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].

Identifiers on the Web

In the World-Wide Web, the most common kind of identifier is a URL. URLs are probably familiar to anyone who has used a web browser or read the papers in this conference, where the references include URLs. While the name "URL" seems to indicate that it locates the object (says `where it is'), in fact, a URL is more like an `access method': it tells you how, on the Internet, to access the object. As many have observed, there is a serious problem using URLs when information or web resources move. There is a strong desire to create a new scheme for URNs that name an object independent of its location. Some kind of distributed URN -> URL location service (for which there is not yet an accepted design) would then be employed to find out the actual location of objects. Several proposals have been brought forward and are being evaluated.

Issues in Document Identifiers

There are a number of open design issues in the area of document identifiers. These design issues are present for dealing with electronic documents, whether in a library, a workgroup, or on the Internet.

Fragments, relationships

How does one identify a piece of something else? For example, if there is a volume of collected papers, do the individual papers get separate identifiers? If so, is the identifier for each element somehow syntactically related to the identifier for the whole? If not, how is the relationship established? Is there a database that links the part to the whole?

When an object is revised, does it retain its identifier? For example, in System 33[23], every document had two identifiers: one that was assigned to `this version' and another that specified `the latest version of whatever this becomes'.

In the office environment, a document with a cover memo attached might be considered a different object. However, in some situations, the `cover' material is merely an external attribute, and the document hasn't changed and should not get a different identifier.

In general, there are a large number of relationships between objects that can be expressed as relationships of the identifiers of the objects, and relevant design decisions are currently made in an ad hoc fashion. Publishers are allowed to retain the same ISBN number for minor printing revisions, but the paperback and hardcover of a book are given different ISBN numbers. On the web, the URL of a document doesn't change if the content changes. Moreover, different vendors' document management systems seem to take different approaches to dealing with revision and identity.

Uniqueness

There are a variety of methods used to ensure that different documents do not get the same identifier, even when different entities are assigning names. These methods rely either on a distributed hierarchy, or a probabilistic method of name assignment.

In a hierarchical uniqueness system, there is a tree of 'naming authorities'. Every naming authority guarantees that it will not give out the same identifier to two different documents. If it delegates some of the naming authority to sub-authorities, it also delegates that promise. ("Here, you can give out names, but you make sure you never give out the same name twice.") For example, the Internet's Domain Name Service is a hierarchical service; the owner of "xerox.com" can hand out unique names under that suffix, and to delegate the naming system underneath to the owner of "parc.xerox.com". Many of the proposals for

URNs on the Web are hierarchical.

Some distributed naming systems are hierarchical but have a fixed depth of the hierarchy. For example, ISBN numbers have three parts: a country code (the country of registry for the publisher), the publisher identifier, and, for each publisher, the document identifier. Each publisher is allowed to assign their own ISBN numbers. Some naming systems are not distributed, but guarantee uniqueness by keeping a single source of identifiers; for example, the Library of Congress Control Number is assigned uniquely by the U.S. Library of Congress.

A random naming authority is one in which names are given out using random numbers; each authority uses enough information to make the probability of two documents getting the same identifier quite small. For example, some schemes use the one-way hash (MD5, SHA) of the document as the document identifier. The LIFN system [24] uses a randomly assigned document identifier in this way.

Resolution

Given a name for an object, how does one go about finding information about that object? How much information is packed into the name? For example, ISBN numbers give you some clue about who the publisher is, and there is a global registry of publishers. If you can't find the document in your catalog, you can check the publisher. On the other hand, the random schemes give no hints. Using URLs, the identifier contains nearly complete information to access a resource across the global Internet. Usually, though, the more information contained in the identifier, the harder it is to for the resolution system to find objects when they have moved.

2.2 MetaData

In document management, digital libraries and the web, it is common to want to record information about documents that is not part of the documents themselves. These assertions are sometimes called `document attributes'; sometimes they are called `metadata' to signify that they are data about data rather than the information itself. Metadata assists in the description, organization, discovery and access to network information resources.

Metadata in Document Management

Most document management systems include mechanisms that permit at least the system administrator to define, according to the application, a set of attributes that are common to the documents in a repository or at least a variety of classes of documents. For example, many systems record the user identity of the originator of the document, the date and time of origination, other information external to the documents themselves, or some other attributes of the documents in the repository, as determined by the system administrator. A law office might index its documents by the name of the client; a manufacturer, by the product or parts codes affected within.

Metadata in Digital Libraries

Libraries have traditionally been quite concerned with cataloging -- a process which associates metadata with bibliographic material. The card catalog entries for an item in the library provides metadata about the item. There are a variety of standards used for online cataloging. The most prominent is USMARC. Various attempts have been made to extend and enhance USMARC to deal with online material[25][26]. The Z39.50 standard contains extensive mechanisms for both communicating search parameters (requested metadata) and document attributes (output metadata.) More recently, attempts to define online document

standards for the humanities arrived at a standard set of metadata for humanities texts[28].

Metadata on the Internet

The Internet community has several efforts to define a set of metadata tags useful for information on the network. For example, the Internet Anonymous FTP Archives working group of the Internet Engineering Task Force attempted to set a standard for describing FTP-accessible data[29]. In fact, one could think of the standard headers of an Internet electronic mail message as identifying attributes for each message[30]. Every Internet message has required attributes; for example, it must identify who it is "From" and "To" and the "Date" it was sent. In addition, there are optional attributes, such as "Subject" and "Comments". There are rules that specify the kinds of values each attribute can have.

The Uniform Resource Identifier working group[31] has been trying to develop a standard syntax and representation for information citations in a scheme called Uniform Resource Citations (URCs) to describe information on the Internet as a way of discovering or describing more about a referenced resource (via URL or URN) before retrieving the item, as well as a way of cataloging Internet information.

Issues in Metadata

There are a number of design issues in representing metadata for online information, some semantic (what does it mean and how do you say it?), some structural (does metadata have structure?) and some syntactic (how do the semantics and structure get represented as a sequence of characters or bytes?) These issues span the three application areas.

Semantic issues

Are there well known attributes? MARC takes a strong stand: MARC defines a set of well-known attributes with descriptions of each. Some of them take on values within a controlled vocabulary. There are standards for the completeness and quality of a catalog entry. The set of attributes is defined and used universally by nearly all online library catalogs. In document management systems, on the other hand, the system administrator for a workgroup generally establishes conventions for the attributes used and what they mean. When multiple document management systems are brought together, though, combining the semantics of the disparate sources is a serious problem. The Internet community is struggling with standardization of semantics for attribute sets. While there are some attributes that are well-known (content attributions in mail messages, mapping to ISO protocols in X.400), these are by no means universal.

If there is not a single well-known set of attributes that spans all known objects, then it is still possible to create a system of *entities* -- classes of documents which share the same schema of attributes. For each class, the attribute set can then be defined. For example, a document management system might allow for 'memo' and 'spreadsheet' and 'expense report'. Every memo might be catalogued by its distribution list, while an expense report might be required to have a budget center and a signature status. More complex schema systems allow for inheritance and specialization of classes, as is found in object-oriented programming. There are variations among different implementations, just as there are in different object-oriented programming systems.

Structural issues

Frequently it is difficult to tell the `boundaries' of an online electronic work. If one describes a site's `home page', does the description apply to the site, or just to the introductory `splash page'? If an object contains parts, do the parts have separate attributes? For example, if a report in a document management system has

a cover memo, in what way are the author of the report and the author of the cover memo distinguished or reported in the description of the overall object?

Metadata itself can also have structure. It is sometimes necessary and occasionally critical to know the author of an attribute or the time when the attribute was assigned. If metadata itself can be updated and revised, then the history of its editing may be of relevance. How does one distinguish between `the title' and `the title, translated into French', and `the title, translated into English from Italian by D.H.Lawrence'. The relationships between elements of the metadata are problematic for some flat attribute-value representation schemes like MARC.

Syntactic and system issues with metadata

While it might seem straightforward, standardization of the syntactic mechanisms for representing the semantics and structure of attributes is quite difficult. First, attributes might have a fixed, extensible, or uncontrolled set of values. The mechanisms for assigning the allowable elements of the controlled set are difficult to establish. Each attribute or field might need to deal with alternative syntaxes (e.g., for names, is it last name first or given name first?), multiple character sets (names in Chinese or Arabic), or even non-textual data.

2.3 Authentication, Authorization, Accounting (AAA) and Related Issues

There are several related issues having to do with security, rights, privacy, confidentiality and access that arise in all of the application areas. Authentication is the process by which the identity of a person (or system) is ascertained and assured. Authorization is the process of determining whether a given operation is allowed, such as reading a document or updating metadata. Accounting is the process of recording operations and the payment due for them. An audit trail of records of past operations might be kept, as a way of checking the integrity of the system.

AAA in Document Management

In document management systems, the critical elements of AAA are concerned with managing the permissions to access the information in a set of documents and maintaining the integrity of these documents. Some documents are confidential, others are public, others belong to particular workgroups. Most of the early work in authorization followed the military model of classified information and clearance levels; this model has been found to be inappropriate for many non-military applications. Frequently, the authorization system of the document management system is inadequate to represent and enforce the company's access control needs; for example, the actual work practice in many organizations will relax rules and guidelines in specific situations.

Despite the more complex needs, some document management systems rely on either their database manager or the host network operating system to provide authentication and access control, if for no other reason than to avoid providing a separate authentication and administrative domains.

AAA in Digital Libraries

In the library setting, the requirements for AAA often focus on copyright, payment methods, and usage rights; in addition, there is a significant concern for the privacy of the reader and information about what is being read by whom. The situation is made more complex by the difficulties in interpreting copyright law originally designed for physical material in a world of electronic reproduction and distribution. In many countries, the copyright law and practice around it is being reexamined in the age of electronic distribution.

In any case, it is clear that digital library systems will need to address issues of copyright and intellectual property rights before they can be widely deployed.

AAA in the Web

The Internet community has a large number of separate efforts defining security standards. The web community is exploring two systems, Secure HTTP (S-HHTP)[32] and Secure Socket Layer (SSL)[33]. S-HTTP is a modification of the HTTP web protocol that includes security features. SSL is an application-independent protocol for negotiating secure network communication. Recently these efforts have joined forces. In addition, new authentication mechanisms for web access (other than simple passwords) are being proposed using Digest Access Authentication[33] and Multi-party Digest Authentication[XX].

In addition, the Internet mail community has produced two complementary systems for secure electronic mail, Pretty Good Privacy (PGP)[36] and Privacy Enhanced Mail (PEM)[37]. PGP is a public key cryptosystem with a number of utilities for dealing with keys and mail. PEM is a system for providing privacy enhancement services (confidentiality, authentication, message integrity assurance and non-repudiation of origin) using either symmetric (secret-key) and asymmetric (public-key) approaches for encryption of data encrypting keys. There is some hope that all of these separate efforts will eventually converge.

Beyond the mechanisms for dealing with security, copyright and intellectual property, the Web is capable of providing for spontaneous financial transactions. A number of mechanisms for handling payment and billing are being explored, either through credit card settlement methods or digital cash[38].

The most serious issue is the design of an authorization scheme that will scale to the size of 'all users on the Internet', given the enormous international scope of the Internet and the wide variety of needs and policies requiring support.

Finally, US export control laws that govern the export of cryptographic software have been perceived as a difficult impediment to widespread deployment of secure software solutions to the Web's problems.

AAA Issues

It is clear that the main issues in each domain (intellectual property in libraries, complex authorization needs in document management, and secure communication for spontaneous transactions on the web) will also become important in the others. In particular, as enterprises grow their document management needs, the need for cross-domain authentication mechanisms grows. Likewise, the web will need richer methods for expressing access control and authorization than most web services currently provide.

One common issue in all of the systems is detecting the boundary of the item to which a particular authorization might apply. Access control and authorization might need to apply to a different granularity of object than is denoted with a single identifier.

In general, one of the most troubling elements of AAA design is that it is difficult to retrofit security in an architecture that doesn't already have it. The analysis of likely threats often requires revisiting optimizations made for performance reasons. For example, a design which employs distribution and caching of documents close to the site of access for performance reasons needs to account for the risks embodied in having a repository of cached documents which might be compromised.

2.4 Document types

Generally, digital libraries, document management and the web manage documents and not files. The unit of communication, the items being stored and retrieved are representations of intellectual content, not merely a sequence of bytes. However, documents are *represented* as one or more sequences of bytes in a file system. The representation is tagged with an indication of what kind of object it is. This labeling is itself an issue in each area.

Document types in Document Management Systems

Individual vendors of document management systems have frequently created their own ad hoc registries, to allow their systems to deal with multiple document types in a consistent way. More recent work in the electronic mail vendors association and ODMA group have created registries of well-known document types. Most generally, though, document management systems restrict themselves to dealing with the document types that either are common in desktop applications in the workplace or else are registered by the system administrator of the document management system.

Document types in Digital Libraries

The range of kinds of media and digital objects that potentially might be stored in a digital library is enormous. Currently, most attempts to catalog material have used fairly ad hoc descriptions of the files and their formats. A critical issue in the library community, though, is *preservation*[39][40]. It is important to make sure recorded material will be available in 10, 20, or 100 years. This is an issue not only of the longevity of the storage medium (which can be mitigated by refreshing the media), but, more importantly, the longevity of any particular storage representation. If one were to preserve a file that was created with Microsoft Word in 1995, how long is it expected to have a Microsoft Word-capable reader in the future? [39]

Document types in the web

The method for indicating the media type of an object in the Internet arose from work on MIME: the Multipurpose Internet Mail Exchange standard. MIME extended Internet electronic mail -- formerly confined to the interchange of ASCII text -- by allowing for a rich representation of objects and object types. The MIME standard allows for the labeling of an object by its media type. Media types are defined as a two part name (e.g., "text/html" or "application/postscript") along with optional parameters. Media types are categorized into several top-level types ("text", "image", "audio", "application", "multipart") and then, within each top-level type, an extensible set of subtypes. Each type can also define parameters; for example, "text" types can have a "charset" parameter where the character encoding used for the text is given. There is a formal process for defining new media types, where information about the type and required and allowed parameters are supplied.

Issues in Document Types

There are difficulties with the current mechanisms used for specifying document types that are common in all of the application areas, and affect the long-term interoperability and capability of the typing system.

Type attributes

In many scenarios of use, it is important for one system element to be able to interrogate across the network the type of a digital object to determine if the local system is capable of processing or rendering the object. For example, a reading machine might not bother to retrieve an image-rich rendition of a document, but prefer one with structural markup. In some cases, the coarse denotation of 'image' is not

sufficient; for example, it is important to note externally whether the image is color or black-and-white, its resolution or other attributes. Text documents may need to be annotated with a description of the character encoding employed or the fonts used. These sub-type attributions are difficult to deal with in many document type definition systems.

A related problem is that many document types are merely references to specifications that are evolving over time. For example, when the "application/postscript" type was originally proposed, there was one version of Postscript. Now, there are two levels. The GIF specification for images has two versions and a third under development. A system element might be able to deal with some versions and not others. Many type specification systems do not explicitly allow for versioning.

Resources used

Many representations of documents implicitly rely on external resources to actually define the interpretation of the file(s) that comprise the document. Thus, a Postscript file also requires the definitions of the fonts that it names; a TeX or nroff file also requires the definitions of any macro packages it invokes. These resource definitions are often assumed implicitly in the environment rather than being called out separately. In the case where one wishes to externally identify the media type, it may be necessary to also name the resources assumed in a more explicit manner.

Preservation

The issues of preservation in the library community are of growing concern in other areas. Companies with large repositories of electronic documents are discovering that they have great difficulty accessing them over time, not just because the storage media has become obsolete, but also because conversion of old document formats to new is difficult, unreliable and time-consuming.

Open vs. Proprietary

In a number of cases, the definition of the document type is not available outside the package that produces the type. While it might be is reasonable within a limited context to define a document as a being a `WordPerfect file', without a preserved specification of the actual interpretation of WordPerfect files, this labeling may not be useful decades hence. This is especially true because over time, there may be many different versions, configurations for multiple platforms, or localizations for different countries.

Compound objects

In many cases, the object being cataloged, manipulated and described is a compound object: a sequential concatenation, a collection of independent documents, or a compound object with some items nested inside or referenced from others. Any system of externally labeling and describing the type of the objects in use must be able to deal with expression of the types of compounds.

Encapsulations

Some system representations are transformations of others. For example, the `compress' program applies LZW compression to an object, binhex is a mechanism used on Macintosh computers for encoding binary data in ASCII. A language for describing types of objects needs to be able to describe the `binhex of a compressed postscript file', that is, the various encapsulations of one format within another.

2.5 Searching

All of the systems employed in each of the application areas allow for some method of searching a large collection of documents for those of particular interest or relevance.

Search in Document Management systems

Most personal, workgroup and enterprise document management systems offer the ability to search not only the externally assigned document attributes but also the content of the document. However, since the nature of the attributes and the natural search parameters differ so widely, many systems allow for site configuration of search methods. The metadata for documents is generally entered by office workers, and the quality of that information may vary. Metadata derived from document context (user ID of creator, time of last modification, workflow system case assignment) and from the content (title of presentation derived from initial slide) is usually more reliable than that entered manually.

Search in Digital Libraries

If online information resources are to be as useful as libraries of books and stacks, a number of tasks that are simple in physical libraries need to be made simpler in the online world. Much of digital library focuses on the capabilities necessary to find the most relevant information for a user who comes to do a search. Since the repositories are assumed to be extremely large and have full content available for search, new methods are being explored. As libraries are moving from providing bibliographic search (for words in the title, abstract, author) to full-text search, the algorithms for full-text retrieval are being reexamined. In addition, there is much research and development into the ability to search libraries of images, sound and video by a variety of techniques.

Search in the Web

The web community relies on search of existing corpora from digital libraries or information publishers to provide a search access, primarily by using gateway functions (web pages that interface to search engines) as well as supporting WAIS and Z39.50 directly.

Some organizations are offering services to search the Internet, by traversing the known Internet web, gathering together the pages, and indexing them. The search capability is offered as a service, for a fee, as a demonstration of text retrieval capabilities or as a way of advertising other products and services[42].

Design issues in search

Whether in a digital library, a document management system or on the web, there are a number of common design issues in expressing search operations.

One fundamental choice, made differently by different applications, is whether search is expressed by a search language or by a programming interface or some combination. Search languages include SQL (originally designed for relational databases) or enhancements of it, intended to deal with full text search, geographic information, etc. For example, Documentum's DQL[42] is a query language extended with versioning. The WAIS system originally left the `question' as a full text (presumably English) query. On the other hand, interfaces such as DEN allow the programmer of an interface to construct a query using API calls, without an expression in a query language. This has several advantages; it allows for more extensibility than is generally found in predefine syntax, allows for the query to be expressed in non-textual terms and does not require a parser in the search engine.

Much effort in each domain is being placed on enhancing user interface systems to deal with multiple sources. When a user queries more than one database at a time, it is necessary to merge the results from those sources. If two search engines have quite different capabilities, however, it is difficult to know how to express a combined search in a simple manner. Also, if the query language allows the expression of capabilities that are not present in the search database, there is a conflict. Some systems attempt to gloss over this or return results that are only approximately what the original search entailed.

Most models of database query and search allow for a single call/return sequence, where a search produces a result set, and then the result set is sequentially accessed to get back individual documents. However, in many cases, searching a corpus is a time-consuming process. Advanced user interfaces allow better feedback on the operation of the system and the state of the search; in order to provide that feedback, though, the search engine needs to provide updates as to the state, and these updates from multiple sources need to be merged.

3. Opportunities for integration

The first part of this paper described the different applications for Document Management, Digital Libraries and the Web; the second part laid out some of the areas where the design considerations of components, standards and protocols for each application area are the same.

The boundaries between these separate domains are blurring. Most digital library projects are exploring ways of making their libraries available to the entire Internet community, usually in spite of the perceived limitations of the current suite of web protocols and standards. As enterprise boundaries become more flexible with corporate outsourcing, dynamic enterprise construction and the increasing use of the Internet in the commercial sector, there is growing pressure to blur the boundary between an enterprise and workgroup repositories and those accessible on the Internet. And as companies and workgroups build larger repositories of archival quality documents--beyond those useful only momentarily--the distinction between an enterprise document management repository and a digital library is being blurred.

There is an opportunity to merge the interfaces for systems originally intended for document management, digital libraries or deployment in the web, in a way that will allow for several kinds of synergy. More specifically, there are several near-term opportunities.

For example, those charged with building and maintaining an Internet presence for an organization are discovering that, with the growth of their site, they have a large collection of documents with interdependencies, and need tools to help them manage their sites. One possible scenario is to use a tool originally designed as a document management system as the back-end to a web site. The version management, check-in and check-out, access control features of the document management system can be used by the web development staff, while the results are exported to the world over the Internet. Some explicit support for this kind of operation has been announced by a handful of document management companies.

Because workgroup document management systems are designed to integrate with office applications, it would be useful, for those office workers, to also be able to access other resources in repositories, whether in online libraries or other kinds of Internet resources. This could be accomplished by connecting the document management standard interfaces with Internet services.

Another possibility is to extend current Internet protocols for the web access (HTTP and current browsers) to add protocol elements for document management, including check-out, check-in, and a more rigorous approach to document attribute management. This effort has also begun in some quarters.

Other combinations of these technology elements are also possible, as long as the protocols and system architecture of the systems are not architecturally incompatible. Bringing together document management, digital libraries and the web is an important goal.

References

[1]
The Document Management Guide, Interleaf, c.1994. <URL:http://www.ileaf.com/docman.html>

[2]
Paula Rooney, "PC document management catches eye of big business", *PC Week*, May 18, 1992, v9 p45.

[3]
Lisa Nadile, "Document-management standards pave `open' path", *PC Week*, v11, n28, p8(1), July 18, 1994.

[4]
J. Garris, "Digging through your data", *PC Magazine*, v13, n19, pNE1(4), Nov 8, 1994.

[5]
Open Document Management API (ODMA) 1.0 specification, WordPerfect Corporation. <URL:ftp://ftp.wordperfect.com/pub/wpapps/windows/odma/>

[6]
Document Enabled Networking, DEN 0.86 API Specification, DEN Special Interest Group, 1994. <URL:http://www.xerox.com/DEN/DEN.html>

[7]
S. Teague, "Document management standards group formed", *InfoWorld*, V17, n16, p16(1), April 17, 1995.

[8]
James H. Billington, "Libraries and the NII", *Delivering Electronic Information in a Knowledge-Based Democracy (DEIKBD)*. <URL:http://iitfcat.nist.gov:94/doc/Library.html>

[9]
Ed Fox, ed., *Source Book on Digital Libraries. Version 1.0*, December 6, 1993. <URL:http://fox.cs.vt.edu/DLSB.html>

[10]
Digital Libraries '94; Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries, College Station, Texas, June 19-21, 1994. <URL:http://atg1.wustl.edu/DL94/>

[11]
1994 Workshop on Digital Libraries: Current Issues, Rutgers University, May 18-20, 1994. <URL: >http://superbook.bellcore.com/DBRG/DL94/>

[12]
Special Issue on Digital Libraries, Communications of the ACM, April, 1995. <URL: >http://cs.brandeis.edu/CACM/CACM_apr95.html>

[13]
The USMARC Formats: Background and Principles, American Library Association, 1989. <URL:gopher://marvel.loc.gov:70/00/.listarch/usmarc/usmarc.pri>

[14]
Robert E. Kahn, *Deposit, Registration and Recordation in an Electronic Copyright Management System*, Corporation for National Research Initiatives, Reston, VA, August 1991.

[15]
John R. Garrett and Patrice A. Lyons, "Toward an Electronic Copyright Management System", *Journal of the American Society for Information Science*, 44(8):468-473, 1993. CCC 0002-8231/93/080468-06.

[16]

Handle Management System, CNRI, 1995. <URL:http://www.cnri.reston.va.us/home/cstr/handle-intro.html>

[17]
Robert Kahn and Robert Wilensky, *Architecture of the Digital Library: Accessing Digital Library Services and Objects: A Frame of Reference (Draft 4.4 for discussion purposes, February 2, 1995).* <URL:http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>

[18]
Z39.50 Implementors Group Minutes, 1992-1995. <URL:http://ds.internic.net/z3950/minutes.html>

[19]
Kurt Piersol, "A Close-Up of OpenDoc", *BYTE*, March 1994. <URL:http://www.austin.ibm.com/developer/aix/library/aixpert/june94/aixpert_june94_closeup.html>

[20]
Jed Harris and Ira Ruben, *Bento Specification, Revision 1.0d5*, July 15, 1993. <URL:http://www.cilabs.org/pub/cilabs/tech/bento/Bento-Spec/postscript/>

[21]
Kraig Brockschmidt, *OLE Integration Technologies*, Microsoft Corporation, 1994. (adapted from article Dr. Dobbs Journal, December, 1994.) <URL:http://www.microsoft.com/pages/services/technet/ddjole.htm>

[22]
Proceedings of the Seminar on Cataloging Digital Documents, University of Virginia, Charlottesville, October 12-14, 1994. <URL:http://lcweb.loc.gov/catdir/semdigdocs/seminar.html>

[23]
Steve Putz, *Design and Implementation of the System 33 Document Service*, Xerox PARC P93-00112, 1993. <URL:http://www.xerox.com/PARC/dlbx/other-papers/system33.ps>

[24]
Stan Green, Keith Moore, and Reed Wade, *Bulk File Distribution.* <URL:http://www.netlib.org/nse/bfd/>

[25]
Mapping the Dublin Core Metadata Elements to USMARC, Discussion paper no. 86, Library of Congress, May 5, 1995. <URL:gopher://marvel.loc.gov/00/.listarch/usmarc/dp86.doc>

[26]
Hunter Moore, *Alex: A Catalog of Electronic Texts on the Internet*, July, 1994. <URL:gopher://vega.lib.ncsu.edu/00/library/stacks/Alex/About%20Alex >

[27]
Z39.50-1994 Information Retrieval: Application Service Definition and Protocol Specification, completed preliminary ballot draft, ANSI/NISO, August 1994. <URL:http://ds.internic.net/z3950/z3950.html>

[28]
C. M. Sperberg-McQueen and Lou Burnard, eds. *Guidelines for Electronic Text Encoding and Interchange*, May 16, 1994. <URL:http://etext.virginia.edu/TEI.html>

[29]
Jill Foster, ed., *A Status Report on Networked Information Retrieval: Tools and Groups*, RFC 1689, FYI 25, Internet Engineering Task Force, August 1994. <URL:ftp://ds.internic.net/rfc/rfc1689.txt>

[30]
David H. Crocker, *Standard for the Format of ARPA Internet Text Messages*, RFC822, Internet Engineering Task Force, August 13, 1981. <URL:ftp://ds.internic.net/rfc/rfc822.txt>

[31]
Roy Fielding, *IETF Uniform Resource Identifiers (URI) Working Group (home page)*, 1995. <URL:http://www.ics.uci.edu/pub/ietf/uri/>

[32]
E. Resclora, A. Schiffman, *The Secure HyperText Transfer Protocol (work in progress)*, December

1994. <URL:ftp://ds.internic.net/internet-drafts/draft-rescorla-shttp-00.txt>.; See also <URL:http://www.eit.com/projects/s-http/faq.html>.

[33]
Kipp E.B. Hickman, *The SSL Protocol (work in progress)*, April 1995. <URL:ftp://ds.internic.net/internet-drafts/draft-hickman-netscape-ssl-00.txt>; see also <URL:http://home.netscape.com/newsref/std/SSL.html>.

[34]
Dave Raggett, *Mediated Digest Authentication*, March 27, 1995. <URL:http://www.ics.uci.edu/pub/ietf/http/draft-ietf-http-mda-00.txt>

[35]
Jeffery L. Hostetler, John Franks, Phillip Hallam-Baker, Ari Luotonen, Eric W. Sink, Lawrence C. Stewart. *A Proposed Extension to HTTP: Digest Access Authentication*, March 23, 1995. <URL:http://www.ics.uci.edu/pub/ietf/http/draft-ietf-http-digest-aa-01.txt>

[36]
Simson Garfinkel, *PGP: Pretty Good Privacy*, O'Reilly & Associates, Inc. ISBN: 1-56592-098-8, December, 1994. See also <URL:http://www.ifi.uio.no/~staalesc/PGP/home.html >.

[37]
J. Linn, *Privacy Enhancement for Internet Electronic Mail* RFC 1421, Internet Engineering Task Force, February 1993. <URL:ftp://ds.internic.net/rfc/rfc1421.txt>

[38]
Electronic Cash, Tokens and Payments in the National Information Infrastructure, XIWT (Cross-Industry Working Team), Reston, Virginia, 1994. <URL:http://www.cnri.reston.va.us:3000/XIWT/documents/arch_doc/title_page.html>

[39]
Michael Lesk, *Preservation of New Technology*, Commission on Preservation and Access, Washington, D.C., October, 1991. <URL:gopher://palimpsest.stanford.edu:70/00/ByOrg/CPA/Reports/lesk.preservation.new.technology.txt>. See also <URL:http://www.cpa.org>

[40]
Task Force on Archiving of Digital Information (Web page). <URL:http://www.oclc.org:5046/~weibel/archtf.html>

[41]
Jeff Rothenberg, "Ensuring the Longevity of Digital Documents", *Scientific American*, January, 1995. See also <URL:http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/index.html>.

[42]
Glyn Moody, "Get crawlers to do your hunting through the Web", *Computer Weekly*, p43(1), March 2, 1995. See also <URL:http://asearch.mccmedia.com/embed.html> for a list of web search tools.

[43]
"Using DQL", in *The Documentum Server User's Guide*, Documentum, Inc.,Pleasanton, CA, 1995.

Author information

Dr. Masinter is a principal engineer at the Xerox Palo Alto Research Center. He has been working in the area of document management system architecture since 1988, the Web standards groups from their inception, and the research area of Digital Libraries since 1993.

Acknowledgements

Thanks to Geoff Nunberg, Marti Hearst, Carl Hauser, Ken Pier, Emil Rainero, Bill Anderson, Bill Crocca,

Ron Kaplan, Geoffrey Sejourne, David Elliott, Mary Ellen Zurko and Andreas Paepcke for their help with this paper.