Paper submission for ASIS Mid-Year 1996 meeting:

Digital Library Models and Prospects

by Gregory B. Newby
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street; Champaign; IL; 61820
Email: gbnewby@uiuc.edu / Telephone 217-244-7365

ABSTRACT

Digital libraries are the means by which people of the next millennium will access materials found in current libraries, yet the nature of digital libraries is only now being shaped. Different visions of digital libraries include the digital library for electronic access to materials previously found in print form, access to data stores, and access to scholarly (self-published) materials. In order to bring about significant progress towards the realization of full-scale digital libraries, existing publishers must participate. Yet economic impediments to full participation of publishers exist, including fears of unauthorized duplication, distribution, or modification. Uncertainty about whether digital access to published materials may yield acceptable fees or royalties may be addressed through adoption of “pay as you go” or “buy once, use many” models, yet significant infrastructure development is needed to make these models feasible. The proliferation of scholarly reprint archives, moderated discussion groups, and electronic conferences and proceedings is helping to address some of the publisher concerns, as are developments in network access, browser standards, bibliographic standards, data security, and encryption.

THE DIGITAL LIBRARY

Conferences, journals, books, and projects have emerged in 1995 and 1996 on a topic which has been of only limited interest previously: digital libraries. The basic concept of the digital library may be traced back at least to such visionaries as Bush (1945), Licklider (1965) and Lancaster (1969). The sudden surge of interest in digital libraries is due primarily to the modern state of the Internet. The

Internet has matured to the point of near-ubiquity in higher education institutions, and sophisticated cross-platform browsing and display tools for the Internet are widely available (notably Mosaic and its offshoots). It's therefore to be expected that long-held visions of digital libraries can be implemented in earnest, with expectations of considerable progress by the millennium.

The basic concept of a digital library is uncertain, and subject to debate. All notions of digital libraries include electronic ("digital") storage of materials for retrieval or processing, but beyond this common base there are several major approaches to digital libraries. Some of the major approaches include data stores, electronic access to traditional library materials, and scholarly archives.

The data store is the least glamorous approach to digital libraries, yet is concerned with topics critical for the other approaches. For a data store, the goal is to make large collections of numeric, textual, images, or other data accessible in electronic form. Examples include an academic slide archive or a collection of nightly news programs from around the world. Important concerns for the creation of such a digital library include the storage of the data (what file formats to use, how to maximize for quick retrieval, how to make the data available across a local- or wide-area network); indexing and retrieving (what sorts of descriptions should be applied to the data, what retrieval methods should be used -- for example, for browsing or for keyword searching); and standards (what standards for data organization should be used in order for people to be able to display or process the data?). None of these areas are trivial, and examples of the type described above are at the limits of what researchers and commercial developers are able to accomplish, especially on a large scale.

Electronic access to traditional library materials is the approach to digital libraries which may be the most obvious. The challenge is to make books, journals, vertical files, indexes, and other (print) materials found in libraries accessible to patrons in an electronic format, preferably from any location. While the storage of, for example, a book in a computer file is not necessarily challenging, since, after all, the book probably started out as a computer file, the surrounding issues are challenging. Making the contents of an electronic library searchable, insuring that access to materials is gained only by those to whom access is granted, delivering materials, agreeing on standards for display, search, and retrieval, and so forth are necessary issues to be addressed in order for this approach to digital libraries to come to pass. Note these issues are largely similar or identical to those for providing access to a data store -- the main differences are that the data store is intended for post-processing, but traditional library materials are to be used more or less “as is,” and traditional library materials are geared more towards the general public, versus a more specialized user group for the data store.

The final approach to digital libraries to be considered here is that of the scholarly archive. The primary idea is to bypass much or all of the existing publishing industry to make scholarly articles available. To this end, the main areas of development to date are scholarly pre-print archives and networked discussion groups (Harnad, 1996). But development must continue to achieved the goals of quick, ready, and equitable access to scholarly works. The main impediments to development include the technical problem of maintaining good quality control, editorial oversight, and other benefits that journal publishers provide, and the social problem of insuring that scholarly publications which bypass publishers are seen as legitimate by the scholarly community and by tenure review committees. A further area of great concern is maintaining archival access to the scholarly publishing -- insuring that a journal article will be available many years in the future is of concern of both the author and the researcher.

Common threads for these major approaches to digital libraries include primary areas of inquiry of information science. How can data collections be successfully searched by a variety of users with a variety of information needs? What indexing, abstracting, and organizational techniques should be applied to materials? What mechanisms should be utilized for timely retrieval and display of materials?

This paper will continue with an analysis of economic models which may be applied to digital libraries, with a focus primarily on the electronic access to traditional library materials and scholarly archives approaches. Then, prospects and challenges for developing economic models and other aspects of digital libraries will be examined.

ECONOMIC MODELS

Books cost money. The cost of producing a book includes fees, salaries and royalties paid to authors, editors, indexers, and others; cost of printing and binding; distribution costs; and others (Freeman, 1996). Because books cost money to produce (as do journals, indexes, newspapers, and most other items found in libraries), fees are charged to recoup the costs.

Individuals who purchase books receive a copy of the book to use at their discretion. Barring reproduction, almost any use is usually legitimate: giving the book away, sharing the book, storing the book indefinitely for later access, destroying the book, etc. For a library, similar rules apply: the library may make its own policies for how the book may be used, loaned, etc. For most materials found in libraries, responsibility for how the material is used is up to the individual library. Only in some instances (e.g., for CDROM databases or other materials which are updated frequently) does the publisher include stipulations for how a material may be used or disposed of.

For digital libraries, though, additional guidelines are needed to determine what sort of use is legitimate. According to Freeman (1996), the actual physical printing and production cost of a book is only 10-20% of the total cost of producing a book. Publishers want to make sure they are able to profit from their books in digital forms, as they do in print forms. But "buy once, then use as you wish," if applied to digital books or other materials, creates a threat to publishers recouping their costs due to the ease with which electronic materials may be copied, shared, and redistributed.

Consider the simple case of a book which has been transcribed to plain ASCII text, such as those produced since 1971 by Project Gutenberg (see ftp://uiarchive.cso.uiuc.edu/pub/etext/gutenberg). If one desires a copy of, perhaps, Henry Thoreau's Walden, she may retrieve it by anonymous FTP or other means. She may then print the text, reformat it, redistribute it, or destroy her copy. Gutenberg places almost no restrictions on what one may do with an etext they publish. But if they did, how would restrictions be enforced? There are no good methods to insure that the text is not printed, reproduced, retransmitted, modified and resold, etc.

As the forms of electronic books or other materials get more sophisticated, some of the things which people can do with plain ASCII texts become more difficult:

- SGML or other text markup schemes can make it more difficult for the layperson to modify a text.

- Larger bodies of work (such as a CDROM encyclopedia) are less likely to be transmitted or copied due to their size.

- Some browsers may limit printing, or create problems with reproduction. For example, an online hypermedia book may contain electronic movies, sounds, or other materials which are more difficult (again, for the layperson) to reproduce or retransmit.

- Legal copyright restrictions can also help to prevent unauthorized duplication or distribution, as they do with the photocopying of books and journal articles.

- Texts stored in non-copyable (or less-copyable) formats can include encrypted CDROM data or hardware ROM-based data. Or, text readers may be utilized which keep track of when a text is used.

Clearly, from the above, there is no guarantee that a publisher's work can be fully protected from uses the publisher would rather not permit. In the world of physical books, publishers cannot prevent pirate publishing operations from stealing their work wholesale, then reproducing and selling it on their own. However, such instances are relatively rare -- as, for example, counterfeiting of US currency is relatively rare -- due to the expense and expertise required.

In the world of the electronic book or other electronic material, unauthorized duplication and distribution is much easier. As the (estimated) billions of US dollars lost to software piracy annually attests, people are willing to duplicate and distribute electronic materials even when the law prohibits it.

The imperative for publishers to profit from their work in the electronic world is therefore at odds with the ease with which materials in electronic forms may be put to uses which, in the print world, are not legitimate. A major issue for digital libraries -- and a crucial impediment for large-scale digital library implementation -- has to do with how publishers can make their materials available in electronic form with confidence that the materials will not be put to uses they do not wish to permit.

Two major models for publishers to sell their products for digital library use exist. A third model sidesteps the publishers. The two publisher models are "pay as you go," and "buy once, use many." The model which sidesteps the publishers is the author as publisher/distributor -- as discussed for scholarly archive creation, above.

"Pay as you go" refers to the charging scheme currently employed for online databases, pay-per-view movies, and other things. In this model, publishers (or their agents -- perhaps the digital libraries themselves) would make electronic materials available on-demand, but not transfer ownership of a copy to the user. In a world with ubiquitous access to the Internet or another network, the most likely scenario is that the central distribution point for a material would let a user view the material, but not actually provide a usable copy of the material.

One possible technical scheme for such distribution would be to have browsers/viewers that are capable of displaying materials sent using a one-time encryption key. This is not a perfect scheme, of course. The idea is to require the user to communicate with the distributor (and, presumably, pay some sort of fee) every time she wants to view a book, journal article, or other material. As with pay-per-view movies, the publisher would rely on a combination of the technical difficulty of making a personal copy and distributing it, the illegality of doing so, and the ease (and, one hopes, relatively low cost) of obtaining legitimate access to the material.

The necessary infrastructure for pay as you go schemes for digital libraries barely exists. Although the Internet may be used for access to publishers' (or others') materials, the Internet's emphasis is on free and open transmission of data. Access restrictions, encryption schemes, etc. for the World Wide Web, for example, are fairly limited and not yet reliable. Browsers for WWW data, as for email, Gopher, word processed files, etc. are not geared towards encryption, restrictions on such things as printing or editing, or restrictions on storage or redistribution.

Pay as you go is, however, probably the method most desired by publishers to sell their wares. It is fairly close in nature to existing models for how books and journal articles are sold, and is able to keep revenues fairly well synchronized to the amount of use for a particular material. Developers of digital libraries would need to invent the necessary infrastructure in order for pay as you go to work, should this prove to be a dominant model for digital library access.

The "buy once, use many" model is how libraries usually work. Libraries purchase a book and decide for themselves how to provide access. For most libraries, books and other materials are made available for loan to any authorized person. The digital library, in such a scenario, can know for certain the cost of a particular material. In the pay as you go model, the library cannot predict the cost of a material, since it cannot know how many individual uses of the material may be made in the future. (One of the major issues that becomes apparent here is whether a digital library is something which is an independent organization, as traditional libraries are, or something run by the publishers or their agents.)

Under the buy once, use many model, a library would need to insure that restrictions of the publisher were enforced. For example, publishers might dictate that only one person may make use of a material at any one time (typical of software or CDROM licensing schemes currently in place). Or, libraries might be forced to insure that the sorts of encryption schemes and reproduction impediments described above are in place.

The "scholar as publisher" model is one which requires relatively few technical advances and is so occurring at a good pace already. Through pre-print servers, electronic conference proceedings, electronic mailing lists, and electronic journals in a variety of formats, scholars are now able to make their work widely available without the participation of publishers. In some cases, editorial boards and editorial processes are used. In others, there is little opportunity for editorial oversight, peer review, etc.

The main limitation of the scholar as publisher model for the creation of digital libraries is that the audience is relatively narrow: A digital library content would not be complete with only self-published scholarly works. Even if it were, scholarship does not take only the form of the journal article or conference paper -- books, monographs, edited conference proceedings, and so forth are also important. Without good opportunities for compensation for the time spent at putting together such works, such materials would probably not be produced in any great quantity.

An additional important limitation of the scholar as publisher model is that indexing and abstracting services may be unaware of the published works, or unwilling to add such "gray" literature to their databases. Without being searchable, indexed, and abstracted, the value of scholarly works are lessened by their decreased accessibility.

This section has presented a summary of the issues surrounding the practical creation of materials to be stored and distributed by digital libraries. From the standpoint of publishers, who produce the vast majority of materials found in the collections of existing traditional libraries, important risks to revenues would be encountered by providing electronic copies of print materials for use in digital libraries, even on an experimental basis.

The problems are not insurmountable, but unfortunately there does not yet appear to be an immediate financial imperative for publishers to solve them. Instead, the scholarly community is attempting to resolve the issues that face publishers and other stakeholders in digital library development. Some of these prospects and developments are discussed in the next section.

PROSPECTS AND DEVELOPMENTS

The analysis in the previous sections has focused on the role of the publishers -- who create the majority of materials which are found in current libraries, and so could be expected to contribute greatly to digital library development. There are other areas of work, though, which this section will briefly address. These areas include data browsers, network development, network security, electronic publication of scholarly materials, and developments in resource sharing for libraries.

Data browsers for network access have been ubiquitous on the Internet since about 1990. Gopher, WAIS clients, and the World Wide Web all became available at about that time. The advantage of these types of data browsers over older methods (cf. telnet and FTP) are that they remove the need for an information seeker to know the network location of the items he is looking at. The first graphical browser for the World Wide Web, Mosaic, became available in 1993 and added greatly to the capability of network users to "publish" (usually informally) work with visual appeal, greater depth, and better organization. In the near term, network data browsers (the progeny of Mosaic, Netscape, and others) will incorporate the ability to display data from a variety of formats, most notably SGML. SGML is the standard employed by most publishers in the creation of their work. Thus, the union of SGML capability with networked browsers will yield browsers which are capable of directly accessing the data in a digital library in their most likely format.

Thanks to the ease of use for the Internet which tools such as Mosaic have helped to bring, to the favorable press coverage the Internet has received, and to the increased role of the microcomputer for home and office use, Internet access is now commonplace throughout US society. In higher education, Internet connectivity and basic access for faculty, students, and staff is nearly universal. This widespread access to the Internet from the home, office, schools, etc. means that access to digital libraries, when they exist, will be easy to obtain.

Along with the interest that many individuals have taken in the Internet over the past few years has been a surge of interest in the private sector. The largest area of growth on the Internet in 1994 and 1995 has been the commercial domain (according to the Internet Society: http://www.isoc.org/). The outcome of this growth which most closely effects digital library implementation is the need of commercial applications on the Internet for security for data delivery. Security is especially needed for customers to make credit-card purchases on the Web, but also for return delivery of data and for ensuring customer confidentiality. As discussed above, the ability of materials found in digital libraries to be encrypted and transmitted securely may be important to publishers. The ability of digital libraries (or publishers) to identify and bill users of materials appropriately is also likely to be extremely important. These needs of digital libraries will be directly met by commercial interests in security for Internet transactions.

The future of scholarly self-publishing for digital libraries is not entirely clear, as it may be that such publishing is superseded by the digital libraries of the future (hopefully to the satisfaction of all parties and purposes involved). Current efforts for edited electronic journals, refereed preprint collections, and moderated electronic discussion are all relevant practice for large-scale digital library development.

A final area of development relevant to digital libraries is continued progress with bibliographic standards (MARC, as applied to electronic materials) and bibliographic interchange formats (Z39.50, used to query databases remotely, including over the Internet). These developments have been underway for a long time, partially due to a relative lack of real materials being used in real libraries that require such services. Both academic libraries and library automation vendors, among others, are making gradual progress towards remote queries and bibliographic standards for electronic materials.

Other areas of development for digital libraries are underway as well. Issues for effective storage and delivery hardware are not trivial, as are increases in network speed and reliability. Interface design is another area of research with key implications for digital libraries. And, many user studies are underway which can help to identify the information needs, situations, individual factors, and other components that must be combined appropriately to create environments for effective information retrieval.

CONCLUSION

The exact role of the digital library for the 21st century is far from certain. In order to develop digital libraries sufficiently so that their role in public, organizational, and academic life might begin to emerge, many steps must be taken. Research is underway, most notably at universities (ACM 1995), which will address some of the fundamental issues for digital libraries.

The role of the traditional publisher for the digital libraries of the future is of particular interest, due to their centrality to libraries of the present. But in order for publishers to start making a transition to participation in digital library formats, solutions to the problems of illicit copying, distribution, access, and modification of publishers' materials are needed. Publishers might prefer a "pay as you go" model, in which only temporary access is given to electronic materials, with some sorts of usage meters or restrictions. Libraries are more familiar with a "buy once, use many" model, in which ownership of a copy of a material is made to a library, then the library loans or otherwise provides access to that material as it sees fit.

Developments in scholarly self-publishing may help to guide the transition to digital access to materials from commercial publishers by creating standards, identifying demand, and providing materials for a test market. The creation of public-access data stores -- for example, for US Census data, Supreme Court decisions, and geographic data -- also can help lead to standards and provide a test market.

There is no doubt that the digital library, not the print library, will be the basic access point for information of all types within only a few years -- for scholars, for business uses, and for the general public. The nature of the digital libraries of the future is currently being shaped by researchers, scholars, government funding agencies, publishers, and commercial interests. This work has attempted to characterize some of the important areas of concern for the development of digital libraries, and discuss current work and some anticipated outcomes.

REFERENCES

Association for Computing Machinery (ACM). 1995. Communications of the ACM special issue on digital libraries. June.

Bush, V. 1945. "As we may think." Atlantic Monthly 176: 101-108.

Freeman, Lisa. 1996. "The university press in the electronic future." in Peek, Robin P. & Newby, Gregory B. (Eds.). Scholarly Publishing: The Electronic Frontier. Cambridge, MA: The MIT Press.

Harnad, Stevan. 1996. "Implementing Peer Review on the Net: Scientific Quality Control in Electronic Journals" in Peek, Robin P. & Newby, Gregory B. (Eds.). Scholarly Publishing: The Electronic Frontier. Cambridge, MA: The MIT Press.

Lancaster, F.W. 1969. Conceptual Alternatives to the Scientific Journal. Bethesda: ERIC.

Licklider, J.C.R. 1965. Libraries of the Future. Cambridge, MA: The MIT Press..