Moving from Isolated Digital Collections to Interoperable Digital Libraries

Howard Besser
Associate Professor
UCLA School of Education & Information Studies
besser@ucla.edu

Abstract:
Online collections do not yet function like conventional libraries. Many digital collections are experimental and lack service components, and few have preservation components. The function of searching across collections is a dream frequently discussed but seldom realized at a robust level. This paper discusses how we might move from isolated digital collections to interoperable digital libraries. It first examines how early efforts to construct digital collections were perceived as experiments rather than operational libraries. It then discusses various conventional library components that are necessary to deployment of operational digital libraries. Finally, the author points to functions (such as infrastructure, robust metadata, and preservation components) that can be deployed to move us from isolated digital collections to interoperable digital libraries.

What is a Library?

As we try to plan for digital libraries, it is useful to review the roles and functions of traditional libraries. Writing about public libraries, McClure has outlined a critical set of roles they fulfill (McClure 1987) including: community activities center, community information center, formal education support center, independent learning center, popular materials library, preschoolers' door to learning, reference library, and research center. This set of roles needs to be rethought in an age when physical location and service can be separated from one another; some of these roles are more tied to the library's physical presence in the community, while others may function very well if delivered from remote sites. Besser has tried to update McClure's roles for the digital age (Besser 1998), claiming that the four core missions of a public library are: that it is a physical place, that is a focus spot for continuous educational development, that it has a mission to serve the underserved, and that it is a guarantor of public access to information. But we need to go beyond public libraries to make some generalizations that can apply to most types of libraries.

Traditionally, libraries have been more than just collections. They have components (including service to a clientele, stewardship over a collection, sustainability, and the ability to find material that exists outside that collection)[1] as well as traditions (including free speech, privacy, and equal access).

Almost all conventional libraries have a strong service component. All but the smallest libraries tend to have a substantial "public service" unit. Library schools teach about service (from "public service" courses to "reference interviews"). And the public in general regards librarians as helpful people to whom they turn to meet their information needs.

Many libraries deliver information to multiple clienteles. They are very good at using the same collection to serve many different groups of users, each group incorporating different modalities of learning and interacting, different levels of knowledge of a certain subject, etc. Public libraries serve people of all ages and professions, from those barely able to read, to high schoolers, to college students, to professors, to blue collar workers. Academic libraries serve undergraduates who may know very little in a particular field, faculty who may be specialists in that field, and non-native English speakers who may understand detailed concepts in a particular domain, but have difficulty grasping the language.

Most libraries also incorporate the component of stewardship over a collection. For some libraries, this is primarily a matter of reshelving and circulation control. But for most libraries, this includes a serious preservation function over at least a portion of their collection. For research libraries and special collections, preservation is a significant portion of their core responsibilities, but even school, public, and special libraries are usually responsible for maintaining a core collection of local records and works over long periods of time.

Libraries are organizations that last over long periods of time. Though occasionally a library does "go out of business", in general, libraries are social entities that have a great deal of stability. Though services may occasionally change in slight ways, people rely on their libraries to provide a sustainable set of services. And when services do change, there is usually a lengthy period where input is solicited from those who might be affected by those changes.

Another key component of libraries is that each library offers the service of providing information that is not housed within that library. Libraries see themselves as part of a networked world of libraries that work together to deliver information to an individual (who may deal directly only with his or her own library). Tools such as union catalogs and services such as inter-library loan have produced a sort of interoperable library network that was able to search for and deliver material from afar long before the advent of the WorldWide Web.

Libraries also have strong traditions. These include fervent protection of readers' privacy, equal access to information, diversity of information, serving the underserved, etc.

The library tradition of privacy protection is very strong. Librarians have risked serving jail time rather than turn over whole sets of patron borrowing records. Libraries in the US have even designed their circulation systems to only save aggregate borrowing statistics; they do not save individual statistics that could be later data-mined to determine what an individual had borrowed.

Librarians believe strongly in equal access to information. Librarians traditionally see themselves as providing information to those who cannot afford to pay for that information on the open market. And the American Library Association even mounted a court challenge to the Communications Decency Act because it prevented library users from accessing information that they could access from venues outside the library. Librarians have been in the forefront of the struggle against the privatizing of US government information on the grounds that those steps would limit the access of people who could not afford to pay for it.

Librarians also have a strong tradition of assuring diversity of information. Libraries purposely collect material from a wide variety of perspectives. Collection development policies often stress collection diversity. And librarians pride themselves on being able to offer patrons a rich and diverse set of information.

As we move towards constructing digital libraries, we need to remember that libraries are not merely collections of materials. They have both services and traditions that are a critical part of the functions they serve. The digital collections we build will not truly be digital libraries until they incorporate a significant number of these services and traditions.

Brief Digital Library History

The first major acknowledgement of the importance of Digital Libraries came in a 1994 announcement that $24.4 million of US federal funds would be dispersed among 6 universities for "digital library" research (NSF 1994). This funding came through a joint initiative of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA). The projects were at Carnegie Mellon University, the University of California-Berkeley, the University of Michigan, the University of Illinois, the University of California-Santa Barbara, and Stanford University.

These six well-funded projects helped set in motion the popular definition of a "digital library". These projects were computer science experiments, primarily in the areas of architecture and information retrieval. According to an editorial in D-Lib Magazine, "Rightly or wrongly, the DLI-1 grants were frequently criticized as exercises in pure research, with few practical applications" (Hirtle 1999).

Though these projects were exciting attempts to experiment with digital collections, in no sense of the word did they resemble libraries. They had little or no service components, no custodianship over collections, no sustainability, and no traditions. But because they were the first group to receive such widespread acknowledgement under the term "digital library", they set a popular impression for that term that persisted for many years. By 1996, social scientists who had previously worked with conventional libraries began trying to widen the term "digital libraries" (Bishop & Star 1996; Borgman et. al. 1996), but the real breakthrough came in late 1998 when the US federal government issued their highly funded DL-2 awards (Griffin 1999) to projects that contained elements of traditional library service, such as custodianship, sustainability, and relationships to a community of users. Around that time, traditional libraries began building serious digital components.

As librarians and social scientists became more involved in these digital projects, we moved away from computer science experiments into projects that were more operational. By the late 1990s, particularly under the influence of the US Digital Library Federation, projects began to address traditional library components such as stewardship over a collection and interoperability between collections. But even though these issues are finally being addressed, they are far from being solved. Though we have made great progress on issues such as real interoperability and digital preservation, these are far from being solved in a robust operational environment. In order to really call these new entities "digital libraries", we will need to make much more progress in moving analog library components such as sustainability and interoperability into the digital realm. And we need to begin to seriously address how we can move our library traditions (such as free speech, privacy, and equal access) into the digital realm as well. The remainder of this paper examines important efforts to move us in those directions.

Moving to a more user-centered architecture

Both the early computer science experiments in digital libraries and the earlier initial efforts to build online public access catalogs (OPACs) followed a model similar to that in figure #1. Under this model, a user needed to interact with each digital repository independently, to learn the syntax supported by each digital repository, and to have installed on their own computer the applications software needed to view the types of digital objects supported by each digital repository.

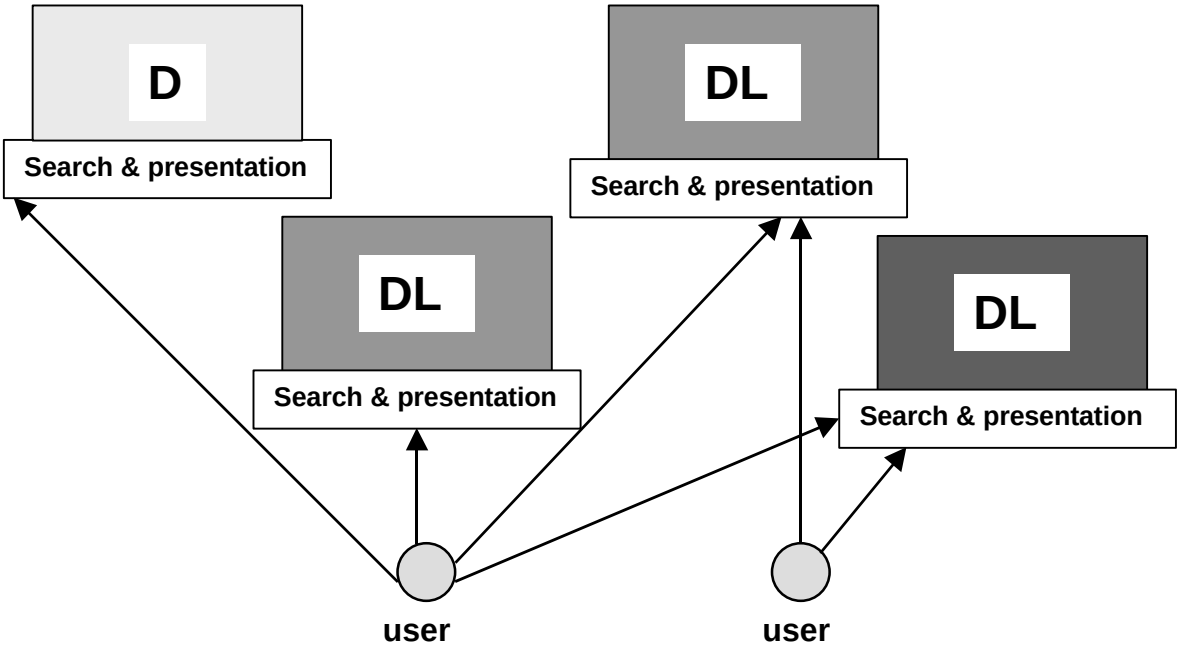

Figure #1

So, in order for a user to search Repository A, s/he would need to first adjust to Repository A's specialized user interface, then learn the search syntax supported by this repository. (For example, NOTIS-based OPACs required search syntax like ***A=Besser, Howard***, while Inovative-based OPACs required search syntax like ***FIND PN Besser, Howard***.) Once the search was completed, s/he could retrieve the appropriate digital objects, but would not necessarily be able to view them. Each repository would only support a limited number of encoding formats, and would require that the user have specific software installed on their personal computer (such as viewers for Microsoft Word 98, SGML, Adobe Acrobat, TIFF, PNG, JPEG, or specialized software distributed by that repository) in order to view the digital object. Thus users might search and find relevant works, but not be able to view them.

The user would then have to repeat this process with Repository B, C, D, etc., and each of these repositories may require a different syntax and different set of viewers. Once the user searched several different repositories, they still could not examine all their retrieved objects together. There was no way of merging sets. And because different repositories supported different viewing software, any attempt to examine objects from several repositories would likely require going back and forth between several different applications software used for display.

Obviously, the model in Figure #1 was not very user-friendly. Users don't want to learn several search syntaxes, they don't want to install a variety of viewing applications on their desk, and they want to make a single query that accesses a variety of different repositories. Users want to access an interoperable information world, where a set of separate repositories looks to them like a single information portal. A more user-friendly model is outlined in Figure #2. Under this model, a user makes a single query that propagates across multiple repositories. The user must only learn a single search syntax. The user doesn't need to have a large number of applications software installed for viewing. And retrieved sets of digital objects may be looked at together on the user's workstation. The model in Figure #2 envisions a world of interoperable digital repositories, and is a model we need to strive for.

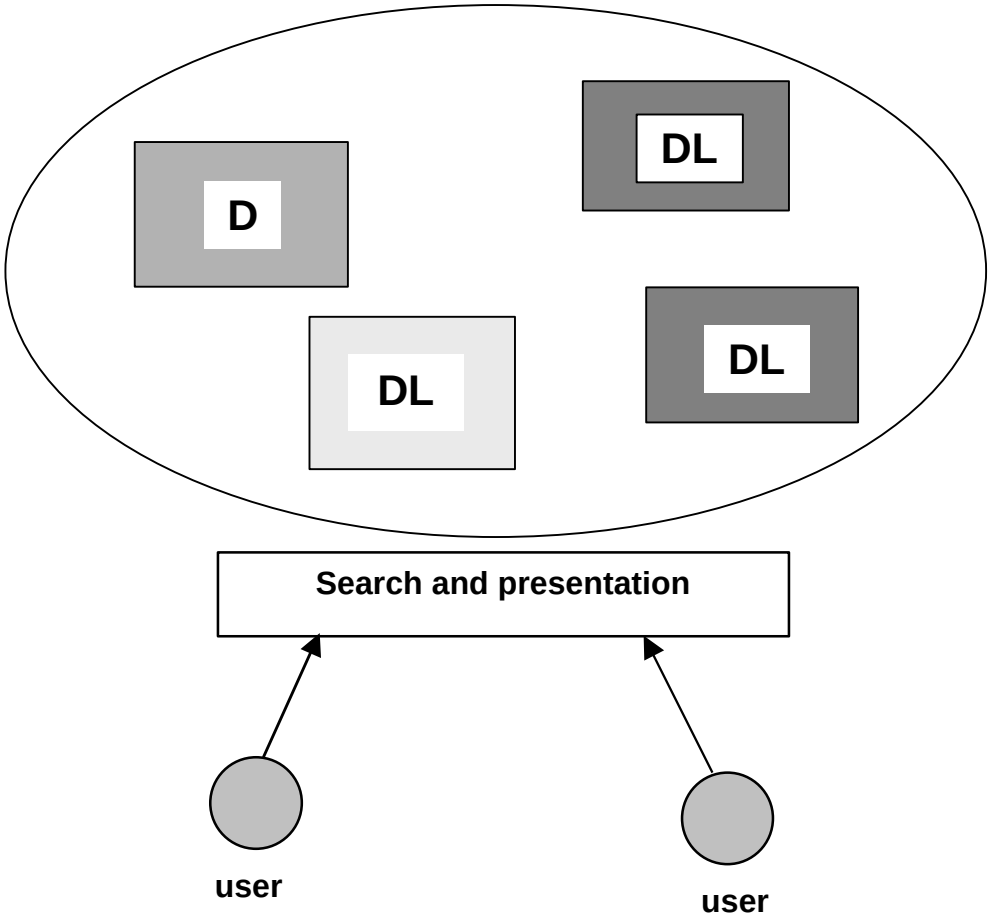

Figure #2

Over the years, we have made some significant progress towards the Figure #2 model, particularly in the area of OPACs. Web browsers have given us a common "look-and-feel" between different repository user interfaces. The Z39.50 protocols have allowed users to employ a single familiar search syntax, even when the repository's native search syntax appears foreign. Z39.50 has also promised to let user queries propagate to different repositories. But when one leaves the world of OPACs and enters the world of digital repositories, much work still needs to be done to achieve real interoperability. Most of this work involves creation and adoption of a wide variety of standards: from standards for the various types of metadata (administrative, structural, identification, longevity), to ways of making that metadata visible to external systems (harvesting), to common architectures that will support interoperability (open archives).

General processes and stages of technological development

The automation of processes often follows a series of pragmatic steps as well as a series of conceptual stages.

Pragmatic implementation steps usually begin by using technology to experiment with new methods of performing some function, followed by building operational systems, followed by building interoperable operational systems. And at the later stages of this, developers begin trying to make these systems useful for users. We have seen this pattern (experimental systems to operational systems to interoperable systems to useful systems) repeat in the development of OPACs, Indexing and Abstracting services, and image retrieval. The automation of each of these has begun with experiments, followed by implementations that envisioned closed operational systems (with known bodies of users who needed to learn particular user interfaces and syntaxes to interact with the system), followed by implementations that allowed the user to more easily interact with multiple systems (and sometimes to even search across various systems). Today's "digital libraries" are not much beyond the early experimental stage, and need much more work to make them truly interoperable and user-centered.

The conceptual steps typically include first trying to replicate core activities that functioned in the analog environment, then attempt to replicate some (but not all) of the non-core analog functions, then (after being in use for some time) discovering and implementing new functions that did not exist within the previous analog environment. This final step is a major shift in terms of creating something different that makes good use of the new functional environment enabled by the new technology. So, for example, word processors were initially built as typewriters with storage mechanisms, but over time grew to incorporate functions such as spell-checking and revision-tracking, and eventually enabled very different functions (such as desktop publishing). Our early efforts at creating MARC records began as ways to automate the production of catalog cards, then moved to the creation of bibliographic utilities and their union catalogs, then to OPACs. Functionally, our OPACs began as mere replicas of card catalogs, then added boolean searching, then title-word searching capabilities, and now are poised to allow users to propagate distributed searches across a series of OPACs. Today's digital collections are not much past the initial stage where we are replicating the collections of content and cataloging that existed in analog form, and just beginning to add minor functions. In the future, we can expect our digital libraries to incorporate a variety of functions that employ the new technological environments in ways we can hardly imagine today.

The Importance of Standards

In moving from dispersed digital collections to interoperable digital libraries, the most important activity we need to focus on is **standards**.[2] This includes standards and protocols for open archives and metadata harvesting. But most important is the wide variety of metadata[3] standards we need. We need to widely employ descriptive metadata for consistent description, discovery metadata for finding works, administrative metadata for viewing and maintaining works, structural metadata for navigation through an individual work, identification metadata to determine that one has accessed the proper version of a work, and terms and conditions metadata for compliance with use constraints.

Having consensus over metadata and other standards is important for a variety of reasons. We need administrative and longevity metadata to manage digital files over time, to make sure we keep together all the necessary files, and to help us view these files when today's application software becomes unusable. Because of the mutability of digital works, we need standards to ensure the veracity of a work, and to help assure users that a particular work has not been altered, and is indeed the version of the work that it purports to be. And we need a variety of types of metadata and standards to allow our digital collections to interoperate, and to help users feel that they can search across groups of collections. One side benefit of reaching consensus over metadata that will be recorded in a consistent manner is that vendors will have an economic incentive to re-tool applications to incorporate this metadata (because they can spread their costs over a wide variety of institutions who will want to employ these standards).

The Various Metadata Types

Libraries have had agreements on metadata standards for many decades. The Anglo-American Cataloging Rules defined a set of descriptive metadata for bibliographic (and later, other) works, and the MARC format gave us a syntax for transporting those bibliographic records. Likewise, Library of Congress Subject Headings and Sears Subject Headings have for many years provided us with discovery metadata to help users find relevant material. In the last quarter of the 20th century, other types of discovery metadata emerged to serve specialized fields, including the Art and Architecture Thesaurus (AAT) and the Medical Subject Headings (MeSH).

Though both AAT and MeSH envisioned use in an online environment, both were developed in an era when indexing and cataloging records might sit on a computer, but that the works they referred to would not. And both were developed at a point in time when the merging of records from these specialized fields with records for more general works was unlikely to take place on a widespread basis.

The rapid acceptance of the WorldWide Web led a number of us to consider how one might allow users to search across a variety of online records and resources, particularly when some of those resources received extensive cataloging, while others received little or none. This led to the March 1995 meeting that defined the Dublin Core as a type of discovery metadata that would allow users to search across a wide variety of resources including both highly cataloged (often legacy) material, and material (much of it new and in electronic form) that was assigned only a minimal amount of metadata. We envisioned the Dublin Core as a kind of unifying set of metadata that would permit discovery across all types of digital records and resources. Library cataloging records or museum collection management records could be "dumbed down" to look like Dublin Core (DC) records, while DC records for resources like an individual's research paper might either be easy enough for the individual to create, or might even be automatically generated by the individual's word processor. The not-yet-realized promise of the Dublin Core (see next section on Harvesting) was to provide discovery-level interoperability across all types of online indexes and resources, from the highly-cataloged OPACs to the websites and webpages of individuals and organizations. And because the Dublin Core has been in existence for approximately 7 years, it is more developed and better known than any of the other types of metadata created for electronic resources.

Though the Dublin Core was developed as a form of digital metadata to be applied to works in both digital and non-digital form, a variety of other metadata types have more recently been developed specifically for collections of works in digital form. Below we will briefly discuss efforts to define structural metadata, administrative metadata, identification metadata (particularly for images), and longevity metadata. All these metadata types are critical for moving from a set of independent digital collections to real interoperable digital libraries. Hence they all incorporate either functions likely to lead to increased interoperability, or to the fuller and more robust services that characterize a library rather than a collection.

Structural metadata recognizes that, for many works in digital form, it is not enough to merely display the work; a user may need to navigate through the work. Structural metadata recognizes that users expect certain "behaviors" from a work. For example, imagine a book that is composed of hundreds of digital files, each one the scan of a single book page. Structural metadata is needed for a user to perform the normal behaviors s/he might expect from a book. The user will expect to be able to view the Table of Contents, then jump to a particular chapter. As they read through that chapter, they will expect to turn the page, and occasionally go back to re-read the previous page. When they come to a citation, they will want to jump to the bibliography to read the citation, then jump back. And when they come to a footnote marker, they may want to jump to where they can read the footnote contents, then jump back. These are all just normal behaviors we expect from any type of book, but these behaviors all require structural metadata. Without such structural metadata, the book would just be a series of individual scanned pages, and users would have a great deal of difficulty trying to even put the pages in the correct order, let alone read the book. Structural metadata has also been applied to a variety of other types of works that can benefit from internal navigation, including diaries and journals.

Administrative metadata maintains the information necessary in order to keep a digital work accessible over time. In the case of a digitized book, the administrative metadata would note all the individual files needed to assemble the book, where the files were located, and what file formats and applications software would be necessary in order to view the book or its individual pages. Administrative metadata becomes particularly important when moving files to a new server, or engaging in digital longevity related activities such as refreshing or migration.

Instead of employing open standards for structural and administrative metadata, many individuals and organizations choose to encode their documents within commercial products such as Adobe Acrobat. While this is highly convenient (particularly given the proliferation of Acrobat readers), it could be a dangerous practice for libraries and similar repositories to engage in. Commercial products are proprietary, and focus on immediate convenient access (rather than long-term document access). Hence, there is no guarantee of continued compatibility or future access to works encoded in earlier versions of commercial software. In order to cope with long-term preservation and access issues as well as to provide a higher level of structural functionality, in 1997 a group of US libraries began the Making of America II Project (Hurley et. al. 1999) to define structural and administrative metadata standards for library special collection material. These standards were further refined within the Technology Architecture and Standards Committee of the California Digital Library (CDL 2001a), and have since been renamed the Metadata Encoding and Transmission Standards (METS) and taken over by the US Digital Library Federation and are maintained by the US Library of Congress (http://www.loc.gov/standards/mets/).[4]

Identification metadata attempts to address the proliferation of different versions and editions of digital works. In the print world, the publishing cycle usually both enforced an editorial process and created time lags between the issuance of variant works. But for a digital work which is highly mutable, often those processes and time lags are eliminated, and variants of the work are created quickly and with very little thought about their impact on what we used to call bibliographic control. In addition, the networked digital environment itself leads information distributors to provide a variety of different forms of a given work (HTML, postscript, acrobat, XML, and Microsoft Word forms of documents to support different user capabilities and needs; thumbnail, medium-sized, and large images to support image browsing, viewing, and study).

To illustrate the identification metadata problem, let us turn to Figure #3, an illustration of variant forms of images. The original object (a sheep) is shown in the upper left corner. There are 4 photographs of the original object, taken from 3 different angles. Two of the photographs (A and D) are taken from the same angle, but photograph D has captured a fly on the side of the sheep. The images to the right of photograph D are variant forms of that photograph after image processing was done to remove the fly from the side of the sheep, while the images below photograph D show variant forms that include the fly. Spread throughout the figure are variant forms including different sized images (thumbnail to high resolution), different compression ratios (including both lossy and lossless) and different file encoding formats (PICT, TIFF, JFIF). Certain users may only want particular resolutions or compression ratios (e.g. a serious researcher may require an uncompressed high resolution image). All the images illustrated in this figure share a base set of metadata that refers to the initial object of origin (the sheep), and adapting Leazer's idea of Bibiliographic Families (Leazer & Smiraglia 1999), we can say that all the images in this illustration form an "Image Family" which shares a common set of metadata. Each image instantiation in the family also inherits metadata from its parents, and knowledge about that inheritance can often be critical to someone viewing a particular instantiation (for example, someone using one of the lower-right corner instantiations to study wool characteristics should be able to ascertain that image processing was done on a parent or grandparent of this image [to remove the fly], and that this image processing might affect the matting of the wool). Thus, it is critical that any image inherits important metadata from its lineage, or that systems at least provide ways that a researcher can trace upwards in lineage to discover metadata that might affect their use of the work. A first step in this direction is in the US National Information Standards Organization efforts at creating Technical Imaging Metadata Standard that incorporates "change history" and "source data" (NISO). But our community has much more work to do in order to provide the type of identification of variant forms that users (particularly researchers) have come to expect from libraries of analog materials. And we still need to come to grips with the problem of how to preserve dynamic documents -- documents that are essentially alive, and changing on a daily basis.

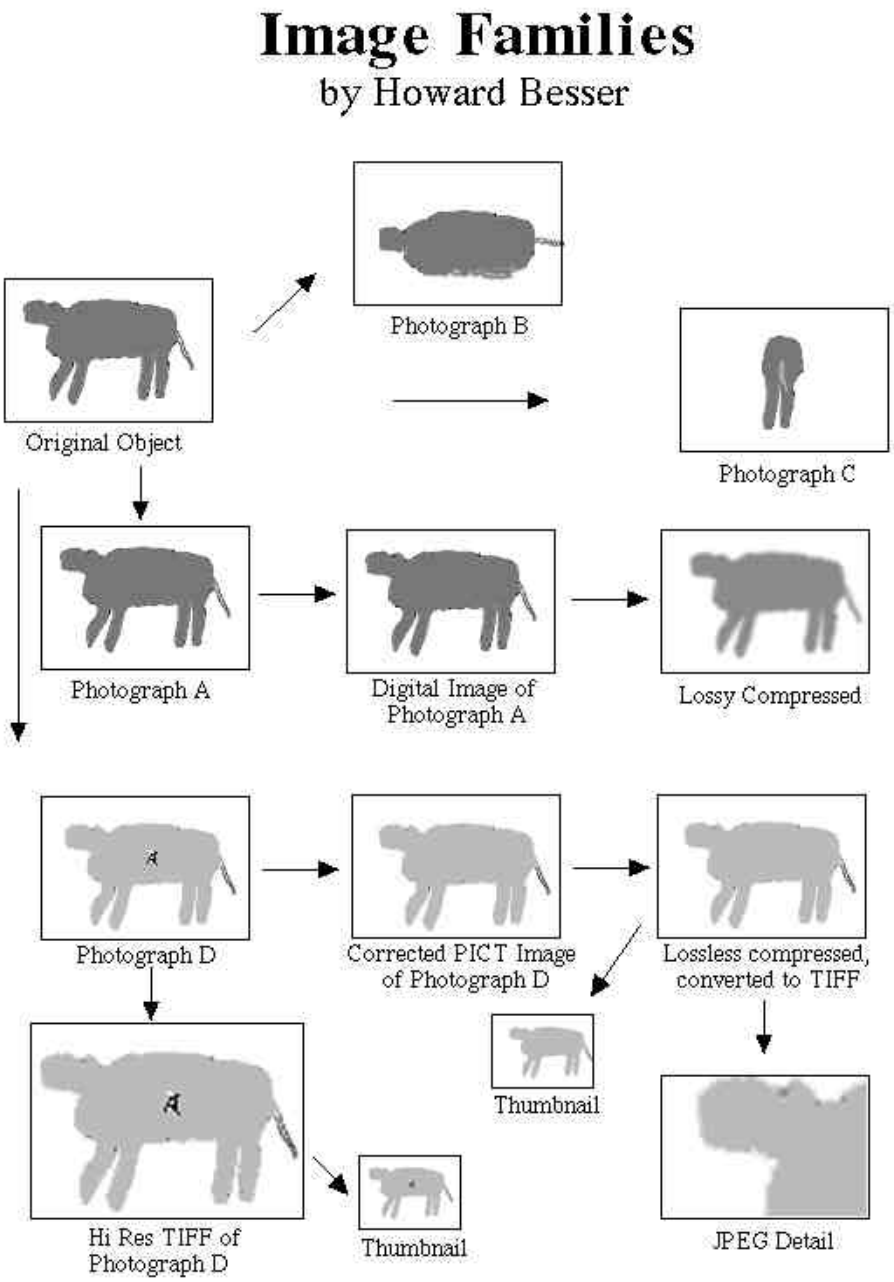

Figure #3

As we construct large collections of material in digital form, we need to consider how digital works will provoke changes in long-standing practices that have grown up around analog works. Elsewhere this author has outlined how electronic art will likely provoke changes in conservation and preservation practices (Besser 2001a) and how the growing body of moving image material in digital form is beginning to reshape the role of film archives and archivists (Besser 2001b). But at a very pragmatic level, all repositories of digital works need to worry about the persistence of those works over time. Longevity metadata is necessary in order to keep digital material over long periods of time. While saving bits may be fairly straightforward, saving digital works is not. Digital works are very fragile, and pro-active steps need to be taken in order to make sure that they persists over time. This author has also outlined five key factors that pose digital longevity challenges (the Viewing Problem, the

Scrambling Problem, the Inter-relation Problem, the Custodial Problem, and the Translation Problem), and has suggested that community consensus over metadata can be a key factor in helping digital works persist over time (Besser 2000a). Recently, the major US bibliographic utilities have begun serious efforts to reach consensus on preservation metadata for digital works (OCLC/RLG 2001a, 2001b). Widespread adoption of this type of standard will make the challenge of digital persistence much more tractable. And late in 2001 the Library of Congress, with the help of the Council on Library and Information Resources, began a planning process for a National Digital Information Infrastructure and Preservation Program.

Widespread support for the emerging metadata standards mentioned here will greatly improve interoperability between collections and sustainability of those collections over time. This will help us move away from isolated experiments with digital collections towards sustainable digital libraries.

Metadata Philosophies and Harvesting: Warwick vs. MARC

Though agreement on a variety of metadata standards is a necessary prerequisite for interoperable digital collections, implementation of interoperability also requires a set of architectures and a common approach to making that metadata available to other collections, middleware, and end users. In this section we will discuss two philosophical approaches to metadata, as well as methods for sharing metadata with applications and individuals outside the collection.

Libraries have traditionally employed the MARC/AACR2 philosophical approach to metadata. This approach employs a single over-arching schema to cover all types of works and all groups of users. As new types of works arise, new fields are added to the MARC/AACR2 framework, or rules for existing fields are changed to accommodate these new works. And as communities emerge with new metadata needs, these are also incorporated into the existing schema. The MARC/AACR2 philosophy maintains that one big schema should serve all user needs for all types of works. Critics of this approach point out that the schema has become so overly complex that only highly trained specialists (librarians) are able to assign metadata using it, and that the system is too slow to adapt to emerging types of works. They also claim that groups of users often have sets of metadata needs that the controllers of MARC/AACR2 are unwilling to accommodate.

In recent years, a rival philosophy has emerged from within the Dublin Core community. This philosophy, based upon the Warwick Framework, relies upon interlocking containers and packages of metadata, each maintained by a particular community. According to this philosophy, each community can support the packages of metadata it needs for its own particular uses, while still interoperating with the metadata packages from other communities. In this philosophy, the Dublin Core serves as a unifying set of metadata to allow discovery across all communities. And even within the Dublin Core (DC), certain communities can employ qualifiers that meet their own detailed needs, while still providing useful metadata to other communities. (For example, the library community could use qualifiers to reflect the nuances of differences between main title, alternate title, transliterated title, and translated title, while other communities could find any of these as part of a search under unqualified title.) This philosophy supports metadata packages that are modular, overlapping, extensible, and community-based. Advocates believe that they will aid commonality between communities while still providing full functionality within each community. This approach is designed for a networked set of communities to inter-relate to one another.

No matter which philosophical approach one follows, any collection faces the pragmatic issue of how to make their metadata available to other collections and external searching software. The traditional library model was to export MARC records to a bibliographic utility (like OCLC or RLIN) and to have all external users search through that utility. While this works fine for MARC-based records, increasingly users want to search across a much wider base of information from a world not circumscribed by bibliographic records. Therefore, most digital collections are beginning to consider how to export reduced records into a space where they can be picked up by Internet search engines. For records following the MARC/AACR2 approach, this means extracting simple records (probably in DC format) from complex MARC records, and exporting these. For both the Warwick and the MARC/AACR2 approaches, this means developing methods for metadata harvesting that allow the appropriate exported records to be found by Internet search engines. A number of projects are currently underway to test metadata harvesting.

Best Practices

Along with standards and architectures, community agreement on Best Practices is another important ingredient in helping make collections of digital materials more interoperable and sustainable. Best practices assure that content and metadata from different collections will meet minimum standards for preservation purposes, and that users can expect a baseline quality level.

A key best practice principle is that any digital project needs to consider users, potential users, uses, and actual characteristics of the collections (Besser & Trant 1995). This means that decision-making both on digital conversion of analog material and on metadata assignment needs to be carefully planned at the start of a digital project. The pioneering best practices for digital conversion developed by the Technology Architecture and Standards Committee of the California Digital Library (CDL 2001c) introduced important concepts designed to aid in the longevity and sustainability of digital collections. These concepts included differentiating between masters and derivatives, inclusion of greyscale targets and rulers in the scan, using objective measurements to determine scanner settings (rather than matching the image on a nearby monitor to the original object), storing in common formats, and avoiding compression (particularly lossy compression). This document also suggested that collections strive to capture as much metadata as is reasonably possible (including metadata about the scanning process itself). The relationship of scanning best practices to longevity of digital works is more fully explained in the Digital Library Federation's draft benchmarks for digital reproductions (DLF 2001).

The Making of America II Project (Hurley et. al. 1999) introduced the idea that metadata could begin in fairly raw form, and, over time, move toward being seared and eventually cooked. This notion of incrementally upgrading metadata appears to have assuaged the fears of some groups that metadata schemes like METS were too overblown and complicated for them to undertake. In effect, this notion appears to have increased the adoption level of more complicated metadata schemes.

Other standards issues

A number of other standards issues need to be addressed in order to bring interoperability and other conventional library services to our emerging digital libraries. These include open archives, metadata harvesting, persistent identification, helping users find the appropriate copy, and user authentication.

Metadata stored in local systems is often not viewable by external applications or users that may be trying to discover local resources. This seriously inhibits interoperability, and the ability of a user to search multiple collections. The Open Archives Initiative (http://www.openarchives.org/) is tackling this problem by developing and testing interoperability protocols that will allow applications to harvest metadata, even that residing in deep archives. The success of a project like this is critical to providing users with the type of access outlined in Figure #2. In addition, this project's focus on e-print archives should provide users with a diverse body of content free of onerous constraints.

Persistent naming is still an important issue for building real digital libraries. Though the WorldWide Web has brought us increased access to works, Web architecture has violated traditional library practices of providing relative location information for a work by instead providing a precise location address. The Web's precise location addressing system (the URL) has led to the most common error message that Web users experience (404--File Not Found). Most of these error messages result from normal maintenance of a Website (renaming higher-order folders or directories, re-organizing file locations). Librarians would never consider telling a user that to find the book they're seeking they must go to the third tier of the stacks, in the 8th row, the 5th bookcase, the 3rd shelf, and grab the 7th book from the left; they know that once someone removes the 3rd book from the left, the entire system of locating will break down. Yet, this is the type of system that URLs are based upon. In recent years there has been much work done on indirect naming (in the form of PURLS, URNs and handles). But to replicate the power that libraries have developed, we need truly persistent naming. This means more than just the indication of a location for a particular work. Sophisticated persistent naming would include the ability to designate a work by its name, and to distinguish between various instantiations of that work and their physical locations. Just as conventional libraries are able to handle versions and editions and direct users to particular copies of these, our digital libraries will need to use identification metadata to direct users to an appropriate instantiation of the work they are seeking.

Ever since the advent of indexing and abstracting services, conventional libraries have had to face the problem of answering a user's query with a list of works, some of which may not be readily available. Conventional libraries have striven to educate users that some sources are not physically present in the library, and have developed both interlibrary loan and document delivery services to help get material to users in a timely fashion. But the advent of licensed full-text electronic resources has greatly complicated this problem. It may be very difficult to match a user with an appropriate document they are licensed to use for a variety of reasons: certain users may be covered by a given license while others are not; the same document may be provided by several aggregators under different licenses; much of the online content is physically stored by the licensor (content provider) rather than by the licensee (library). Recent standardization efforts have begun to address part of this problem; the US National Information Standards Organization has formed the OpenURL Standard Committee AX (http://www.niso.org/commitax.html) to allow a query to carry context-sensitive information. This will help a library authenticate their licensees to a remote content site. But there are still

many more complex problems to solve in getting users appropriate copies (particularly when content licenses are complex and overlapping).

Still another critically important standards and protocols area that is under development is that of authentication of users. With more and more licensed content being physically stored by content providers, those providers want assurances that users accessing the content are indeed covered by valid licenses to do so. Yet conventional methods of user authentication (such as password or IP addressing) would allow content providers to track what an individual reads and develop complex profiles of user habits. Legal scholars have warned of the dangers this poses (Cohen 1996), and it flies in the face of the important library tradition of privacy. Work has begun on a project that lets an institution authenticate users to a resource provider without revealing individual identities. It still remains to be seen whether the Shibboleth project (http://middleware.internet2.edu/shibboleth/) will be acceptable to resource providers, yet still provide the privacy and anonymity that libraries have traditionally assured users. Success becomes more questionable in the wake of the September 11 US building destructions, as the US federal government has increased the pressure to eliminate anonymous library access (ALA 2001).

Moving from Isolated Digital Collections to Interoperable Digital Libraries

Conventional libraries have both components and traditions. The digital collections we are constructing will not truly be "digital libraries" until they incorporate a significant number of the components of conventional libraries, and adhere to many of the important traditions of libraries. And though our digital collections have made significant progress in these areas in the past 7 years, they still have a long way to go.

For the component of interoperability, projects such as open archives, metadata harvesting, and structural and administrative metadata hold great promise. Within the component of stewardship over collections, digital preservation projects have finally begun, but face a huge challenge before we will be able to confidently say that we can preserve the portion of our culture that is in digital form. Additionally, we have only recently begun to grapple with the issue of economic sustainability for digital libraries (CLIR 2001). For other components such as service to clientele, we have barely scratched the surface in one important area that conventional libraries do very well -- deliver information to different groups of users (by age level, knowledge base, particular need, etc.) in ways that are most appropriate to their group.[5]

Those constructing digital collections have spent less energy trying to build library traditions into our systems, and in many cases have relied on those outside the digital library community[6] to work on upholding library traditions such as free speech, privacy, and equal access. But as Lessig has made clear, the choices we make in the architecture and design of our systems will limit the social choices we can make around use of those systems in the future (Lessig 1999). For example, some of our online public access circulation systems purposely only saved aggregate user data so that no one in the future could attempt to track individual reading habits. While recent project such as Shibboleth are trying to design library traditions into the technological infrastructure we build, the digital libraries we're building still have not addressed many of our important library traditions.

As we build our digital collections we also need to uphold library traditions of equal access and diversity of information. Both of these are threatened by the commercialization of intellectual property. As we see the increased commodification of information and consolidation of the content industry into fewer and fewer hands, less and less works enter the public domain and more require payment to view them (Besser forthcoming). Commodification and consolidation also bring with them a concentration on "best-seller" works and a limiting of diversity of works (Besser 1998; Besser 1995). The builders of digital collections need to go beyond content that is popular and become aggressive about collecting content that reflects wide diversity; instead of the opportunistic approach that has characterized decision-making over content to convert to digital form, we need to develop carefully planned digital collection development policies. We also need to involve ourselves in efforts to assure that we and our users will have continued access to a broad set of content that eventually leaves the realm of the marketplace and enters the public domain. We need to involve ourselves in struggles like the American Library Association's current effort to build a coalition to protect the "information commons" in cyberspace.

We also need to continue to be vigilant about making sure that other forms of “equal access to information” extend to our digital world. We need to exert pressure to extend the “library bill of rights” into cyberspace, and we have to struggle to keep the digital world from having any greater divide than the analog world has between "haves" and "have-nots".

As we move towards constructing digital libraries, we need to remember that libraries are not merely collections of works. They have both services and traditions that are a critical part of their functions. Libraries interoperate with each other to serve the information needs of a variety of different user groups today, and expect to sustain themselves and their collections so that they can serve users 100 years from now. They defend their users' rights to access content, and to do so with some degree of privacy or anonymity. The digital collections we build will not truly be digital libraries until they incorporate a significant number of these services and traditions.

Citations

American Library Association et. al. (2001) "Library Community Statement on Proposed Anti-Terrorism Measures", October 2 (http://www.ala.org/washoff/)

Bishop, Ann and Susan Leigh Star (1996), "Social informatics for digital library use and infrastructure" in Martha Williams (ed.) ***Annual Review of Information Science and Technology***, 31, Medford NJ: Information Today, pages 301-401

Besser, Howard (forthcoming). "Commodification of Culture Harms Creators", http://www.gseis.ucla.edu/~howard/Copyright/ala-commons.html (to be published as part of a special "Information Commons" website by American Library Association)

Besser, Howard (2001a). "Longevity of Electronic Art", in David Bearman and Franca Garzotto (eds.) ***ICHIM 01 International Cultural Heritage Informatics Meeting: Cultural Heritage and Technologies in the Thrid Millennium***, Volume 1 - Full Papers (Proceedings of the September 3-7, 2001 Milan meeting), Milan: Politecnico di Milano, pages 263-275

Besser, Howard (2001b). "Digital Preservation of Moving Image Material", ***The Moving Image***, Fall

Besser, Howard (2000a). "Digital Longevity", in Maxine K. Sitts (ed.) ***Handbook for Digital Projects: A Management Tool for Preservation and Access***, Andover Mass: Northeast Document Conservation Center, pages 155-166

Besser, Howard (1998). "The Shape of the 21st Century Library", in Milton Wolf et. al. (eds.), ***Information Imagineering: Meeting at the Interface***, Chicago: American Library Association, pages 133-146

Bessser, Howard (1995). "From Internet to Information Superhighway", in James Brook and Iain A. Boal (eds.), ***Resisting the Virtual Life: The Culture and Politics of Information***, San Francisco: City Lights, pages 59-70

Besser, Howard and Jennifer Trant (1995), ***Introduction to Imaging: Issues in Constructing an Image Database***, Santa Monica: Getty Art History Information Program

Borgman, Christine (1997), "Now that we have digital collections, why do we need libraries?" in Candy Schwartz and Mark Rorvig (eds.), ***ASIS '97: Proceedings of the 60th ASIS Annual Meeting, Volume 34***, Medford NJ: Information Today

Borgman, Christine, et. al. (1996), "Social Aspects of Digital Libraries", Final report to the National Science Foundation
(http://dli.grainger.uiuc.edu/national.htm)

California Digital Library, Technology Architecture and Standards Committee (2001a), "California Digital Library Digital Object Standard: Metadata, Content and Encoding", May 18 http://www.cdlib.org/about/publications/

California Digital Library, Technology Architecture and Standards Committee (2001b), "California Digital Library Digital Image Format Standards", July 9 http://www.cdlib.org/about/publications/

California Digital Library, Technology Architecture and Standards Committee (2001c), "Best Practices for Image Capture", February
http://www.cdlib.org/about/publications/

California Digital Library, Technology Architecture and Standards Committee (2000), "California Digital Library Technical Architecture and Standards", June 16 http://www.cdlib.org/libstaff/technology/tas/Standards/

Cohen, Julie (1996), "A Right to Read Anonymously: A Closer Look at 'Copyright Management' in Cyberspace", ***Connecticut Law Review***, vol. 28, pp. 981-1039

Council on Library and Information Resources (CLIR) (2001) " Building and Sustaining Digital Collections: Models for Libraries and Museums ", Washington: CLIR, August (http://www.clir.org/pubs/reports/pub100/pub100.pdf)

Digital Library Federation (2001), "Draft benchmark for digital reproductions of printed books and serial publications", July 30 (webpage)
(http://www.diglib.org/standards/draftbmark.htm)

Digital Preservation and Archive Committee (DPAC) (2001) "Draft Final Report", October 8, 2001

Griffin, Stephen M. (1999*),* "Digital Libraries Initiative - Phase 2: Fiscal Year 1999 Awards", ***D-Lib Magazine***, 5:7-8, July/August
(http://www.dlib.org/dlib/july99/07griffin.html)

Hirtle, Peter (1999), "A New Generation of Digital Library Research" (editorial*),* ***D-Lib Magazine***, 5:7-8, July/August
(http://www.dlib.org/dlib/july99/07editorial.html)

Hurley, Bernard, John Price-Wilkin, Merrillee Proffitt, and Howard Besser (1999). ***The Making of America II Testbed Project: A Digital Library Service Model****,* Washington: Digital Library Federation, Council on Library and Information Resources, December (http://sunsite.berkeley.edu/moa2/wp-v2.html)

Leazer, Gregory and Richard Smiraglia (1999), "Bibliographic Families in the Library Catalog: A Qualitative Analysis and Grounded Theory", ***Library Resources & Technical Services 43***:4, October

Lessig, Lawrence (1999). ***Code and other laws of cyberspace****,* New York: Basic Books

McClure, Charles et. al. (1987) ***Planning and Role Setting for Public Libraries***, Chicago: American Library Association (figure #11).

National Information Standards Organization. "Technical Metadata for Digital Still Images, Standards Committee AU" (website)
(http://www.niso.org/commitau.html)

National Science Foundation (1994?), "NSF Announces Awards for Digital Libraries Research; $24.4 Million to Fund Advanced Research and Technology Development by Teams from Universities, Industries and Other Institutions", September (http://elib.cs.berkeley.edu:80/admin/proposal/nsf-press-release.html)

OCLC/RLG Working Group on Preservation Metadata (2001a*),* "Preservation Metadata for Digital Objects: A Review of the State of the Art", January 31 (http://www.oclc.org/digitalpreservation/presmeta_wp.pdf)

OCLC/RLG Working Group on Preservation Metadata (2001b), "A Recommendation for Content Information", October
(http://www.oclc.org/research/pmwg/contentinformation.pdf)

Endnotes

[1] In her discussion of why conventional libraries will not disappear simply because we develop online collections, Christine Borgman states that the convention library's role is to "select, collect, organize, preserve, conserve, and provide access to information in many media, to many communities of users" (Borgman 1997).

[2] Bill Arms has said that interoperability can be achieved on any of 3 levels: the technical, the content or by organizational agreement (or by a combination of these). In fact, all 3 of these involve standards.

[3] The most extensive metadata activities have focused on discovery metadata (such as the Dublin Core). But metadata also includes a wide variety of other functions: structural metadata is used for turning the pages of a digital book, administrative metadata is used to assure that all the individual pages of a digital book are kept together over time, computer-based image retrieval systems employ metadata to help users search for similar colors, shapes, and textures, etc.

[4] The Research Libraries Group recently announced plans to lead the METS development effort.

[5] The California Digital Library and the UCLA/Pacific Bell Initiative forf21st Century Literacies have recently begun a project to explore this problem (http://www.newliteracies.gseis.ucla.edu/design/).

[6] such as the American Library Association filing lawsuits on privacy and free speech, or the Internet Engineering Task Force developing protocols to preserve privacy