Document Management, Digital Libraries and the Web

June 9, 1995

Larry Masinter <masinter@parc.xerox.com>

Abstract

Document management systems are used by individuals, office workgroups and enterprises to organize and keep track of the documents being produced as a part of their work. Digital Library technology is being developed by many organizations to make the world's knowledge available through computers and communication technology. The World-Wide Web is an Internet application being used by individuals, companies and other organizations for promoting themselves, their products, doing electronic commerce, and for providing information to the vast number of Internet users around the world. These three application areas have much in common and also significant differences. The paper notes the common elements and some of the technical issues common in these areas, and explores the opportunities for synergy when these applications merge.

Contents

---

1. Introduction

The terms "document management system", "digital library" and "World-Wide Web" describe applications with a number of common architectural elements, though they are distinct in many of their features, in their domains of use, and in the systems and protocols they involve. This first section of paper describe each of the areas, their critical properties, some examples of their use, and the systems, standards, and organizations involved in developing them. Section 2 then explores many of the common design issues that are facing developers in each of the areas. Finally, Section 3 sets out some of the opportunities for integrating the three application areas.

1.1 Document Management Overview

Document management systems are software packages designed to help individuals, workgroups and large enterprises manage their growing number of documents stored in electronic form[1][2]. Document management is seen as a way to help companies manage the intellectual property that is locked up in the company's documents, currently hidden away in a morass of directories and subdirectories in scattered file servers across their networks. Document management systems may be used for a workgroup (a group of users connected via a local area network) or an enterprise (everyone in a company, connected via a corporate network).

Document management is used to manage the entire life cycle of a document, from creation through multiple revisions and finally into long-term storage and records management. For example, workgroup document management systems often offer library services for preserving update consistency, similar to check-out and check-in capabilities of software source code control systems. When a user checks out a document, the system locks the document from other users' changes. When the document is checked back in, the document management system makes it available for others to revise. Along with maintaining update consistency, the document management application tracks revisions in a multi-author/editor setting.

Document management systems usually feature searching in repositories of documents both by externally applied information about the documents (e.g., user who entered it, date of revision, or version relationship) and by content (e.g., search on words contained within the document.)

Frequently, document management systems are integrated with imaging capabilities: the ability to deal with scanned raster images (fax quality or higher) of documents that originated in paper form, as well as with documents that originated in electronic form. While imaging applications traditionally had been a separate domain, the line between image management and general document management has been increasingly blurred in recent years. In image document management systems, optical character recognition (OCR) is used to analyze the document content and index the corpus for content retrieval, even when the documents themselves are retained in image form.

Document management systems are usually integrated with the desktop applications. That means that the user's application program -- word processor, spreadsheet, graphic editor -- is modified to work directly with the document management system. For example, if a user running WordPerfect pulls down on the "File/Open" menu, a search interface to the document management repository might appear rather than the standard file system dialog interface.

Document management systems are sometimes connected to or integrated with workflow systems, though the latter is strictly speaking a different application. While document management systems deal with storing and searching documents in repositories, workflow systems are organized around work processes. Thus, a workflow system contains a model of the tasks of an organization and the roles that individuals play in that organization, and routes the work according to the model of the work process. Of course, the results of that process are often stored in document management repositories, and document management operations are often steps in the tasks managed by the workflow system.

Applications of Document Management Systems

To make clear the function of document management applications, it may help to give some typical examples of how these systems are used:

- A large multinational law firm manages all of its correspondence and contracts in a document management system. Because the firm believes it has an obligation to offer similar legal advice to all

clients in similar situations, the company wants the system to keep track of all correspondence, contracts, and so forth as produced in each of its offices.

- A large aerospace company finds that almost every plane off their assembly line is different in configuration. The documentation for the repair and maintenance of the plane needs to match the configuration shipped. The document management system allows the configuration of the shipped documentation to match the product. As more and more manufacturers move into custom product delivery and just-in-time manufacturing, it has become increasingly important to have a system that can allow documentation to track the changes in the products.
- Offices accumulate large repositories of general correspondence and often look for smaller document management applications for tracking correspondence and business documents.

There are a large number of vendors of document management systems. Some of the major products and vendors include Documentum, PC Docs, SoftSolutions from WordPerfect/Novell, FileNet, Visual Recall from Xerox, and Mezzanine from Saros. Many other products include document management capabilities, including offerings from Verity, Oracle, and Lotus (Notes).

As document management products have developed, there has been a growing demand for standards to allow interoperability between them. Large enterprises discover that different workgroups within their organization have, for various reasons, chosen different document management products. As they attempt to integrate these products across the enterprise, enterprise-wide standard interfaces and interoperability become increasingly important.

To this end, consortia have organized to define standards for document management. For example, the Open Document Management API (ODMA) is a simple Application Program Interface (API) designed to let desktop applications (such as an editor or spreadsheet) integrate with any of a number of document management systems[3][4][5]. It redefines file access menu items such as "Open", "Save", and "Save as..." to call the document management system (if one is installed) instead of the file system.

At another level, there have been recent attempts by industry groups to define a middleware layer between the user interface and back-end document repositories, so that users in an enterprise can access documents stored in multiple document management systems across their enterprise. The two efforts by the Shamrock Document Management Coalition (Shamrock's Enterprise Library Services) and the Document Enabled Networking[6] specification are being merged into a new Document Management Alliance (DMA)[7] to promote a single standard interface. These initiatives are creating a set of standard interfaces that define system elements such as "document", "repository", and "attribute" as well as as operations such as searching, checking out a document, and retrieving it.

1.2 Digital Libraries

What is a digital library? The term is sometimes used in a relatively literal way to refer to a system or application whose function is chiefly to extend the reach of a conventional library, for example by making its collection available in electronic form to remote users. More abstractly, the term is used to describe any application or system aimed at providing access and services for a large electronic document corpus. Usually the users of such corpora are thought of as members of a general or specialized public, rather than the personnel of an organization or enterprise. Over the last few years there have been research and development projects of both types; see, for example, [8][9][10][11] and special issues of journals[12]. For all their differences and particularities, these projects have certain general characteristics in common.

Key Features of Digital Libraries

Digital libraries usually possess large corpora of information of generally high value. Not only is the

material of high quality, but also some care is placed on cataloging the material, and making sure that the origin, date, and other external descriptive information is accurate. Many digital library projects are concerned with providing digital access to material that already exists within traditional library collections, and thus concentrate on material that was originally intended for analog media: libraries of scanned images of photographs or printed texts, digitized video segments and so forth. Other projects extend the library metaphor to other collections such as scientific data sets, software libraries or multimedia works. A great deal of work in this area concentrates on providing enhanced content or access methods, with the problem often couched as one of providing a way of satisfying the individual's particular "information needs". This might be a chemistry graduate student looking for information for a research project, a high-school student downloading a multi-media chemistry text, or a market researcher looking for information about chemical companies.

Digital library systems and standards

While much digital library work is in its early phase of development, there is a rich tradition in the library community that has influenced the thinking and design of systems for Digital Libraries. Historically, library automation has taken the form of Online Public Access Catalogs (OPACs). The standards for online library catalogs include MARC[13] and Z39.50[27]. Another kind of metadata is represented by the Scientific and Technical Attribute Set (STAS), which defines a standard for metadata elements to describe scientific datasets as opposed to traditional bibliographic material.

More recently, a number of research initiatives have proposed systems and mechanisms for future digital libraries, including the six NSF/ARPA/NASA joint initiative projects, initiatives of the national libraries and library system vendors. Previous work in copyright management[14][15], document identifiers[16], and the Computer Science Technical Report project [17] also contribute to digital library technology.

1.3 The web: an overview

These days, it is hardly necessary to define "the web" at an Internet conference. (It's hardly necessary to define "the web" to the cab driver who takes you to the conference from the airport.) For the sake of contrast, though, it will be useful to lay out the web's key features here.

Key Features of the web

By "the web", I mean information on the Internet, as is accessed by individuals using a World-Wide Web or some other network information access tool. The web is accessed using one of the many web browsers now available. The web provides a *document interface* to information. That is, a users is presented with a document which includes links to follow and forms to fill out. By interacting with the document, the user causes a new document to be presented. The web, as an Internet service, is primarily public. A web site can provide access to a very large number of users across the world.

Example applications of the web

The web is used for institutional public relations and product information, personal communication, online publishing, and scientific, technical and scholarly interchange. For example, companies put up web sites about their products and services; a growing number of newspapers and information service providers are producing web sites. Students put up `home pages' covering their hobbies. Professional organizations and educational institutions give out information about their organizations and their resources.

Web systems and standards

There are a growing number of web systems and software packages, including those produced by sponsored research, university researchers and commercial vendors. Dozens of start-ups compete for attention.

The web systems and protocols, originally defined in the research community, are being refined by a number of companies and consortia (the W3C consortium, for example) and being standardized by working groups of the Internet Engineering Task Force (IETF). The IETF is developing standards for Uniform Resource Locators (URLs), Uniform Resource Names (URNs), the HyperText Transfer Protocol (HTTP), and the HyperText Markup Language (HTML). These elements are the principal elements of the World Wide Web. The web also includes other network search protocols and access systems. For example, the Gopher protocol defined by the University of Minnesota is part of the web, while the Internet use of the Z39.50 standard is defined by the Z39.50 Implementors Group (ZIG)[18].

2. Common Elements in Document Management, Digital Libraries and the Web

The three application areas of document management, digital libraries and the web share common technology elements. This section describes some of these common elements, how they're deployed in each area, and the general design problems that are shared by all three areas. With more coordination between the groups designing the systems and protocols in these areas, solutions that are deployed for one set of applications might be reapplied in others, duplicate effort avoided, and the opportunities for synergy enhanced.

2.1 Document Identifiers

In any computer system for manipulating information, it is important to allow objects to contain persistent references to other objects. These references are used from inside databases, in bibliographies, hypertext links, and in a variety of other ways. The approaches used in document management, digital libraries and the web have differed.

Identifiers in Document Management systems

Commercial document management systems all employ some kind of document identifier mechanism, so that pointers to documents in the document management system can be saved and referenced independent of that system. For example, ODMA has a document ID -- a persistent, portable identifier for a document -- that is accepted or returned by ODMA functions. It is used to save away references to documents, to refer to documents in electronic mail or by other processes. Other examples of document identifiers include those used in OpenDoc[19] and OLE. The OpenDoc standard uses the Bento file format[20], which incorporates globally unique identifiers to make references from one document to another. OLE use a variety of identifiers to keep permanent references valid between composite objects[21].

Identifiers in Libraries

Traditionally, the library community has developed a number of mechanisms to uniquely identify a work. These mechanisms include "call numbers" (e.g., the Library of Congress Call Number system which yields identifiers that are printed like PS3566O815.W4.1987), ISBN numbers (originally intended for inventory) and ISSN numbers (which identify serials, i.e., material that is updated regularly.) More recently, librarians have tried to apply this apparatus to digital works, which do not always lend themselves to traditional treatment and which raise a number of design issues involving the use of document identifiers[22].

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.