" license or policy, which essentially was a 180-degree swing on copyright and amounted to a disclaimer of the author's copyright (Stallman, 2002).

Many viewed Stallman and his adherents as a group of antibusiness extremists and troublemakers. In an era during which the cost of developing commercial software products easily could run into millions of dollars, the suggestion that the power should be in the hand of the software user (or others—including competitors—who might fix, enhance, or modify the code for the user) rather than in the hand of the developer was a radical notion not readily embraced by established software companies. This opposition group was strong, well-financed, and powerful.

Eventually, some of those involved in the "movement" realized they needed a platform the business community would find more acceptable. This led to the founding of the Free Software Foundation, which still exists and is headquartered in Boston, Massachusetts. (The Free Software Foundation Web site may be found at http://www.fsf.org.) The basic credo of the FSF is that software is a form of digital information technology that contributes to the world by making it easier to copy and modify information. According to the FSF, the copyright system grew up with mechanical printing and fit in well with that technology because it did not take freedom away from readers of books. An ordinary reader, who did not own a printing press, could copy books only with pen and ink, and few readers were sued for that. Digital technology is more flexible than the printing press, however: When information is in digital form, it can be easily copied for sharing with others. Easy copying makes a bad fit with a system such as copyright, they hold. The interests of users in having the ability to use and change digital information (i.e., programs) outweighs any interest of the creator, they suggest. The FSF advocates that society needs information to be truly available to its citizens—for example, programs that people can read, fix, adapt, and improve, not just operate (see the FSF Web site). They assert that software owners typically deliver a black box that can't be studied or changed. According to the FSF, as stated on its Web site, "Free software is a matter of the users' freedom to run, copy, distribute, study, change and improve the software. More precisely, it refers to four kinds of freedom, for the users of the software:

- The freedom to run the program, for any purpose (Freedom 0).
- The freedom to study how the program works, and adapt it to your needs (Freedom 1). *Access to the source code is a precondition for this.*
- The freedom to redistribute copies so you can help your neighbor (Freedom 2).
- The freedom to improve the program, and release your improvements to the public, so that the whole community benefits. (Freedom 3). *Access to the source code is a precondition for this.*

A program is free software if users have all of these freedoms." (Emphasis added; see The Free Software Definition, n.d., at http://www.fsf.org/philosophy/free-sw.html.)

In 1984, with the BSD experience as a precursor, Stallman started the GNU (standing for "GNU's Not Unix," a clever, recursive acronym) project under the FSF umbrella, to put into practice his beliefs (see the GNU project

Web site: www.gnu.org). The goal of the GNU project was, essentially, to make an operating system that would be and remain (economically and legally) free and open. He chose to make the operating system Unix-compatible so that it would be portable and to attract converts to it.

The GNU Public License

How can this "ideal" be effectuated? Having a freedom is one thing, what one does with it is another. Note that Freedom 2 is expressed as a permission, not as a mandate. If a developer distributes software into the public arena and includes the source code, nothing in Freedom 2 prevents enhancements from being kept private or from being restrictively licensed. They could be treated as proprietary, thwarting the idea of creating a free software community, or constraining the community and its freedom. Needed was both the development of a successful free software product as well as an agreement that would ensure the maintenance of downstream freedom.

Ironically, in the past when source code was made freely available, without the protection of copyright laws or trade secret laws, commercial developers were free to incorporate free source code into their proprietary products for their own profit motive and close off much of the freedom Stallman advocated. This, of course, is what happened in some instances under the BSD license; BSD source code found its way into proprietary products. To enforce continuing "freeness," Stallman developed a new legal construct that came to be known as the GNU Public License or the GNU General Public License, or more commonly, the GPL. The GPL creates obligatory "non-proprietariness" via a concept he called "copyleft." "Copyleft uses copyright law, but flips it over to serve the opposite of its usual purpose: instead of a means of privatizing software, it becomes a means of [making and] keeping software free" (Stallman, 2002, p. 59). That is, it indicates that the licensee may copy and distribute source code provided it does not impose restrictions on the sublicensees that would prevent them from doing likewise. Thus, there may be no extra charge for the source code and no restrictions on further licensing. Moreover, works created from the licensed code also must be distributed under the GPL terms (Stallman, 2002) to the extent they constitute "derivative works" under the U.S. Copyright Act. Section 101 of the Copyright Act defines a "derivative work" as "a work based upon one or more preexisting works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which a work may be recast, transformed, or adapted. A work consisting of editorial revisions, annotations, elaborations, or other modifications, which, as a whole, represent an original work of authorship, is a 'derivative work'." If part of the work is not a "derivative work" from the original GPL-distributed work, then under the GPL "Terms and Conditions for Copying, Distribution and Modification," subsection 2.c, it is not required to be (re-)distributed under the GPL.

The GPL established the legal foundation for the free software movement and eventually led to what we now call OSL. The following four principles were embodied in the GPL:

1. The right to use the software and have access to its source code,
2. The right to modify the software,
3. The right to redistribute the software, and
4. The obligation to distribute modifications under the same "free" terms.

By coupling the obligation of Point 4 with the rights of Points 1–3, the GPL broke new ground and induced new behavior requirements. Essentially, the GPL provided a new framework for software development and licensing.

FROM "FREE" UNIX TO OPEN SOURCE AND LINUX

The GNU project took off and garnered a significant participation, particularly in the academic world where Unix had taken hold and where experimentation and lack of funds were the norm. Even as of the 1990 timeframe, however, it had not crossed the bridge into the personal computing world. So it was still of interest to only a relative few. Then, in Finland, in 1991, this all changed. A college student named Linus Torvalds grew dissatisfied with the performance of his personal computer (PC) and its MS-DOS operating system. His dissatisfaction became the crossover point. He was experienced in the world of Unix operating systems and wanted a similar operating system for his computer as well as one that would be available generally for PC users. (Among the relative advantages of Unix, it need only load into memory a small kernel and utilities it needs at a particular time. This would allow Linus's PC to operate faster and more efficiently that would its manufacturer's operating system.) Torvalds initiated an OS development project and invited others to join him, via postings on an Internet Usenet newsgroup. He posted the code he was developing, inviting others to offer comments and improvements. As a group formed, he took on some assignments himself to lead the writing of different parts of the operating system for different parts of the PC hardware and allowed others to lead the development of other parts (Pavlicek, 2002).

The project took on the name Linux (a contraction of Linus and Unix). However, it was not Torvalds' ego that named the project but the host of the server that held the source code.

Although Linus focused his effort on writing the kernel for the operating system, there is a lot more to Linux than its kernel. Many of the other components of Linux came directly from the GNU utilities, thanks to the porting of the GNU C compiler to the Linux system.

Arguably more significant than Linus's initiation of the project and his code-writing leadership was his enthusiastic encouragement of others to make contributions and his public acknowledgement of those contributions. (For an excellent exposition of the development of Linux and, in particular, Torvalds's leadership role, see Raymond, E. [2001]). Because most individuals yearn for recognition, this aspect of Linus's leadership generated a great deal of positive reaction and brought others on to the team. All were volunteers. Had they not felt their contributions were recognized by their peers, perhaps many would have left the project. Torvalds both showed leadership behavior and demonstrated that a community of capable individuals with dedication and shared values and objectives was capable of creating an end product equal to or better than would normally have been produced by a company investing millions of dollars. He also substantiated the hope that collaborative development could produce rapid solutions to design problems and reduce the number of bugs in the end product. The success of the Linux project was perhaps the first real indicator that the open source movement could have more than *de minimus* impact.

Linux initially was adopted by a large number of computer professionals and researchers, academics, and students. It required knowledge and skill to install. To broaden the appeal of Linux, one obstacle that had to be overcome was the complexity of putting Unix on a personal computer. The various components to a complete Linux operating system were not nicely integrated in one place. A user often had to download the kernel from one site, various drivers, and other components from others, for example. Then the user had to configure those components to work together. After a few short years, the demand for Linux was sufficient to justify the creation of a business to pull together and integrate the components in one branded package, a so-called Linux "distribution" (Raymond, 2001). Companies such as Red Hat Software realized that users would pay for the convenience of having the components assembled and integrated, as well as for documentation and support. Computers running Linux, as of 2002, probably number several million, and the number is rising fast. The most popular search engine, Google, runs on a Linux "farm," as does Amazon, the most popular e-commerce site. Linux also has become popular in the world of embedded systems because a scaled-down, limited functionality kernel can be built that runs on just one megabyte of memory (Fink, 2003, pp. 33–34). Also, the lack of royalty obligations makes it desirable for use in products that will be manufactured in large volume.

Departing From the Free Source Culture

The culture of the group (often referred to as a "community" for its shared bond, mutual support, and other community-like attributes) developing Linux was, interestingly, different from that of the free software movement from which it evolved. The free software movement was motivated by a view of moralism and a rejection of the notion of intellectual property as applied to software. Their mantra was that all software should be free and unprotected by intellectual property laws or other constraints. This principle was embodied in the "copyleft" expression. Software distributed under a copyleft notice was, in contrast to copyrighted software, free for anyone to do with as they please. In general, this was, in the 1980s, a point of view that the business community could not accept as viable. In fact, it may be safe to say that a major segment of the business community was repulsed by the idea. A company that has invested millions of dollars in developing a software product hardly wishes to have others be able to copy the product freely with impunity, because that completely undercuts their ability to obtain a return on their investment. Consequently, there is no motivation for anyone to invest in a project to develop free software unless there is some way other than the direct sale of that software to obtain profitable revenue.

The Birth of Open Source

Many saw the GPL as going too far, even advocates of source code freedom. They judged this arrangement antibusiness. That concerned some in the leadership of the free software community. In the winter of 1998 (long after the Linux project started as a free software project), several of them assembled in California to address this concern, because they wished to obtain broader adoption of the ideas underlying their movement for collaborative development and "no handcuffs" licensing. They believed that free software was a powerful concept and the perception of the free software movement as antibusiness was hampering its expansion. This discussion percolated the idea that a marketing campaign was necessary and that new terminology was needed to facilitate the marketing effort. The terminology that came to be adopted is that the kind of software they promoted should be called "Open Source" to get away from the loaded and misunderstood word "free." Along with this new term, they developed criteria for its use. One participant, Bruce Perens, previously had headed an effort called the Debian Project. The Debian Project managed a Linux distribution (i.e., a coordinated collection of a Linux kernel and utilities provided as a package) that included only software furnished under GNU principles. Perens wrote up this approach in a document called the "Debian Social Contract" (see http://www.debian.org/social_contract.html). Essentially, the group accepted the "Debian Social Contract" as defining the construct for open source software. Without a controlling regulatory framework, however, individuals were free to state they were using an open source license even if their terms deviated from those of Perens. An organization known as the Open Source Initiative (OSI) therefore was formed to provide an "official" definition and certify licenses as complying with the definition. This "official" definition is formally titled the Open Source Definition, or OSD, and can be found at http://www.opensource.org/docs/definition_plain.php. The OSD permits more licensing freedom than the GPL, particularly as to mixing proprietary and open source code. One could, for example, add proprietary code to an open source Linux distribution and not have to notify or secure permission from the various programmers or organizations that had developed the Linux code; and there is no requirement to give them credit. OSD terms have been embraced by companies seeking to open their source code to achieve wide distribution. For example, Netscape did not use the GPL when it decided to open its source code. Instead, it fashioned its own unique Open Source License, complying with the OSD requirements. This license became widely used as the Mozilla Public License, which can be found at the OSD site.

If a license has been approved by OSI as meeting the OSD requirements, the licensed software may be identified as "OSI Certified Open Source Software" and an OSI certification design trademark may be displayed on the license, product screen, and package.

A variety of licenses are certified as meeting the open source definition. Some of them will be explored later. In many ways, the fact that there exists the flexibility to craft a license to meet different circumstances is a hallmark of the Open Source Definition. It does not require a "one size fits all" approach.

Version 1.9 of the OSD lays out the following 10 principles for certification of a license:

1. *Free Redistribution.* The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.
2. *Source Code.* The program must include source code, and must allow distribution in source code as well as compiled form. The source code must be the preferred form in which a programmer would modify the program. . . . Deliberately obfuscated source code is not allowed.
3. *Derived Works.* The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.
4. *Integrity of the Author's Source Code.* The license may restrict source-code from being distributed in modified form *only* if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modified source code. The license may require derived works to carry a different name or version number from the original software.
5. *No Discrimination Against Persons or Groups.* The license must not discriminate against any person or group of persons.
6. *No Discrimination Against Fields of Endeavor.* The license must not restrict anyone from making use of the program in a specific field of endeavor.
7. *Distribution of License.* The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.
8. *License Must Not Be Specific to a Product.* The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.
9. *The License Must Not Restrict Other Software.* The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.
10. *The License must be technology-neutral.* No provision of the license may be predicated on any individual technology or style of interface.

Open Source Development

The development of program code for an open source project proceeds in a different way from the development of closed source code. A closed source project typically runs according to a traditional corporate product development model. There is a hierarchy to the development team and communication tends to run in the vertical direction. By contrast, an open source development community is self-organizing, with little hierarchy and a great deal of communication horizontally. It is a relatively flat structure. There generally are individuals who are nominally responsible for managing different pieces of a project and for making final decisions among competing code offerings, there is no true captain at the helm of the ship. Any member of the community may contact any other member of the community and direct contact to a code developer is encouraged as the most efficient way to communicate bugs, desires for modification and suggestions. Mechanisms may be established to formalize some of this communication to track the development process, but there is generally no hierarchy presenting an obstacle for one member of the community to communicate with another.

One of the driving motivations behind the open source movement is the concept that two heads are better than one and that more is to be gained from cooperating than from competing. That is, the larger the team, the better the result. The collaborative development of source code is expected to result in a high-quality product, produced in a shorter development cycle.

Open source developers have an expression: "Given enough eyes, all bugs are shallow." The meaning of this expression is simply that an open source code approach to development enlarges the development team for the project. Many users will be developers with appropriate expertise to both identify bugs and fix the source code to remove them. If one person does not recognize the fix, then another one will. Furthermore, companies that develop computer software typically commit to support only a given version or a few recent given versions of their products. Earlier versions are unsupported. However, source code distributed through OSL has a team of developers that is available for support so long as the product is in use.

OPEN SOURCE ECONOMICS

Some have called the open source movement "disruptive" in an economic sense. That is, a point of departure from the past that creates a marked change in direction. In fact, it may well be that the open source movement has taken hold in large part because a sufficient number of students who had experienced in college and graduate school the benefits of software developed under free and open source principles. Since becoming employed in industry, many have sought ways to incorporate open source approaches into their software development and marketing. The central question arises: "How does a company make money if it is making its software products available under an open source license?" The answer, of course, is that something other than the code has to have value to users. This may, for example, include support, customization, and training. As of this writing, the prototypical example of a success story is probably Red Hat Software, which markets its flagship product, Red Hat Linux, under an open source license. Red Hat packages a complete Linux distribution with a manual. To a large degree, it is selling convenience and documentation, as well as service. The Linux kernel and Linux utilities can be downloaded for free from various sources and can then be configured to work together if one has the expertise. By marketing a complete Linux distribution, Red Hat relieves its customers of the time-consuming effort of finding and downloading the various components of a Linux operating system from a variety of source and eliminates the need for them to take time to integrate the various components. In addition to convenience, Red Hat is selling a trusted brand name and a known commitment to its product and customers. The pricing for its product is such that the benefit of paying for something that actually is freely available makes a great deal of sense.

Now, one might say that Red Hat owns a large share of the Linux market by virtue of having seized the opportunity at an early stage and as a result of having managed that position well, to the point of developing brand recognition. Red Hat also develops, and supports the development of, new code to enhance its Linux product on an ongoing basis. Customers appear to be willing to pay a modest price to support those efforts and the convenience of a packaged Linux distribution. But Red Hat has only recently become profitable. One may ask whether there is an opportunity for a second, third, or fourth profit-making company in the Linux space. IBM makes substantial revenue from Linux-related services and equipment sales and UnitedLinux (the merger of a number of other companies) has significant market share. The open source nature of the software product appears to establish a very low entry barrier; however, it limits the price the market will support for the product (although not necessarily for service), and effective brand management by Red Hat (and maybe others) may create another entry barrier that could prove even more difficult to overcome. There are strong reasons to believe that service revenue may be the key to profitability for an open source software company.

One successful model in the open source arena is that of virtually giving away the product and selling support, training, and customization. Companies following this model include not only Red Hat (although it does generate revenue from selling the product as well), but also VA Research, Penguin Computing, LinuxCare, Send Mail, and MySQL AB. In some cases (e.g., that of Send Mail), an "enhanced" proprietary version of a free, open source product is available for a fee.

How does IBM profit from open source software? One way is by providing service and support. For example, IBM installs and supports open source Apache Web server software. Apache is an open source Web server product that is widely recognized as having very stable code. IBM gets all the profit from installation and support activities without incurring full development costs. Moreover, its engineers are kept in the communication stream on the Apache development community. Other companies are open sourcing their code to enlist the aid of the programming community to make their products more robust while achieving larger market penetration.

IBM has also been porting its "DB2" database program, its Websphere line of e-business solutions and other mainframe products to the Linux operating system, has ported Apache to platforms including the popular AS 400, and is developing new Linux applications. IBM has a Linux portal on its Web site highlighting its Linux products and commitment.

About one third of the Web servers currently in use are Linux-based machines; in the overall server market, Apache owns greater than 50% market share.

In the eyes of advocates, everyone benefits. This is not always a businessperson's perspective, however. In particular, if a product has been developed with proprietary investment and expectations by a business, pressure from the open source community to distribute the product under open source terms is sometimes greeted as an attempt at mob-supported theft. Some efforts have been made at hybrid licenses, part being open source and part being restrictively licensed under copyright or patent protection. Examples will be discussed subsequently. The latest approach is so-called "dual" licensing programs. Dual licensing allows a licensee to choose between (a) a free or very inexpensive open source license and (b) a more expensive license which does not impose open source obligations. The second choice permits the program code to be included in a proprietary product.

The Economic Case for Acquiring Open Source Software

To an individual, saving $50 to $100 on a personal computer system by using an open source operating system (OS) instead of a proprietary OS may not provide much motivation in the selection decision. Corporate customers, however, may have many thousands of PCs. A savings of $50 to $100 per machine for an OS license can translate into a substantial savings. Moreover, there are also cost savings on compilers, application software, servers, databases, for example. Additional cost savings may flow from the absence of license restrictions. The user can configure the system to its own needs, simplifying usage. Utilities can be written inexpensively for special situations, owning to the open code base. Consequently, the total cost differential between an open source approach and a closed source approach can be substantial.

Some users particularly value not being held hostage to software companies that are constantly pushing upgrades and improvements for additional revenue or that provide bloated code requiring newer and more costly machines when the user will never require many of the features of the bloated code.

Additionally, although there is perhaps no inherent reason for this result, there is evidence that the OS code for Linux is, indeed, better and more efficient and requires a lesser machine (i.e., one with a slower, less advanced central processing unit and less memory) than commercial, proprietary OSs such as Microsoft's Windows family. One may speculate that the commercial motivation to get a proprietary product completed at a minimum of expense militates against investing additional time to optimize and minimize program code, providing an inherent (although not inherently realized) potential advantage to open source rivals.

A 2002 survey by TheOpenEnterprise.com of more than 200 IT managers at major corporations showed that 86% of IT departments now use open source software (OSS), and about 30% use it for 40% or more of their applications. Of those surveyed, 90% expected to be using OSS in 2003, and 35% expected it to amount to 40% or more of their applications. Two thirds reported a savings of at least 25% in total cost of ownership over proprietary software. Three quarters characterized OSS as important to their corporate IT strategy, and only one in seven complained about the quality of OSS (TheOpenEnterprise.com, 2002).

Additionally, government initiatives, regulations, and even legislation have been promulgated or are under consideration in a number or states and countries including California, the Republic of Korea, Germany, and Peru, as well as the U.S. Government, authorizing or approving acquisition of open source software.

Non-Open Source "Alternatives"

It may be instructive, in contemplating the impact and economics of open source licenses, also to consider, from the side of the licensor, some arguable near-open alternatives reflecting a desire to capture much of the value of OSL but not all of its strictures. The most prominent of these alternatives is probably the license available from Microsoft under its Shared Source Initiative (SSI). Examples may be found, for example, at http://msdn.microsoft.com/msdn-files/027/001/901/ShSourceCLIbetaLicense.htm and http://www.microsoft.com/windows/Embedded/ce.NET/evaluation/sharedsource/eula.asp, both retrieved April 27, 2003. In Microsoft's words, its Shared Source philosophy is "a balanced approach that enables commercial firms to share source code with their customers and partners while preserving intellectual property rights" (http://www.microsoft.com/licensing). Although it does provide source code to the licensee, largely for the purpose of studying integration issues and debugging the licensee's product, a license under the SSI does not permit commercial product to be redistributed from the source code. Neither does it permit modification of the code. Some versions of shared source licenses Microsoft has announced will allow modification of code provided the rights to the modifications are transferred back to Microsoft. The Shared Source Initiative is in its infancy but Microsoft is pressing forward with it. Arguably, the SSI is an anti–open source effort. Microsoft, in some of its license agreements, forbids the introduction, commingling, or use of its products with open source code and fears the inclusion of open source code into its own products by its developers, intentionally or accidentally.

Other noteworthy near-open source licenses include the Netscape Public License and the SUN Community Source License (Fink, 2003, pp. 50–51). The former was used by Netscape before the Mozilla open source code base, for its browser, and the latter requires, among other things, royalty payments to Sun Microsystems upon redistribution of source code.

IS THE OPEN STANDARDS MOVEMENT RELATED TO THE OPEN SOURCE MOVEMENT?

The open standards movement is probably an even more recent phenomenon than the open source movement and is directly related to the Internet.

To take a 30,000-foot view, all communications is, of necessity, based on a priori agreement on standards. Otherwise, one party or device would not be able to be understood by another party or device. To take an elementary example, the fact that we can buy a new telephone (wired or wireless) and send or receive calls is made possible by adoption of standards and manufacture of telephones so that all can send and receive the same signals. The Internet, being a communications medium, is a physical network and a set of standards for communicating information over that network. The evolving development of standards is controlled by the Internet Engineering Task Force (IETF) and, in the case of the World Wide Web, the World Wide Web Consortium (W3C), based at the Massachusetts Institute of Technology. Both organizations have expressed a preference that Internet-related standards be "open." There is, however, a different meaning they apply to "openness" than that used in the "open source" context, although an overlap exists. No organization controls the definition of the term "open standard." (Although it is certainly beyond the scope of this article to expound in detail, an "open standard" is to be distinguished from a more common "industry standard." The latter may or may not involve software, may be free or royalty-bearing, usually pertains to selecting among competing pre-existing technologies rather than developing a new technology, and may (usually are) not open-specified. The significant shared trait is vendor-neutrality.) Industry literature suggests that a good place to start the effort to define an open standard is the definition of "open system" used by the Institute of Electrical and Electronics Engineers (IEEE), as the goal of an open system is the establishment of an open standard. According to that definition, an open system is one that "provides capabilities that enable properly implemented [software] applications to run on a variety of platforms from multiple vendors, interoperate with other applications, and present a consistent style of interaction with the user." Various authors have identified the points they believe to characterize an open standard, overlapping with this definition of an open system. (See, for example, http://perens.com/OpenStandards/Definition.html.) I conclude that there are four identifying characteristics. A standard is open provided it

- Is royalty-free,
- Is vendor neutral,
- Is open-specified, and
- Excludes predatory practices.

The royalty-free requirement means that there can be no use-related tolls. There may, however, be charges imposed for copies of the standard and for certification that software or hardware complies with the standard. The requirement of vendor neutrality means that a proprietary (e.g., patented) scheme normally will not be adopted unless licenses are made available royalty-free by the proprietor. It also means that no single-source material or product can be required. “Open-specified” means that the standard is not closed; extensions can be added and subsets may even be possible. Finally, the prohibition of predatory practices means that a license to the standard cannot require, for example, nonuse of other, competing technology or forced grant-back of rights to modifications.

Thus far, the only examples of open standards development known to the author have been in the Internet sphere. Likely this is because open standards development is focused on new technologies (whereas more common industrial standards most often address the harmonization of, and selection from among, existing technologies). Most recently and prominently, the W3C has had a Patent Policy Working Group addressing the issue of what W3C policy should be with respect to inclusion of patented technology in WWW standards. The current draft policy requires patented technology to be royalty-free if it is to be knowingly incorporated into a W3C standard. (Further details of the policy are beyond the scope of this chapter and may be read at http://www.w3.org/2001/ppwg/.)

OPEN SOURCE AND INTELLECTUAL PROPERTY

The relationship between OSL and the developers’ intellectual property rights is widely misunderstood. A common, first-level understanding is that all intellectual property rights—be they patent, copyright, or trade secret—must be relinquished to have one’s code distributed “open source.” This is not quite correct. One must consider each type of right separately to understand the compatibilities and incompatibilities.

Copyright confers on the “author” of the code the exclusive right to copy, modify, and so on. Consequently, it allows the owner of the copyright to control whether—and under what conditions—others may copy, modify, and so on. the copyrighted work (e.g., program code). The freedom to copy and modify, of course, is at the heart of OSL. So it would appear that the author must relinquish copyright rights in code distributed under an open source license. This need not be an all or nothing proposition. If the author or developer wishes to charge for the code, or to impose other conditions such as obtaining a reciprocal license, it is the grant of permission to exploit the underlying copyright that provides the basis for making those conditions enforceable, for imposing a fee on the licensee, and supplies the consideration needed for an enforceable contract. Others, who do not agree to the conditions, do not receive the permissions. So a copyright still has value in allowing the owner to control and enforce his or its open source plan.

Trade secrets and OSL, by contrast, are essentially antithetical. The sine qua non of trade secret status is the factual predicate of secrecy. No secrecy, no trade secret. Except in the rarest circumstances, however, source code is analyzable to derive from it any trade secret embodied therein. Object code, by contrast, often is used as a distribution medium, in the hope of hiding trade secret algorithms from disclosure. So at least an instance or example of the application of a trade secret will be discoverable from open source code; only a higher level trade secret that was used in generating, but not apparent from the end product code, will remain undisclosed.

Patents pose a far more complicated situation because there usually is not a one-to-one correspondence between a patent’s scope and particular program code. For example, a patent may have dozens of claims defining islands of protection for an invention. A program product might embody the invention of only a few of those claims. Conversely, a patented invention often will be useful beyond the instantiation of a specific program. So the owner will want to be cautious about the scope of any license granted and the impact of that license on other opportunities. This “caution” may manifest itself in the source code for an open source distribution or an open standard implementing a base technology which the entire industry and user community is free to practice, whereas the patent owner reserves a proprietary position with respect to certain enhancements.

Competition versus Coopetition

Even more fundamentally, intellectual property is, at its core, a vehicle supporting a competitive market. The overriding objective of the patent and copyright systems is to benefit society by encouraging and rewarding innovation. That is, to put the innovator in a better position to introduce new technology, products, and so on. Depending on the scope of protection that copyrights and patents afford the innovator, others may or may not be able to compete (without infringing), so in this sense the innovator is placed in a favorable competitive position. By contrast, the open source movement relies on cooperation rather than competition among the product developers in a market. To create open source software, the developers must cooperate with one another, including competitors. Thus, concurrent cooperation and competition are employed, a concept that has been labeled “coopetition.” They compete for customers (through price, unique features, or additions to their code, service, etc.) and cooperate in developing program code they all may use.

Doesn’t such cooperation mean giving up the intellectual property tools of competitive advantage which are so necessary when competing for customers? The answer is both yes and no. Yes, some intellectual property rights must be relinquished or licensed freely. But this code, by definition, will not be the code that gives competitive advantage. By prior agreement, the code developed in an open source project will be available to all. So each party will only give up what it has agreed to give up. Nothing precludes the development of additional code for enhanced features not required as part of the open source. In contrast with the GPL, the additional code will not, under a typical open source license, also be forced to be open source. Consequently, the “coopetive” approach allows a company to distinguish between work product not relying on intellectual property protection and work product that maybe kept under the intellectual property umbrella.

At least arguably, coopetition helps reduce development cost and time, lowers development cost to each participant, and allows the participants to influence the final product design.

There are, however, other laws that are called into play by coopetition arrangements—most notably the antitrust laws. Agreements that restrict the freedoms of competitors may be considered to have monopolistic or other anticompetitive effects, violating, for example, the venerable Sherman Act. Both open source and open standards projects pose challenges under antitrust law. Those involved require counsel from the beginning of their involvement. The U.S. Federal Trade Commission and Department of Justice have a major investigation in process, looking into the interplay between the intellectual property laws and the competition (antitrust) laws. The outcome of this investigation may result in new rulemaking or even legislation and certainly will be of great importance to the future of open source and open standard development.

Third-Party Rights

A discussion of the relationship between intellectual property and the open source movement would not be complete without noting two other situations. First, although open source software development was initially a group effort, there are now several examples of a single company developing a program and then deciding to distribute it under the terms of an open source license. The first major company to do so was probably Netscape Communications, which decided to employ an open source approach with respect to its Netscape browser in order to attract the developer community and better compete with Microsoft. Others have followed suit. Naturally, there is expected to be a business model fostered by this approach.

Second, if a third party owns intellectual property rights infringed by a particular open source software, the situation is more complex than a simple two-party dispute. Most likely, the intellectual property will be a patent, because a copyright only rarely can be innocently or unknowingly infringed. (In general, copyright infringement is established by showing that the accused infringer had access to the protected work and that the accused work is substantially similar to the protected aspects of the copyrighted work. Independent creation, which means there was no access, is a defense to the charge of copying.) If the patent owner does not dedicate its rights to the public, then either the open source development team must design around the rights or the owner has the right to insist on payment from each user (or an injunction against all use). The leader of the open source project usually will try to negotiate a resolution with the patent owner. In short, source code can only be distributed openly if the developer(s) has(have) the right to do so.

MANAGEMENT CHOICES

Although the initial reactions of most management to participation in open source projects and to the use of open source products were mostly strongly negative, those reactions have been changing and becoming both more favorable and more nuanced.

Of potentially greater importance than smaller companies such as Red Hat developing profitable business models based around sales of open source products is the fact that larger companies, including IBM, Oracle, and others, have embraced the open source model. A substantial percentage of certain IBM computers now are shipped with the Linux OS, for example. Indeed, in 2001 approximately 11% of mainframe computing power sold by IBM was Linux-based, and this number is increasing. IBM is now offering Linux on everything from PCs to mainframes (Bray, 2002). This allows IBM to sell computer hardware for a lower price than would be the case if they had to develop and deliver a proprietary OS. IBM is reputed to be spending more than $1 billion on open source initiatives.

Here, attention is again drawn to the distinction between the open source *development* process and the open source *distribution* (licensing) process. Some software products have been developed within the closed confines of a company only to then be distributed under OSL terms. That means the original development work was done in a closed source environment but post-release development can continue in an open source framework.

Is open source the way to go for all software? Clearly not. For example, there is a real danger that open source development, even its community or committee nature, can become a distortion of the originator's viewpoint. The Linux project, as noteworthy as it is, has benefited enormously from having strong leadership. Absent strong leadership, an open source development team could become quite dysfunctional. That is one reason an organization may choose to develop code in a closed environment and then distribute the "finished" product as open source.

Why develop code only to give it away? For recognition. For the altruistic purpose of sharing. For personal empowerment. To fulfill a personal need for code that has not yet been written. To facilitate completion of a larger project and secure its benefits. To improve the quality of the product and reduce its development cost and time, as well as subsequent maintenance cost. Also, to assure users of the continued availability of support. Perhaps to seed a market for, and generate interest in, a dual licensing agreement.

CHOOSING AN OPEN SOURCE LICENSE

As explained, or at least implied, here, any license that meets the open source criteria may be considered an open source license, whether it complies with the OSI Open Source Definition or not. To state the obvious, owning to its importance, note that the OSD is not a license in itself. It is only a definition with which a license does or does not comply. If it complies with the OSD, a license may be so certified. At this writing, there are more than three dozen OSI certified open source licenses. As part of the certification process is a requirement to show why a previously certified license will not satisfy the proponent's needs, it is immediately apparent that the OSD allows a considerable range of variation in licensing terms.

The GNU library of C program utilities is made available under a special kind of open source agreement called the GNU Library General Public License (LGPL). The LGPL allows the user to link *proprietary* software to its library.

According to OSI, until early 1998, the GPL, LGPL, BSD, and MIT agreements were the most commonly used for open source software. At that time, Netscape's efforts to develop an open source license led to the release of the Mozilla Public License. It has since joined the list and might have become the most widely used. Many other licenses have been submitted for review and approval by OSI, and the list of approved licenses is growing. Full text of certified open source licenses is available at http://www.opensource.org/licenses. Following are a few brief observations about some of the licenses published there:

- The Sun Industry Standards Source License (SISSL) may be used, for example, to delineate the limits of a patent license in the open source code and establish that no license is provided apart from the source code.
- The IBM Public License Version 1.0 addresses more fully issues of downstream liability, grants a license under all IBM patents, and requires contributors to grant similar patent rights.
- The Ricoh Source Code Public License includes provisions allowing distribution of code in executable (object code) form.
- The Apple Public Source License requires notice to Apple of modifications, license to Apple of those modifications, and revocability in case of infringement.
- The full text of these licenses and numerous commentaries on them may be found on the World Wide Web.

THE FUTURE OF OPEN SOURCE

The forces behind conventional, "closed" source software cannot be counted out quite yet, despite advances made by the open source community and by Linux in particular. Microsoft remains ardently antiopen. Jim Allchin, Microsoft's Platforms Group vice president, has stated that open source destroys intellectual property and stifles innovation (Holland & Coffee, 2001). In addition to the direct threat to its operating system market posed by Linux, Microsoft is concerned about the possible insidious impact of the GPL forcing others to open their software; it sees this as a threat to innovation. As well, Microsoft is concerned that it could be trapped by the GPL to open source some of its own software if its developers inadvertently combined Linux or another GPL-licensed program with a Microsoft program (Bulkeley & Buckman, 2002). In the past year, Microsoft has waged war on open source software, particularly challenging government agencies in the U.S. and around the world, in an effort to stop them from embracing Linux and other open source products (Bulkeley & Buckman, 2002).

Nevertheless, from the survey discussed earlier, there appears to be little doubt that the open source movement is gaining momentum. In fact, there are efforts to extend the open source model beyond software, toward other forms of creative expression (see, e.g., http://www.creativecommons.org). The OSI, via its promulgation of a useful Open Source definition and certification process, has fostered both adherence to open source principles and adaptability to the licensor's needs. The corporate world has accepted open source products in increasing numbers and is adopting the open source Linux operating system on servers and other computers. Open source is not appropriate for every software project, but it appears to work well in a cooperative development situation, returning high-quality, rapid development, and low cost.

GLOSSARY

Berkeley Standard Distribution (BSD) A Unix distribution from the University of California at Berkeley, licensed under terms that were a precursor to open source licensing (OSL).

Debian Social Contract or Debian Free Software Guidelines A document authored by Bruce Perens that laid down the principles of OSL.

Distribution A package of components comprising a complete version of an operating system (e.g., a Linux distribution) or other software.

Free Sometimes, no cost. In the Open Source context, it equates with "freedom," as in free speech.

Free Software Foundation (FSF) An organization based in Boston, Massachusetts, that is the "successor" to the League for Programming Freedom. The FSF advocates that society needs information to be truly available to its citizens—for example, programs that people can read, fix, adapt, and improve, not just operate.

GNU An operating system project founded by Richard Stallman with the objective of creating a Unix-like OS that would be "free".

GNU Public License or GNU General Public License (GPL) The GPL, designed by Richard Stallman, established the legal foundation for the free software movement and eventually led to what is now called open source licensing. It embodied the four principles of (a) the right to use the software and have access to its source code, (b) the right to modify the software, (c) the right to redistribute the software, and (d) the obligation to distribute modifications under the same "free" terms.

Intellectual property (IP) An umbrella term for patent, trademark, copyright, and trade secret rights.

League for Programming Freedom (LPF) An organization founded by Richard Stallman and others and dedicated to the proposition that software should be IP-free. Predecessor of the Free Software Foundation.

License A legal grant of permission by the owner of property rights (including intellectual property rights) to some other person or entity (the grantee), allowing the grantee to use the property or perform some act that otherwise would infringe the owner's rights. A license grant may be included within a contract, often called a "license agreement," wherein the parties make promises or provide things to each other in relation to the grant (e.g., the grantee promises to pay royalties). The distinction between a license and a license agreement, in the open source distribution context, is discussed in the text.

Linux An open source Unix-like operating system developed under the initiative and leadership of Linus Torvalds.

Mozilla Public License An open source license promulgated by Netscape Communications for its Netscape browser, which includes bundled components. Now one of the more popular models for open source licenses.

Open Source Definition (OSD) A quasi-official definition of the requirements for an open source license, maintained and promulgated by the Open Source Initiative.

Open Source Initiative A California public benefit (not-for-profit) corporation that maintains the OSD, certifies open source licenses, and maintains a Web site of certified open source licenses.

Open Source License In context, may mean a software license conforming to the OSD, certified by OSI as conforming to the OSD. Some may use the term even in the absence to strict adherence to the OSD.

Red Hat Software A public company that markets an open source licensed Linux distribution, complete with documentation, as well as service and support.

Stallman, Richard Founder, member of the League for Programming Freedom and Free Software Foundation; initiator of the GNU OS project and creator of the GPL, one of the first open source licenses. (Stallman does not endorse the term "open source.")

Torvalds, Linus Father of the Linux OS project that some view as establishing the viability of open source licensing.

Unix An operating system that had its origins at AT&T Bell Laboratories and was the basis for Linux.

CROSS REFERENCES

See *Copyright Law; Linux Operating System; Patent Law; Trademark Law; Unix Operating System.*

REFERENCES

Bray, H. (2002, December 9). Linux is no longer just an upstart. *Boston Globe*, p. C1.

Bulkeley, W. M., & Buckman, R. (2002, December 9). Microsoft wages quiet campaign against free software. *The Wall Street Journal*, p. B1.

CONTU (1978, July 31). Final report of the National Commission on New Technological Uses of Copyrighted Works. Chicago, IL: Commerce Clearing House, Inc.

Copyright Act of the United States, Title 17, U.S. Code, Sections 101–810.

Fink, M. (2003). *The business and economics of Linux® and open source*. Upper Saddle River, NJ: Prentice Hall PTR.

History of the OSI (n.d.). Retrieved April 27, 2003, from http://www.opensource.org/docs/history.php

Holland, R., & Coffee, P. (2001, February 26). Microsoft puts more heat on open source. *eWeek*, p. 17.

Pavlicek, R. (2002). *Embracing insanity: Open software development*. Indianapolis, IN: Sams.

Raymond, E. S. (1999). A brief history of hackerdom. In C. DiBona et al. (Eds.), *Open sources: Voices from the Internet revolution*. Sebastopol, CA: O'Reilly & Associates.

Raymond, E. (2001). *The cathedral and the bazaar*. Sebastopol, CA: O'Reilly & Associates. Retrieved April 27, 2003, from the O'Reilly and Associates Web site: http://www.firstmonday.dk/issues/issue3_3/raymond/

Stallman, R. (2002). The GNU operating system and the free software movement. In *Embracing insanity: Open software development*. Indianapolis, IN: Sams.

The free software definition (n.d.). Retrieved April 28, 2003, from http://www.fsf.org/philosophy/free-sw.html

TheOpenEnterprise.com (2002). *Open Enterprise Research Results*. Retrieved April 27, 2003, from http://www.theopenenterprise.com/pages/allcharts.

Wayner, P. (n.d.). *Free for all: How Linux and the free software movement undercut the high-tech titans*. Retrieved April 27, 2003, from http://www.wayner.org/books/ffa

Organizational Impact

John A. Mendonca, *Purdue University*

INTRODUCTION

Embracing the Digital Economy

Two primary components of the digital economy are the Internet and e-commerce. The Internet is the backbone and enabler of the network-based business models and processes that are e-commerce; both have impacted and, with their rapid development, continue to impact organizations in a myriad of ways. The rapid development of the Internet and e-commerce has caused significant changes in organizational management and design, businesses processes, the competitive environment and the ways in which companies compete and adapt new business and revenue models, new products and product delivery channels, marketing processes, customer relations, the characteristics of work and workers, and the ethical and social environment in which companies operate. Netcentric organizations are finding that business-to-consumer, business-to-business, and other applications and business models are converging to form the e-enterprise, in which the entire value chain, from procurement to customer service, is fully digitally integrated.

Technological, economic, and societal factors have contributed to create the Internet, e-commerce, and the modern netcentric organization. The technical capabilities of the Internet, combined with intranets and extranets, make possible new ways to communicate and exchange information at any time, in any place, in a variety of ways. The rapid and continuing decline in technology costs relative to productivity encourages the adoption of these technologies. Economic pressures that support the creation of network-based organizations include the development of the global economy, a competitive environment that demands better, faster, cheaper products and processes, business-to-business alliances, the fast pace of market change, and the increased power of consumers. Societal pressures include changes in the expectations of consumers and workers regarding access to information about commercial goods and commercial and government services. More and more, consumers expect 24–7 service levels accessible via the Internet. As the technology matures, mobile commerce will become an expectation as well.

The extent to which organizations are impacted by the Internet and e-commerce is, of course, a function of how much and how quickly they adopt the netcentric model and embrace its capabilities. Companies that are culturally open to change and that have leaders who understand the digital economy will be affected most. A company that uses the Internet merely to post an online catalog (an early, first-stage implementation of e-commerce) is not as extensively impacted as an organization that participates in enterprise e-commerce. The following discussion assumes the latter. Regardless of the extent of adoption, two major themes form the base of this discussion on organizational impact: first, the Internet is a disruptive technology, primarily via its capability to reduce time and space limitations, thus eliminating traditional boundaries to commerce; and second, e-commerce is a disruptive business paradigm, which forces organizations to change and adapt to new business realities.

Expectations: Realized and Unrealized

Companies that adopt e-commerce seek three basic types of benefits—economic benefits, relational benefits, and strategic benefits. Economic benefits are achieved by creating better, faster, cheaper products through better, faster, cheaper processes that contribute directly to profitability and market success. Relational benefits focus on enhanced, sustained customer, supplier, and other partner relationships that help organizations to grow and maintain stability. Both these contribute directly to long-term strategic benefits, including survival, profitability, and competitive advantages in the marketplace.

It is very important to note that e-commerce is not an end in itself, but rather a means to an end. That is, it is ultimately measured by value contributed. The Internet and e-commerce processes and infrastructure are enablers of beneficial organizational attributes, but they are not guarantors of organizational and commercial success. Although e-commerce is a recent phenomenon, it already has a history. The first period of e-commerce, from the mid-1990s to 2000, was characterized by the explosive growth of dot-com companies and application of the Web to a myriad of existing and new commercial functions. The reality was that decisions often were made too hastily and were poorly designed and implemented. Valuations of pure e-commerce companies were artificially high. There were serious problems in integrating online sales into existing organizational infrastructures, as highlighted by problems with order fulfillment by retailers during the Christmas period of 1999–2000.

History may mark the collapse of dot-com companies in 2000 to 2001 as the end of the first era and the beginning of a second era of e-commerce. Although early e-commerce, through its exploitation of the Internet and the Web, was a strong technological success, from a business perspective it was a mixed bag of success and failure.

The current era of e-commerce, just now beginning, is one tempered by experience. It is business-driven, rather than technology-driven. Organizations will be more cautious in embracing the opportunities and capabilities provided by these new technologies and new business models. CEOs and managers will likely evaluate adoption more from a business perspective, considering carefully the business value added, the return on investment, and the true costs and benefits to the organization.

NOT BUSINESS AS USUAL

Historically, technology has been an enabler of business transformation, and the Internet is certainly not an exception. It is a disruptive technology that changes, sometimes radically, the way people and organizations behave, govern themselves, and interact with external agents. At the core of the value of the Internet is its capability to "collapse" time and space, allowing organizations to dissolve boundaries to better, faster, cheaper commerce (Ashkenas, Ulrich, Jick, & Kerr, 1995). The Internet reduces, and sometimes eliminates, time boundaries by enabling fast, instantaneous, or simultaneous communication and sharing of information. It lessens geographic boundaries that physically separate employees and organizations from one another and from their customers.

The organizational impact of the collapse of time and space that is enabled by Internet-based technologies is extensive. E-mail is a speedy, flexible, and economic method for exchanging text and multimedia files between geographically dispersed individuals or between an individual and a group. Instant messaging technology approaches conversation-like simultaneous communication. Streaming media have the capability of delivering news and information, employee training, and entertainment in real-time mode whenever the viewer is ready, wherever the technology is implemented. Network-based collaborative software allows synchronous or asynchronous communication. Search engines and intelligent agents ("bots" for short) scour Web-enabled files and deliver current information and resources for research and commercial purposes.

The Internet is also a disruptive technology in relation to the balance of power between companies and their customers. Fingar, Kumar, and Sharma (2000) suggest that because of access to information, the Internet "turns the producer–consumer relationship upside down, with the balance of power going to the customer." This is very different from the Industrial Age, when information was tightly controlled by companies.

Just as the Internet is a disruptive technology, e-commerce is a disruptive business paradigm. Several key characteristics of e-commerce have a direct and significant impact on organizations. These characteristics include the following:

Ubiquity. Traditional commerce is limited by place. The e-marketspace is available to consumers nearly everywhere, at times and in locations convenient to them.

Reach. Most traditional commerce is local or regional, concentrated in geographically accessible merchants. E-commerce's reach is extensive, with a potential market size limited only by online access. The Internet is a chief enabler of global reach.

Richness. This refers to complexity and content of a message or product. In traditional commerce there is a tradeoff between richness and reach—in order to provide rich content a seller needed to be face-to-face with a customer. The larger the customer reach, the less rich the content. Through mass customization, e-commerce can dispel that tradeoff.

Interactivity. Internet-enabled e-commerce supports two-way synchronous or asynchronous communication between sellers and buyers. Via access to product databases and digitized mediation mechanisms, customers can easily collect information and communicate their desires.

Personalization/customization. With e-commerce, sellers can target their products or services specifically to smaller groups of potential customers and can personalize products and services based on prior behavior or customer preferences.

Information transparency. This refers to the ease, efficiency, and effectiveness of information collection, distribution, and exchange made possible by the Internet.

The varieties of e-commerce, such as business-to-consumer, business-to-business, and consumer-to-consumer, embrace these characteristics and fundamentally alter the competitive environment. To survive, organizations must adapt to these characteristics of digital business. (See Table 1.)

MANAGING E-COMMERCE

Management: Control and Planning

The fundamental challenge to management from e-commerce is implementing the processes and infrastructure that support information management, rather than product management. The introduction of a disruptive technology and a disruptive business paradigm cannot help but significantly (and appropriately) impact the management structure, management processes, and the need for effective change management.

Through the deployment of the Internet, intranets, and extranets, information quality and richness have

Table 1 Organizational Impact: A Summary

BENEFITS
Decreased costs of creating, distributing, storing, and retrieving information for customers, employees, suppliers, and other business partners
Expansion of markets for products and services, including global commerce
New business models; new ways to compete
Increased communication channels for customers and suppliers
Always open for business (24–7 commerce)
New products and services and innovative ways to deliver them
Faster, better quality, and less costly transactional processing with suppliers and other business partners
Reduced customer transaction costs via disintermediation
Mass customization
Electronic payment
Ease of alliance building (via communications and information exchange)
Reduced time to market
Collaborative work is facilitated
Reduced telecommunications costs (for example, through use of virtual private networks)
CHALLENGES
Organizational redesign for effective e-commerce
Developing and managing the e-commerce infrastructure
Managing the transition to e-commerce
Identifying, evaluating, selecting, and implementing new business models
Cost of implementing fully integrated e-commerce business solutions
Measuring e-commerce as a business proposition and strategic decision
Power shift to customers via efficient and speedy information gathering
24–7 business support
Managing geographically distributed knowledge workers
Security for netcentered transaction
Reduced barriers to market entry by competitors
New set of ethical, political, and social issues must be addressed

improved, and its facile and rapid movement throughout the organization has been greatly enhanced. Information transparency, made possible by networks, can provide lower-level workers the information they need to make decisions with less direction from upper management. Managers and workers who previously merely served as conduits of information will be eliminated. This facility of knowledge transfer, overcoming time and space barriers, tends toward less structural formality, decentralization of decision-making authority, and greater reliance on skills. It therefore promotes a flatter managerial hierarchy because employees have greater independence and managers have a wider span of control. The increased customer focus of e-commerce, customization, and high expectations of flexibility and responsiveness also support a flatter management architecture.

Technological innovations themselves introduce new managerial challenges. The infrastructure of e-commerce is complex, requires high levels of consistency and reliability, and often is globally implemented. For example, managers must develop or acquire new employee skill sets for development and implementation. They need to understand the capabilities and limitations of new technologies and develop strategies for integration into existing processes and systems. Managing the interfaces and relationships in interorganizational systems and networks is particularly challenging.

In planning for e-commerce, organizations need to carefully evaluate their own current capabilities and identify new capabilities required for e-commerce. Selecting an appropriate business model that will be effective and successful is critical. In the rapidly changing digital environment, traditional top-down planning is often viewed as too cumbersome and inflexible. Because of this, the insights of front-line employees have increased significance and tend to support a more bottom-up strategy. Unfortunately, the "just do it" approach, which deemphasizes planning and delivers products more quickly, also has a high level of risk. Arguably, dissatisfaction with early implementations of e-commerce applications is a result of poor planning and design. These applications, primarily front-end sales applications, were poorly integrated with back-office systems such as order fulfillment, product distribution, and customer service applications.

A more successful approach may be continuous planning, in which planning cycles are shortened and there is ongoing performance measurement and feedback. Also effective are trigger-point planning techniques, in which companies devise multiple scenarios for future direction and identify market actions or characteristics that activate a particular scenario plan. Most important is flexibility and speed in responding to changes in the business environment.

To better control the transformation, companies can adopt e-commerce in stages. One schema (Turban, King, Lee, Warkentin, & Chung, 2002) suggests four implementation levels. At the first level, companies establish a presence on the Web with static online brochures. Online ordering and basic customer service are added at the second level. The third level introduces additional transaction capabilities, including customization and personalization. The fourth level exploits fuller capabilities of business transformation, such as supply chain integration and customer relationship management.

Managing Operations: Value Chain Integration

The value chain concept, proposed by Michael Porter, describes organizations as having primary sets of activities—inbound logistics, operations, outbound logistics, marketing/sales, and service. These add value to products and services as they are created and delivered and therefore set the base of value for sales to customers. Porter's examination of the impact of the Internet on these activities (Porter, 2001) shows that each of these can be altered significantly, primarily through information transparency and speed of information transfer. Just as significant, however, is the impact of the Internet on the relationships *between* these functions, specifically the enabling of the enterprise-wide and industry-wide integrative systems that are the commercial backbone of e-commerce. Arguably, functional integration is the key to e-commerce and the enabler of enterprise e-commerce systems (Kalakota & Robinson, 2001). With these systems, ordering, inventory, invoicing, and delivery processes can be completed or triggered simultaneously with a sales event. These systems include enterprise resource planning (ERP), supply chain management (SCM), and customer relationship management (CRM).

ERP systems, the grandfather of integrative systems, is an enterprise-wide, network-based transaction framework that links front-office and back-office processes, including sales, inventory management, production planning, distribution planning and control, finance, and accounting. ERP has been a major player in the replacement of legacy systems characterized by stand-alone file and process structures. They allow organizations to gain better control of their processes through better quality, timely, about-the-business information. They also enable the coordination of operations over widely dispersed geographic areas.

SCM systems, a component of collaborative commerce, integrate supply chain activities in the organization, including purchasing, materials handling, production planning and control, warehousing, inventory control, and distribution. The goals of SCM are greater efficiency and effectiveness through reduction in inventory levels and cycle time, better quality processes, and improved customer service. It reduces organizational uncertainty and therefore risk. Global supply chains allow companies to uncover resources in other countries and take advantage of lower costs, greater availability of materials and labor, and advanced technology or skills not locally available.

CRM systems integrate the marketing, sales, and service activities of the organization. They provide the infrastructure support for the evolution of an organization from a product-centric business approach toward a customer-centric one. The goal is a stronger, more informed relationship with customers through tracking product purchases and usage, providing contact history, and building customer profiles. The expected benefits to organizations include improved identification of customer needs, reduced marketing costs through target marketing, greater customer retention and enhanced capabilities for product personalization and customization.

The fully integrated e-enterprise is moving toward the integration of the above systems, with the addition of collaborative commerce with industry partners (Hoque, 2000).

Managing the Transition

A major concern of corporate leaders and managers is how to effectively and efficiently transform an organization from an older business design and model to one that is competitive and fully operational in the digital world. They must function as change agents, anticipating the need for transformation and carefully guiding the organization through planning and implementation of all the facets of new business paradigms. The transformation may be gradual or quick, narrowly or widely focused, involve only internal components or extensively involve customers and suppliers and other external entities.

To fully exploit the capabilities of e-commerce, interorganizational relationships must be developed, new alliances constructed, and processes and physical networks implemented. Supply chain management systems, for example, contribute value through the sharing of data and processes and require a high level of trust among participants.

In order to be effective, organizations that choose to participate in e-commerce must adopt new technologies, change existing business processes, and transition workers to new roles and skills. Major characteristics and processes for successfully managing the transition to enterprise e-commerce include the following:

Vision building. An interenterprise, customer-focused e-commerce vision that is shared by organizational leaders, workers, customers, suppliers, and other stakeholders is necessary for exploiting the capabilities of digital commerce.

Process re-engineering. Paper-based and functionally oriented processes must be redesigned to be digitally based, cross-functional, and interorganizational.

Architecture redesign. A fully functional digital enterprise requires an overarching design to support integrated internal functions and interorganizational commerce processes. An effective architecture is comprehensive, flexible, and scaleable.

Infrastructure implementation strategy. There must be a thorough and accurate assessment of required technology. An interenterprise architecture, which fits together the pieces or hardware and software into a network, is complex, is costly to build, and is a high-risk endeavor.

A capable strategy for transition will identify and develop all the human resources needed for implementation.

Organizational Redesign

Full exploitation of the capabilities of the Internet and e-commerce requires organizational redesign. Generally, information technology facilitates "radical changes to old organizational structures" (Wang & Shouhong, 1997). Embracing e-commerce accelerates required adaptive change and there is evidence that this change is occurring at the fundamental infrastructure level (Carr, 2001). One way companies redesign themselves is through globalization, which is enabled by the Internet and is a feature of fully implemented digital commerce. At the simplest level of implementation, companies can compete by establishing a Web presence that is accessible throughout the world, wherever there is Internet access. At its most sophisticated level, companies can more easily form international alliances that extend market presence and facilitate acquisition of resources at internationally competitive prices. As an example, the Global Trading Web, established by Commerce One, connects global e-marketplaces to provide partners with information, products, and services and to deliver economic advantages.

Alliance building for e-commerce is becoming a dominant model for large, medium-sized, and even small, companies. Because of the facility of communication and the lowering of coordination costs, companies can more easily exchange information as a resource and as a commodity. The result is increased value-added partnering, such as for supply chain management systems. The e-enterprise includes organizations in the same and other industries that work together through complex processes that bring together customers, suppliers, distributors, and others. An interesting phenomenon is the cooperation of competitors, for example in establishing industry exchanges that facilitate purchasing.

The ease and low cost of communication via the Internet has also facilitated further growth in outsourcing. As an example, application service providers sell access to Internet based software applications. Organizations that contract for these services can reduce costs and eliminate the difficulties of developing and maintaining complex systems, particularly those systems that are fundamentally operational and do not deliver competitive advantage. The virtual organization extends the outsourcing facility to its maximum capability.

THE COMPETITIVE ENVIRONMENT

Competitive Forces in the Digital Economy

The digital economy, through the Internet and e-commerce, has had a major impact on the competitive environment for organizations. Porter's classic model of competition (Porter, 2001) posits five forces that compose that arena: (1) potential entrants; (2) buyers; (3) suppliers; (4) substitutes; and (5) rivalry among existing firms. Digital commerce not only significantly alters these forces, but in addition introduces several new ones that lead to intensified competition.

Historically, the large investment required for marketing and distribution channels, which provide geographic proximity, has served as an entry barrier to new competitors. The Internet makes it possible for new entrants to establish a presence in the marketplace via creation of relatively inexpensive Web sites, thus breaking down the economic barriers of time and space. Hundreds of thousands of storefronts are on the Web. With the organizational infrastructure functionally "invisible" to the customer, companies of vastly different sizes and complexity compete equally, differentiated only by the quality and price of their products and services and their abilities to market. The reduction in the need for sales and customer service forces (via disintermediation) also reduces these barriers.

E-commerce enhances the bargaining power of organizational buyers in business-to-business transactions and the alternatives for the individual buyer in business-to-consumer transactions. Through Web searching, potential customers can efficiently collect product information, including pricing and availability. Customers come to expect 24–7 sales and service. Because of information transparency, they can more easily monitor processes such as product fulfillment. Due to increased information and lower switching costs, brand loyalty may be impacted. The result of this is a tendency to push prices downward and enhance competition via quality, differentiation and personalization in customer service. On the other hand, variable cost per unit (such as for order fulfillment and distribution) for digitized products and services is very low and tends toward a fixed cost even as quantity sold increases.

On the positive side, organizations that participate in Internet-enabled procurement, particularly in e-marketplaces, can gain increased bargaining power over suppliers. On the other hand, suppliers are able to reach more customers, tending to neutralize the advantage of purchasers. Overall the net effect is to give all companies, regardless of size or geographic location, equal access to suppliers.

Through the Internet, companies can greatly increase their product reach (product information disbursed over space and time) and richness (depth of information and customization). Companies can, for example, provide detailed descriptions of their products online, including user manuals and technical specifications. Thus the threat of "substitutes" is increased because potential customers have more information about the availability and characteristics of products. In this environment, a "first mover" advantage, in which a company is first in a market, is magnified.

New competitive forces introduced into the environment include the elimination of national barriers and innovation in technology-based products. In addition to increasing their markets, organizations that have the resources and skills to go global can gain competitive advantage via reduced costs of global procurement. Innovative technology-rich products, for example those that integrate communications and media on personal devices, provide strong competitive advantages for organizations.

Overall, the rivalry among existing competitors in a market is increased, with a rich mix of exclusively online and exclusively nononline companies and those that have both traditional and online components. Viewed from this perspective, the impact to organizations is negative—that is, they are forced to compete in nontraditional ways in an altered and uncertain environment. However, the positive economic effects of e-commerce, such as reduced capital investment and reduced transaction and administration costs, counterbalance the negative impact.

Despite the changes in the competitive environment, companies continue to compete via three generic strategies: cost leadership, differentiation, and niche specialization. Successful companies can use the capabilities of the Internet and e-commerce to enhance these strategies. For example, a company that exclusively uses the Internet for sales and marketing can reduce capital expenses and thus reduce costs. Mass customization and target marketing capabilities of the Internet can provide effective differentiation and niche specialization. Personalization, in which customers build their own profiles and companies provide information based on them, is a common differentiation strategy that distinguishes online services.

From Marketplace to Marketspace

Unlike traditional markets, in which physical goods are created and distributed, digital marketspaces deal with gathering, evaluating, packaging, and distributing information. These virtual marketplaces are characterized by new products and services, new ways of marketing and distributing them, and new business models.

A major difference in operating in a marketspace, as opposed to the marketplace, is the ability to digitize existing products and create new digital ones. Companies can separate information content (the real value-added product) from delivery medium and establish new distribution channels. Newspapers and magazines sell online versions of paper products. Music, video, and software are separated from storage/delivery media for packaging and delivery via the Web. Examples of digitized services include airline and hotel reservations, auctions, and remote education.

Intermediation activity, or the lack of it, has played an important role in the ways e-commerce companies relate to customers and operate competitively. Service intermediaries, particularly those providing search and purchase services, may be eliminated via direct customer interfaces. Disintermediation characterizes automated online environments in which customers conduct their own searches for products, collect product information, place orders, and conduct simple service activities such as monitoring and inquiry. It is the primary factor in reducing transaction costs.

The Web offers many new opportunities for companies to participate in reintermediation by offering value-added digital infomediation services. In the business-to-business e-commerce environment, infomediaries enable the creation of e-marketplaces and exchanges in which many buyers and sellers can interact, negotiate, and sell products and services. The concept extends into business-to-consumer and other types of e-commerce as well. Common examples include search services, in which companies search multiple Web sites to collect and deliver product and service information; broker services, in which companies provide transaction-enabling services such as payment clearing; evaluation services, in which companies collect and distribute evaluative information about products and services; Internet portal services, in which companies serve as gateways to information seekers by collecting, evaluating, and organizing resources; and peer-to-peer contact services for individual transactions (such as auctions) or just social interaction.

Alternative Business Models

There are many e-commerce business models that companies can adopt. Smaller, narrowly focused companies may choose one model, but others may adopt several, with a multitude of variations. New ones are created and abandoned regularly. The basic components of a business model are the same for traditional commerce as for e-commerce. These components include a value proposition (the product or service purchased by a customer), a revenue model (how revenue is earned, assessed, and collected), market opportunity (the market intended to be served), and competitive advantage (how advantage is gained over competitors in the market).

Of the above components, e-commerce revenue models are arguably the most differentiated from traditional revenue models. Laudon and Traver (2001) suggest these types of revenue models:

Advertising revenue. A company receives fees for providing an online forum for advertisers.

Subscription/membership revenue. A company charges subscription fees for online content access.

Service revenue. A company receives a fee for providing an online service, such as payment clearing.

Transaction fee revenue. A company receives a fee for brokering a transaction.

Direct sales revenue. Companies sell products, information, or services directly to consumers.

Affiliate revenue. A company receives a fee for online referrals.

Models may be characterized in a variety of ways—for example, by the type and role of the agents participating in a transaction, or by the nature of the transaction itself. Some models directly support revenue enhancement; others are enablers of better, faster, cheaper products and services and directly or indirectly support organizational goals. Table 2 summarizes some common models and variations.

CUSTOMER CARE INNOVATION

An innovative feature of e-commerce is the ability to conduct one-to-one marketing. With this approach, a company changes its behavior toward a customer based on what it knows about that customer, recognizing that no two customers will have identical needs, profiles, or customer histories. In traditional commerce, being customer-oriented meant focusing on the perceived needs of a typical customer, not the actual one. E-commerce can

Table 2 E-commerce Business Models

Business to business (B2B)
E-marketplace (exchange): Brings buyers and sellers together
E-tailer: Online sales to companies
Service provider: Provides online business services, such as e-procurement
Infomediary: Information broker

Business to consumer (B2C)
Portal: Packages common online services, such as search, news, e-mail.
E-tailer: Online sales to consumers
Clicks and mortar: Combination of online and traditional retailer
Content provider: Entertainment and other online information
Service provider: Provides consumer-oriented services, such as electronic payment

Consumer to consumer (C2C)
Online auctions
Direct sales
Online communities

Peer to peer (for example, file sharing)

Intraorganizational (Intranets for supporting employees, such as for human resources and training)

M-commerce (mobile, wireless commerce)

Collaborative commerce (interorganizational collaboration)

change that. Personalization is the process of collecting the user preferences and behaviors used to build profiles and the matching of products and services to customers via their profiles. From the customer perspective not all personalization is perceived as favorable. Personalization is appreciated by customers when it is perceived to contribute to greater control and facility in future transactions. They sometimes see it as negative, however, when it results in unsolicited offers or is perceived to breach privacy. Personalization that misses the mark can lead to customer dissatisfaction.

Products and services are suitable for individualized marketing where they can be easily produced in multiple, sometimes complex, variations and where customer preferences are easily determined. When applied correctly, expected benefits to the company include customer loyalty, trust, and referrals of new customers.

The Internet has made possible a set of innovative e-services that expand the "when," "where," and "how" of providing customer service. Many of these services adopt disintermediation approaches and are primarily self-service oriented. They include the following:

24/7 services. Online sales and service support all day, every day.

E-mail. Asynchronous communication that eliminates telephone wait time.

FAQ lists. Frequently asked questions, with search capability, an inexpensive way to help customers resolve their own problems and answer questions.

Chat systems. Real-time online chat with customer service representatives.

Online documentation. Customer access to thorough up-to-date documentation to resolve problems.

Automated response systems. Customers receiving e-mailed automatic responses, such as order confirmation, order status, or acknowledgment of e-mailed queries.

Tracking tools. Online resources for tracking transaction status, such as those provided by shippers.

Trouble-shooting tools. Tools that help customers to identify and assess problems.

THE IMPACT ON WORK AND THE WORKFORCE

The Internet and e-commerce merely *allow* some of the changes in the way people work, but *force* other changes. Efficient and effective communications, access to a wealth of various types of information, and the automation of processes encompassed by e-commerce inevitably lead to work redesign and an increase in virtual work. In fact, to fully exploit the capabilities of these technologies and processes, an organization must redesign what tasks workers must perform, where and when they perform them, who does the work, and what kinds of skills are needed.

Some tasks are best done by people, but many others within the e-commerce realm can be done effectively by computers. As more and more organizations adopt e-commerce and its functionality expands, more and more processes will be automated. Disintermediation efforts replace workers who once performed basic sales and customer service tasks and introduce knowledge workers who are capable of collecting, analyzing, and integrating information. They handle the more complex transactions and queries that require more integrated complex skills and cannot be easily automated.

Where and when people work is also changing. Virtual work, which is netcentric and information- and knowledge-based, is less tied to restrictive time and place parameters (Hitt & Brynolfsson, 1997). Mobile e-commerce delivers real-time information to remote locations and enables many types of commercial transactions to be executed outside the traditional work environment and work hours, creating a mobile workforce. Web-enabled personal devices will greatly extend functionality and mobility and VPNs will provide the flexibility and security workers need. Organizations will more frequently face the challenges of managing, evaluating, and rewarding virtual workers who are globally dispersed.

As noted earlier, the changing communication patterns of workers lead to a flatter organizational structure. They also lead to more collaborative work. Cross-functional integration, interorganizational and intraorganizational, is an essential component of e-commerce that is facilitated by collaboration. Task-oriented work teams are not limited by geographic proximity. Virtual communities, uniting managers, workers, and professionals within and across organizations and even across industries, can be more easily developed and maintained for supporting knowledge-sharing and learning.

There are several major negatives in this environment from the worker point of view, with which organizations must deal. One is the worker stress that is an inevitable result of extending work into leisure time and into homes and cars and other traditionally nonwork spaces. The fast pace of change, the need to adopt new skills quickly, and independence in knowledge work are also stress factors.

Another concern to workers is the ease with which processes can be outsourced. Once an infrastructure for virtual work is established, work can be accomplished abroad as easily as it can be done domestically. Whereas easier access to less costly human resources is a financial benefit to companies, "electronic immigration" results in job loss.

ORGANIZATIONAL RESPONSIBILITY: THE SOCIAL CONTEXT

Organizations that use the Internet and e-commerce to conduct business are faced with new social, ethical, and political issues. Because we live in an "information society," information and knowledge are powerful assets and the control, access, collection, and dissemination of information cannot be viewed outside a social context. The collection and storage of vast quantities of data are becoming easier and more cost-effective. Companies that benefit from commercial development of the Internet have a responsibility to use information technologies ethically and responsibly. Expected and acceptable corporate behavior in this regard is just now being defined and determined by new laws and court decisions. These types of issues facing e-commerce companies include the following:

Privacy. What information about customers and customer behavior can appropriately be collected? What information about employee behavior can appropriately be collected?

Information ownership. Who owns the information collected about customers and to what extent can it be treated as a salable asset?

Usage. What are appropriate and inappropriate uses of information?

Security. How can transactions be made secure from unauthorized viewing and from fraud?

Accessibility. As more and more information and services are digitized, what segments of the population become separated from resources and the availability of commerce?

Accuracy. Who ensures the accuracy of information?

Content. What commercial content is not socially appropriate for delivery over the Internet? How can children be protected from offensive material?

Trust. How can companies build trust among actual and potential customers?

Product portfolio. What products or services should be legally banned from online sales?

Content ownership. What are the most effective means for protecting intellectual property, both works that are subsequently digitized and original digital work?

Patents. What e-commerce business methods and techniques are patentable?

Governance. How can distributed content and commercial transactions be appropriately governed by traditional political entities?

Taxation. Should business-to-consumer transactions be taxed like in-store sales, and how can that be accomplished?

THE FUTURE

From a business perspective, two major determinants of the organizational impact of e-commerce in the near future are (1) the failure of the dot-coms at the end of the first era and (2) the slowdown in the U.S. economy at the beginning of the second era.

Instead of the near-frantic pace at which new online companies and new business models were designed and implemented, companies will more carefully consider the business value of e-commerce and adopt its capabilities only when justified. E-commerce must be viewed from the perspective of strategic value, not just technological advancement. Benefits and risks must be analyzed. Traditional business processes must be adapted and integrated with netcentric business processes. As research is conducted into successes and failures, new predictive models will assist organizations in making more informed investment decisions. A particular challenge for managers is to identify, apply, and validate sets of metrics that will assist in evaluation of expected benefits and risks and support ongoing assessment mechanisms.

An interesting phenomenon is that the distinction between Internet-based companies and more traditional ones seems to be lessening. Online companies are creating distribution centers and even opening physical storefronts. Organizations with traditional infrastructures build websites that complement traditional commerce and extend their market reach. Ultimately, e-commerce may be viewed less as a radical new paradigm and embraced as a gradual extension into new capabilities enabled by the Internet.

Technologically, the future of e-commerce is strong. The number of Internet users and the number of products and services available via e-commerce continue to increase. The capabilities of mobile personal devices are just beginning to be exploited and will provide a new frontier for digital business. New systems to support electronic payments are being more widely implemented. The capabilities for rich multimedia content are becoming realized as new broadband technologies are introduced and now separate media are converging. Security and trust issues are being addressed—although more slowly than desired by most participants.

CONCLUSION

The boundary-expanding capabilities of the Internet and e-commerce are gateways to new frontiers for all organizations that choose to embrace them. The potential benefits are great and are available across a myriad of organizational activities. But embracing new ways of doing business means fundamental change—sometimes radical change—that impacts how an organization operates, what it produces, and how it is organized and managed. Indeed, a change in culture and values may very well be needed to exploit those capabilities. Undoubtedly, technological, market, and societal pressures will continue to increase and organizations will be forced to adapt to survive and thrive in the new digital environment.

GLOSSARY

Collaborative commerce Network-based cooperation and information exchange between partners in commerce, for example, partners in a supply chain or in product engineering.

Collaborative work Cooperation and information exchange between individual workers.

Digital economy Economic activity based on digital technologies, including digitally based transactions and all enabling infrastructure. Also called the "new economy" and the "Internet economy."

Disintermediation The removal of intermediaries in commercial transactions, such as elimination of sales order agents in Web-based ordering.

E-enterprise Interorganizational and cross-industry alliance structure for conducting digital commerce.

Electronic marketspace A digitally based market where products, services, and payments are exchanged. Particularly used to emphasize lack of a physical marketplace. Also called an "electronic market" or "e-market."

Infomediary A broker of information services for e-commerce transactions.

Knowledge worker A worker who deals primarily with network-based information and complex information relationships.

Mass customization The use of e-commerce to mass produce products and services that can be tailored to individual wants and needs.

Netcentric organization Organizations that conduct business via networks, including the Internet, intranets, and extranets.

Reintermediation The creation of new ways to provide Web-based intermediation services between sellers and buyers (a characteristic of some business models).

Revenue model Within the context of a business model, a submodel of how a company earns, assesses and collects revenue. Examples include online advertising and subscription service fees.

Virtual organization A company that exploits the communication capabilities of the Internet to the point of outsourcing all but core business functions and minimizing physical location assets.

Virtual work Network-based work not typically tied to location and time restrictions.

CROSS REFERENCES

See *Collaborative Commerce (C-commerce); Digital Economy; E-marketplaces; GroupWare; Virtual Enterprises; Virtual Teams.*

REFERENCES

Ashkenas, R., Ulrich, D., Jick, T., & Kerr, S. (1995). *The boundaryless organization: Breaking the chains of organizational structure.* San Francisco: Jossey–Bass.

Carr, N. (Ed.). (2001). *The digital enterprise: How to reshape your business for a connected world.* Boston: Harvard Business School Publishing.

Fingar, P., Kumar, H., & Sharma, T. (2000). *Enterprise e-commerce.* Tampa, FL: Meghan–Kiffer Press.

Hitt, L. M., & Brynjolfsson, E. (1997). Information technology and internal firm organization. *Journal of Management Information Systems, 14,* 81–202.

Hoque, F. (2000). *E-enterprise: Business models, architecture, and components.* Cambridge: Cambridge University Press.

Kalakota, R., & Robinson, M. (2001). *E-business 2.0: Roadmap for success.* Boston: Addison–Wesley.

Laudon, K. C., & Traver, C. G. (2001). *E-commerce: Business, technology, society.* Boston: Addison–Wesley.

Porter, M. E. (2001). Strategy and the Internet. *Harvard Business Review, 79,* 62–79.

Tapscott, D., Lowy, A., & Ticoll, D. (Eds.). (1998). *Blueprint to the digital economy: Wealth creation in the era of e-business.* New York: McGraw Hill.

Turban, E., King, D., Lee, J., Warkentin, M., & Chung, H. M. (2002). *Electronic commerce 2002: A managerial perspective.* New Jersey: Prentice Hall.

Wang, S. (1997). Impact of information technology on organizations. *Human Systems Management, 16,* 83–91.

0-471-22204-6