

## **An Overview of Digital Libraries: Issues and Trends**

Micah Altman\*  
Senior Research Scientist, IQSS, Harvard University  
1737 Cambridge Street  
Cambridge, MA 02138  
[Micah\\_Altman@harvard.edu](mailto:micah.altman@harvard.edu)

\* Corresponding Author  
[Word count -- main text: 3321 ]

### **INTRODUCTION**

Digital libraries are collections of digital content and services selected by a curator for use by a particular user community. Digital libraries offer direct access to the content of a wide variety of intellectual works, including text, audio, video, and data; and may offer a variety of services supporting search, access, and collaboration. In the last decade digital libraries have rapidly become ubiquitous because they offer convenience, expanded access, and search capabilities not present in traditional libraries. This has greatly altered how library users find and access information, and has put pressure on traditional libraries to take on new roles. However, information professionals have raised compelling concerns regarding the sizeable gaps in the holdings of digital libraries, about the preservation of existing holdings, and about sustainable economic models.

This article presents an overview of the history, advantages, and design principles relating to digital libraries, and highlights important controversies and trends. For an excellent comprehensive discussion of the use, cost and benefits of digital libraries see Lesk [2005], for further discussion of architectural and design issues see Arms [2000], and see Witten [2002] for a detailed example of the mechanics of implementing a digital library.

### **HISTORY**

In 1939, before the first digital computer system was designed, Vannevar-Bush, a professor of electrical engineering at MIT proposed a system that in many ways foreshadowed modern digital libraries. [Bush 1939, 1945] (Bush would become head of the Office of Scientific Research and Development during World War 2 and then one of the chief advocates for the creation of the National Science Foundation.) This system, the "Memex", was designed to microfilm entire libraries of books and journals, combine these with individuals' private notes and indexes, and make them available on the desktop. Bush envisioned that the Memex would enable users and information professionals to create new organizations of knowledge through 'associative trails', links among parts of different documents. Although this system was never built, Bush's ideas inspired generations of future computer scientists, including J.C.R. Licklider, who made fundamental contributions to the development of personal computer interfaces, artificial intelligence, the internet, and digital libraries. Licklider envisioned much of the design of modern digital libraries, including the integration of indexing, search, retrieval, and storage services. [Licklider 1965]

Although lacking the characteristic search and direct access capabilities of modern digital libraries, social science data archives were, in a sense, the first digital libraries, since they maintained and outside users access to large collections of digital material. Many of these collections were started in the 1950's when social scientists realized that it was crucial their research surveys (etc.), recorded in digital form, be preserved for future research. [Bisco 1970] In the 1970's through the late 80's, digital technology was adopted in most libraries, primarily in the form of OPAC's (online public access catalog), which replaced card catalogs. It was not until the early 1990's, when the burgeoning World-Wide-Web, made dramatically more useful by indexing services such as Lycos (one of the early and dramatic successes of Internet search), greatly accelerated the growth of digital libraries and brought the combination of access and content that is their modern hallmark.

Government funding was crucial to early developments in digital library technology, and continues to remain important. For example, the Lycos search engine emerged from work done by the Informedia project at Carnegie-Mellon, and the immensely popular Google search service emerged from Stanford's Interlib project. Both of these projects were initially funded under the Digital Library Initiative, a joint project of NSF, NASA, and DARPA. The two phases of this initiative sponsored some of the most innovative efforts in digital libraries across a decade. (see Griffin, 1998) Other U.S. government programs such as the National Digital Information Infrastructure Preservation program (NDIIPP), funded by the Library of Congress, and the NSF's National Science Digital Library, continue to support innovative research in this area. Other countries have also contributed funding, mostly focused on the digitization of content, although some organizations such as the U.K.'s JISC (Joint Information Systems Committee) have funded a mix of content and innovative research.

Search and information retrieval have long been significant components of digital libraries, and commercial search engines such as Google, Yahoo, and MSN are now extremely popular. Search engines do not, however, constitute digital libraries, which integrate collection management, access, and other services. Some notable examples of modern digital libraries include the arXiv pre-print server [McKiernan, 2000]; and the many on-line electronic journals collections made available through the JSTOR project [Guthrie 2001] and by many of the major commercial and open publishers. (Also see D-lib magazine, which routinely highlights notable digital collections.) Moreover, within the last five years, software systems that provide complete digital library services have become available, including: Greenstone [Witten 2002], VDC [Altman, et. al 1999], Fedora [see Lagoze , et. al 2005], and DSPACE [Tansley, et. al 2003].

## **DIGITAL LIBRARIES – ADVANTAGES AND DESIGN**

Table 1: Some Advantages of digital libraries

- Convenient access
- New forms of search
- Eases information sharing
- Availability
- Lowered costs
- Ease or replication of content
- Enables new forms of content and collection

Digital collections have a number of distinct advantages over traditional physical collections. One of the largest is that they typically allow convenient access, at all hours, to a wide variety of materials, from any location that provides Internet access. By offering different forms of access, digital libraries may also expand the potential user community, by offering, for example, more convenient access to disabled users, digitizing fragile materials for popular use, or offering access to remote populations. (As a case in point, by digitizing collections of older journal articles, JSTOR increased the usage of these articles by the same user communities by a factor of 10. [Guthrie 2001]. Offering open access to journals has been shown to further increase usage. [See Wilinsky 2005])

Cost is another significant motivation. Recent adoptions of digital libraries show that they can dramatically reduce the costs of some types of access. [Montgomery and King 2002] Although not yet uniformly less expensive than off-site depositories of paper materials, in the foreseeable future, for many materials, digitizing and distributing digital collections will be simply less expensive than maintaining the corresponding physical space.

Digital libraries have a number of other advantages They can easily be replicated as a safeguard against loss. They can more easily be shared for collaborative work. They enable new forms of search, since both the metadata and the content of digital objects can be indexed. They enable new forms of information organization, since digital materials are not limited to a single location on a single shelf, and they can be included in many different cross-cutting virtual collections. They facilitate new forms of content, such as newspaper articles with interactive graphs, journal articles with embedded data and simulations, works that include accumulated commentary, and “mash-ups” intermingling content from multiple works in distinctive (and often dynamic) ways.

Architecturally, a digital library system typically includes four types of components: repositories, catalogs, identifier systems, and user interfaces. (This is sometimes called the Kahn-Wilensky architecture [Kahn & Wilensky 1995,2006]) *Repositories* store the raw bits comprising each digital object contained in the library. (Repositories may be little more than file-systems, or may be interfaces to distributed, multi-level storage systems.) *Catalogs* support search, by indexing information in the digital objects, and metadata describing them. *Identifier systems* provide a framework for identifying and locating objects (“resolving” an identifier). *User interfaces* bring together the functions of the other components to perform services for the users – such as, searching, browsing, visualization and delivery.

Many other components, services, and agents are also often incorporated into digital library architecture, such as:

- Discovery services that provide a mechanism (such as a central directory) by which interfaces to the digital library system dynamically discover the services that are available for the user.
- Security services to support authentication of users (perhaps by recognizing authentication from another party), provide framework for making authorization decisions, and secure communication services
- Presentation services that assist the user in viewing complex digital objects that are not otherwise easily viewable at the users desktop or in a browser. This may include on-line systems for data analysis, reformatting services, and streaming media players.
- Software agents, that run independently of the main digital library system, and which perform additional services with respect to the library’s content. These may be used to support services that notify the user of new content of interest, that harvest the content of

one library into another (see Van de Sompel and Lagoze 2000), or that aggregate search results from multiple digital libraries for the user.

## ISSUES AND FUTURE TRENDS

Although digital libraries have yielded significant benefits and gained rapid adoption, a number of significant challenges are foreseeable. A digital library is, as Arms [1995] aptly points out, far more than a collection of digital objects, systems, and software. He identifies a number of core digital library principles, paraphrased here:

- Digital library technology exists within a legal and social framework.
- Digital library concepts are obscured by technical jargon.
- Software systems should be separated from digital library content.
- Persistent identifiers are a basic building block in the digital library.
- Different forms of an object may be necessary for storage, delivery, display, and use.
- Repositories should preserve the information in digital objects
- Users want access to intellectual works, not digital library technology and objects

Many of the trends and challenges surrounding digital libraries are corollaries of these basic principles.

One of the largest challenges is preservation. Archivists face a new paradigm in the digital age. Previously, much of archiving physical object centered on maintaining them in their original forms. This is neither necessary nor sufficient for works in digital form. Both hardware storage media and software file formats are constantly evolving, and thus digital objects must be frequently migrated from one media and format to another in order to ensure future usability. At the same time, these reformatted objects may need to be carefully checked in order to ensure no intellectually significant information is lost in the reformatting (“Universal numeric fingerprints” offer a possible approach to solving this problem [Altman, Gill, McDonald 2003]). Complex dynamic objects such as data-driven websites and software pose particularly difficult technical challenges.

In addition, the responsibilities and rights of institutions with respect to archiving and disseminating digital materials are often hazy. The elimination of the requirement to register or even to include notices of copyright by the Copyright Act of 1976 (moving U.S. law towards the international Berne Convention), implies that copyright applies automatically to almost all digital works. The act’s extension of copyright from dual 28 year terms to the life of the author plus fifty years (changed in 1998 to life plus 70 years) implies that newly created digital objects are likely to remain under copyright until long after any systems capable of displaying them have disappeared. (The Digital Millennium Copyright Act’s prohibition against reverse engineering and decryption, and the Supreme Court’s extensions of patents to software in *Diamond v. Diehr* (1982) have further complicated digital preservation efforts.) To ensure preservation, multiple independent archival copies, migration, and emulation approaches, may be needed. Moreover, many libraries may need to take on additional preservation responsibilities, in order to ensure that one publisher’s demise does not wipe out swathes of important content. [Keller, et. al 2003]

The issue of persistent identifiers is related to preservation. In order to manage digital works over time, to cite those works, and to copy them, those objects need to have identifiers that are unique and persistent. This is essentially a problem of preserving the linkages among different digital objects and catalogs. Digital object identifiers (DOI’s) are a form of identifier [Paskin 2000] that is increasingly being adopted by the publishing community. However, no persistent identifier technology currently has the broad base of institutional and technical support (e.g.

support in end-users' browsers) and offers all of the functionality needed for digital libraries in the future.

Another issue that has generated concerns among researchers, educators, activists, and librarians is the wide variation in the quality and coverage of content in many digital libraries. This variation has a number of sources. Reports in the press have highlighted the willingness of search engines and hosting sites such as Google and MSN to actively self-censor controversial content in response to government pressures [Mills, 2006]. Information scientists who study web search have suggested that a number of more inadvertent biases may also be present in current digital libraries. Wouter [2004] find a bias in both content selection and search results weighting that tilts toward more recent material. This "recency bias" occurs, at least in part, because new material are "born digital" and thus easier to include in digital collections; current journal articles and other similar content demands a price premium and is thus more profitable for the publisher; search engines and digital libraries often do not retain copies of previous versions when material is updated; and search engines may explicitly weight recently updated material more heavily. In addition, Gerhart's [2004] analysis suggests that search engines have an inadvertent tendency to present the sunny side of controversial topics, because the structure of the web in general more strongly reflects organizations than ideas, because controversy may be lost in junk links, and because controversial websites often lack organizational clout

Such biases may be slow to self-correct, because, as Blair and Maron [1985] show in early work, users are not very good at telling how well an information retrieval system actually works, and they tend to overestimate the completeness of the results that such systems deliver. The increasing 'Googlization' of information seeking suggests that users are not becoming more critical of search systems. Underlining this, in a recent large survey, two thirds of college students reported that they started their research projects with a Google search and held Google to be as trustworthy as their library. Furthermore, almost one third of the student respondents, reported that they did not validate the results of web searches against any print sources. [Rosa, et. al, 2006] Lesk offers an interesting perspective on the issues of usability and accuracy: "Users are neither clear about what they want, able to operate the systems well, nor doing much to get help. However, they're satisfied with the results. To the extent that we can tell, the acceptance of new systems is partly based on the inability of users to tell how badly things are actually working, and partly on the probability that older systems were also not being used very effectively." (Lesk 2005, pg. 217)

Despite the complexity of the technical issues, the biggest challenge for future digital libraries are likely to be institutional. Business models for libraries are now a huge unsolved issue. As Lesk aptly summarizes, libraries continue to rely on institutional subvention and have yet to find a way to monetize most transactions between users and libraries in a way that is fair, incentive-compatible, and does not involve exorbitant administrative overhead.

The economic pressures on libraries are created in large part by the disintermediation of publishers and end-users made possible by network technology, and also, in part, by changes in the intellectual property rights regime associated with digital materials. Both driving and driven by the changing intellectual property regime, the increasing use of digital rights management (DRM) software to restrict access to digital materials complicates the libraries dual roles in disseminating information and in preserving it. In the past, libraries would typically own a physical copy of the works that they purchased, and could make a straightforward decision as to whether to circulate it or preserve it, or even to duplicate portions as "fair use." DRM moves libraries from a position of ownership to a more limited and ambiguous position, since, fair use notwithstanding, it may place limits on such things as how many times an object can be viewed,

limiting the number of pages accessed, the ability to print the content, or the ability to export that content to other formats.

The issue of “orphan works” is a telling example of the unintended effects of intellectual property law, and a serious problem for libraries. “Orphan works”, are works for which the current copyright status or copyright holder is unknown. These works constitute a significant portion of works of some types and for some periods, especially: photographs that are embedded in other works, and older films, books, and music that have not yet clearly entered the public domain. Since most created works are not commercially published, and most of those go out of publication quickly, most orphan works are not a potential source of profit to anyone. (Prior to 1976, when registration and renewal were required to gain and retain copyright, only 15% of who registered copyrights were renewed. [U.S. Register of Copyrights, 2006] ) Nevertheless, orphan works cannot be legally distributed (or preserved) by libraries which do not possess a physical copy.

Many libraries would be readily pay reasonable fees for the restricted redistribution and preservation rights that they have traditionally relied upon for works in physical form. However, the administrative costs and complexities involved in DRM and intellectual property management are often prohibitive. (For example when IBM produced a digital commemoration of Columbus’s voyage, they spent a million dollars on copyrights, only 99% of which was administrative costs, *excluding* actual copyright fees. (Garrett and Walters, 1996, cited in Lesk 2005)) Moreover, publishers often do not devote effort to these issues because they perceive libraries to be a relatively small source of profit.

For scholarly libraries, the costs of academic journals have become an increasingly important issue, due to a dramatic increase in prices. Exacerbating this increase, publishers have adopted the practice of ‘bundling’ electronic access to large groups of journals together, which makes it more difficult for libraries to be selective in their subscriptions. As a reaction to this, and to take advantage of the opportunities that the Internet offers for widespread distribution, a movement toward “open access journals” has arisen. Open access journals, proposed in 1995 by Steven Harnad, a cognitive scientist and a specialist in peer review [Okerson & O’Donnell 1995], offer free public access while maintaining professional peer review and quality standards. Many new open access journals have been created, and some, such as the Annals of Mathematics, are even constituted as “overlays” on top of pre-print servers, such as arXiv, where the journal itself essentially constitutes a table of contents and procedure for peer review. In addition, some funding agencies are beginning to mandate open access to publications based on research they sponsor, [See Wilinsky 2005 for a detailed discussion].

These technological, legal and economic trends underscore fundamental changes in the role of the library, as well as in the form of its contents. Possession of information by a library, which was once a primary measure of institutional success, is becoming less important -- individuals may soon have the equivalent of large libraries in their pockets. At the same time, the ability to facilitate an individual’s search for, access of, and collaboration with information is becoming much more important. Such facilitation is not straightforward, however. As Brown and Duguid [2000] point out, networks of information can have great reach, but still fail to promote the types of interactions among users that produces “social knowledge”. How to facilitate effective collaboration within a digital library remains an open technical and institutional question.

Preservation Preservation of complex digital objects Persistent identifiers	Intellectual property Copyright DRM
---	---

Citations  Bias  Search algorithm bias Censorship Incomplete coverage, recency bias Indiscriminating users	Patents Orphan works Open Access to Journals  Role of library Collaboration Selection Preservation
Table 2: A Summary of Critical Issues of Digital Library Technologies	

## CONCLUSION

Digital libraries offer expanded access to information at all levels of complexity -- from recipes to research. These libraries have made it possible for experts to access information far more rapidly and efficiently, and lowered the bar to non-experts who are working to find information of all sorts. Currently, however, such libraries offer only a portion of the collections available at large traditional libraries, and significant institutional and technical challenges remain to be overcome.

## REFERENCES

- Altman, Micah, Leonid Andreev, Mark Diggory, Michael Krot, Gary King, Daniel Kiskis, Akio Sone, Sidney Verba. 2001. "A Digital library for the Dissemination and Rplication of Social Science Research", *Social Science Computer Review* 19(4):458-71.
- Altman, Micah, Jeff Gill and Michael McDonald (2003) *Numerical Issues in Statistical Computing for the Social Scientist*, New York: John Wiley and Sons
- Arms, William Y. 1995, "Key Concepts in the Architecture of the Digital Library", *DLIB Magazine* (July).
- Arms, Williams Y. 2000. *Digital Libraries*, Cambridge, MA: MIT Press.
- Bisco, Ralph, 1970. *Data bases, Computers and the Social Sciences*. New York: Wiley-Intersciences.
- Blair, Davis C., And M.R. Maron, 1985. "An Evaluation of Retrieval Effectiveness for a full-text document retrieval system", *Communications of the ACM* 28(3):289-99.
- Brown, John Seely, and Paul Duguid, 2000. *The Social Life of Information*. Cambridge, MA: Harvard Business School Press.
- Bush, Vannevar, 1939. "Mechanization and the Record", Letter to the Editor. *Fortune*. [Vannevar Bush Papers, Library of Congress], Box 138, Speech Article Book File.
- Bush, Vannevar, 1945. "As We May Think", *The Atlantic Monthly* 176(1):101-8.

Gerhart, Susan L. 2004. "Do Web Search Engines Suppress Controversy", *First Monday* 9(1). URL: <[http://firstmonday.org/issues/issue9\\_1/gerhart/index.html](http://firstmonday.org/issues/issue9_1/gerhart/index.html)>

Garrett, John and Donald Waters, 1996. *Preserving Digital Information*, Mountain View, CA: Commission on Preservation and Access and Research Libraries Group joint publication.

Griffin, Stephen M. 1998. "NSF/DARPA/NASA Digital Libraries Initiative: A Program manager's Perspective," *D-Lib Magazine* (July/Aug).

Guthrie, K.M., 2001. "Revitalizing older published literature: preliminary lessons from the use of JSTOR." In J. MacKie-Mason and W. Lougee (eds), *Bits and Bucks: Economics and Usage of Digital Collections*. Cambridge, MA: MIT Press, URL: <<http://www.si.umich.edu/PEAK-2000/>>

IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998. *Functional Requirements for Bibliographic Records*, International Federation of Library Associations and Institutions. URL: <<http://www.ifla.org/VII/s13/frbr/frbr.pdf>>

Kahn, Robert, and Robert Wilensky. 2006. "A Framework for Distributed Digital Object Services" *International Journal on Digital Libraries* (2006) 6(2): 115–123. DOI 10.1007/s00799-005-0128-x First published in 1995 as: <URL: <http://www.cnri.reston.va.us/k-w.html> >

Keller, Michael A, Victora A. Reich, Andrew C. Herkovic, 2003. "What is a library anymore, anyway?", *First Monday* 8(5) URL <[http://firstmonday.org/issues/issue8\\_5/keller/index.html](http://firstmonday.org/issues/issue8_5/keller/index.html)>

Lesk, Michael 2005, *Understanding Digital Libraries* (2<sup>nd</sup> Edition), San Francisco :Morgan Kaufman.

Lagoze, Carl, Dean B. Krafft, Sany Payette, Susan Jesuroga, 2005. "What Is a Digital Library Anymore, Anyway" *D-Lib Magazine* 11(11) URL: <<http://www.dlib.org/dlib/november05/lagoze/11lagoze.html>>

Licklider, JCR, 1965, *Libraries of the Future*, Cambridge, MA: MIT Press.

McKiernan, Gerry 2000. "ArXiv.org: The Los Alamos National Laboratory E-print Server." *The International Journal on Grey Literature* 1(3): 127-138.

Mills, Elinor, "Google to censor China Web Searches", 2006. *CNET News.com*. January 24.

Montgomery, Carol Hansen and Donald W. King, 2002. "Comparing Library and User-Related Costs of Print and Electronic Journal Collections", *D-Lib Magazine* 8(10) URL: <<http://www.dlib.org/dlib/october02/montgomery/10montgomery.html>>

Okerson, Ann, & James O'Donnell (eds), 1995. *Scholarly Journals at the Crossroads: A Subversive Proposal for Electronic Publishing*, Washington, D.C.: Association of Research Libraries.

Paskin, Norman, 2000. "E-Citations: Actionable identifiers and scholarly referencing", *Learned Publishing* 13(3): 159-168.

Rosa, Cathy De, Janne Cantrell, Janey hawk, Alane Wilson. 2005. *Perceptions of Libraries and Information Resources*, Dublin, OH. OCLC Online Computer Library Center, Inc.

Tandley, R. M. Bass, and D. Stuve, M. Branchofsky, D. Chudnov, G. Mclellen, and M. Smith. (2003), "The Dspace Institutional Digital Repository System: Current Functionality", In Marshall, C., Hengry, G., and Delcambre, L. (eds). *Proceedings of the 20003 Joint Conference on Digital Libraries*, New York: ACM Press. 87-97.

U.S. Register of Copyrights, 2006. *Report on Orphan Works*. Washington, D.C.: United States Copyright Office, Library of Congress.

Van De Sompel, Herbert and Carl Lagoze, 2000. "The Santa Fe Convention of the Open Archives Initiative", *D-Lib Magazine* (6)2.

URL: <<http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>>

Witten, H. Ian, David Bainbridge. 2002. *How to Build a Digital Library*, San Francisco: Morgan Kaufman.

Wilinsky, John. 2005. *The Access Principle*, Cambridge, MA: MIT Press.

Wouters, Paul, Iina Hellsten, and Loet Leydesdorff, 2004. "Internet time and the reliability of Search Engines." *First Monday* 9(10).

## TERMS AND DEFINITIONS

- Work/Edition/Manifestation/Item hierarchy. A set of principles for distinguishing between a distinct *work* of intellectual creation (e.g., Beethoven's 5<sup>th</sup> Symphony), the *edition* of that work (e.g., the 1979 performance by the Vienna Symphony orchestra), the *manifestation* of that work (e.g. an MP3 file created with a 192 bit sampling reate settings), and a particular item (e.g., a copy of that MP3 file that resides in a particular repository).
- Universal Numeric Fingerprint. A universal numeric fingerprint is used to guarantee that a two digital objects (or parts thereof) in different formats represent the same intellectual object (or work). UNF's are formed by generating an approximation of the intellectual content of the object, putting this in in a normalized form, and applying a cryptographic hash to produce a unique key. [Altman, et. al 2003]
- Metadata. Metadata is often defined as "data about data". More specifically, it is information that refers to other digital objects. Metadata often consists of descriptive information (such as a title), administrative information (such as a description of the rights required to view the object), and structural information (such as the organization of page images within a larger book). Metadata can be derived from the object itself in order to used as a surrogate for searching and other services, but is more often information that is not contained in the object itself.
- "Born Digital" vs. digitized objects. Objects that were born digital were created originally in digital forms, for example a word processing file. Digitized objects were created from non-digital forms (e.g., via optically scanning a paper book).
- Simple/Complex Objects. Simple digital objects consist of a single file that can be fully understood by the user and represented by the user's software. Complex digital objects require multiple separate files, and possibly additional metadata, to be properly understood.

- Union Catalog – An online public access catalog (OPAC) formed by indexing descriptive metadata for each item in the library. In the digital library, Union Catalogs are increasingly being supplemented or replaced by a combination of distributed search across multiple independent catalogs, and direct indexing of digital object content.