Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2000 Proceedings

Americas Conference on Information Systems (AMCIS)

2000

Global Digital Libraries: A Historical Perspective and Architectural Considerations

Mahesh S. Raisinghani
University of Dallas, mraising@gsm.udallas.edu

Robert Scott Dupree
University of Dallas, scott@acad.udallas.edu

Global Digital Libraries: A Historical Perspective and Architectural Considerations

Mahesh S. Raisinghani, Ph.D., University of Dallas,
Graduate School of Management, mraising@gsm.udallas.edu
Robert Scott Dupree, Ph.D., University of Dallas, scott@acad.udallas.edu

Abstract

Digital libraries will replicate many of the features of traditional libraries, but they may also draw on techniques developed by video rental companies and bookstores in making information both readily accessible and attractively presented. A generalized schema for global information systems of this type is essential; it must address issues of data structure and system interoperability so that information of all types can be freely exchanged across whatever platforms may develop in the future. The ability to search, identify, and retrieve not only text but recorded sound and complex images depends on a systematic approach to handles and metadata. The mode of access to data will be partly determined by economic and social forces and could require additional tagging. The global digital library may well be an inevitable next stage in the evolution of information sharing, but its implementation requires a plan that addresses the complexities of these issues on a broadly conceived and coherent basis.

"In the end, the location of the new economy is not in the technology, be it the microchip or the global telecommunications network. It is the human mind."
Alan Webber

Introduction to Digital Libraries

The obvious places to look for the form future libraries might take are those enterprises in which its traditional functions are already being performed or similar needs addressed. One could turn to such successful operations as the video rental store, pay-per-view, and other modes of exploiting the new modes of access to familiar forms of entertainment. But many of the issues that are now becoming prominent have to do with legal or economic rather than strictly technological questions. Problems of ownership and copyright, of authenticity and reliability of information, and of custodianship must be addressed before the advances in hardware and software can be adopted successfully.

A viable system of online payments in small sums—a few cents or even fractions of a cent—would revolutionize the way information is propagated. Buyers reluctant to purchase an entire volume don't mind photocopying a page or two for a few pennies. Once it is economically feasible for the sellers of information to collect such small sums, the volume—indeed, even the article—will no longer be the unit of purchase. Money that now goes to copy machines could go instead to the publisher or author.

The traditional solution to the high cost of printed information has been the public library, an exercise in corporate resources to be freely shared by a community who fund it through membership fees, contributions, or taxes. This pooling of economic resources, seemingly cost-free from the individual's point of view, has not, however, resulted in a loss of revenue for the book trade. On the contrary, it has usually enhanced it. In the nineteenth century a number of publishers and booksellers realized that it was to their advantage to rent books, and a number of lending libraries, which one joined by paying a subscription fee, flourished in Victorian England. As Elmer D. Johnson points out, "Rental fees at these collections were usually small, not over a shilling per month. William Lane of London was one of the most enterprising of the circulating library founders. He established chains of bookstores with circulating collections in them, and then published books, fiction and popular non-fiction, to fill them" (Johnson, 1970, pp. 223-24). In fact, some publishers opened commercial circulating libraries in order to provide a market for titles that would not otherwise have been commercially viable. Far from seeing the library as a rival, the bookseller began to see it as an ally—even a publicity agent—and the publisher regarded it as a particularly desirable customer for certain kinds of books. Though few subscription libraries still exist, the Blockbuster phenomenon might well be considered a contemporary manifestation of the same system. Blockbuster has recently begun to seek exclusive distribution rights to certain movies in videocassette format and could conceivably become, as certain television channels have, producers of films in their own right. What the digitization of libraries may bring about is a convergence of traditionally separate enterprises: publishing, bookselling, and public librarianship.

Assuming that some form of online micropayment will eventually change our current models of commercial information exchange, we must suppose that the other problem to be solved is to find a way of guaranteeing the integrity and reliability of information, not in a technological but in a social and scholarly sense. The high cost of printing and the large investment in equipment necessary for publishing insured a certain prestige in the past; the expense of producing a book and the process of peer review or the screening processes of editorial boards in the case of periodical publication create the impression that the information is precious, valuable, or at least

acceptable. Digital information, available to individuals at low cost, does not offer the same guarantee of reliability. There are, therefore, two ways of setting the rate at which one sells digitized information: one can charge by the yard, so to speak, with a price based on the number of characters or other small units that comprise the data; or one can charge according to the value of the information. It is fairly easy to set rates by the yard; it is not so easy to decide what price the market might bear for certain kinds of information.

Nevertheless, these questions of economic and other forms of evaluation need to be considered in any model of global digital libraries. There needs to be a component, in addition to methods of identification and retrieval of data, that takes into account the very real problems that arise once data leaves the area where it is stored and becomes a marketable commodity. Since no one can imagine that information will be given away in the future, any more than it was in the past, a tagging method should be developed so that units of information can be associated with appropriate levels of pricing.

Creating Digital Libraries

There are many reasons for an organization to develop a digital library system. For one thing, digital libraries help preserve fragile documents without denying access to those who want study them. Another major benefit is that the user with can retrieve documents and information in seconds without having to deal with physical inconveniences. Also, a large number of people can consult the same document simultaneously. Finally, a major benefit of digital libraries is that they take up less room then physical books and magazines (Lesk, 1997). However, it is entirely possible that many of these benefits may be forestalled because of economic and social issues involving intellectual property rights and the prestige of traditional printed materials. Universal and free access to all materials will be more difficult to provide than is sometimes assumed.

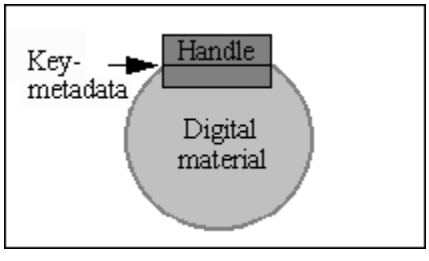
Considering all of these advantages, organizations might well decide to develop their own digital libraries. However many managers may change their minds when they find out how complex these can be to develop and maintain (Fox and Marchionini, 1998). They are so complex because adequate systems demand digital document preservation, distributed database management, information filtering, instructional models, intellectual property rights management, multimedia information services, resource discovery, and selective dissemination of information. However, once an organization decides to develop a digital library, it should make sure that the system is designed for easy usability by both the administrators and users.

A Digital Library Architecture

The main issues involved in the architecture of a digital library are data, metadata, and the processes of the database of the digital library (Nurenberg., Furuta, Leggett,J. Marshall, and. Shipman, 1995). Data is the actual information that is stored in the database, such as books, magazines, journals, video clips, images and sound. The metadata provides information on how the data is stored in the database. The processes are needed for managing the data in the database by adding, searching, retrieving, editing, and deleting. All three must be thoroughly linked in order to make the digital library more efficient.

The metadata used in the information architecture could also be based on three simple concepts that relate to data types, structural metadata and meta-object (Arms, Blanchi, and Overly, 1997). The data type describes the technical properties of the data. Its properties include the format that the data is stored under. It would also include information on the processing method of the data. The structural metadata is the metadata that describes the types, versions, and other characteristics of the data. The meta-object is an object that provides references to a set of digital objects and stores a list of handles for the other data. An example of how a meta-object could be represented is shown in figure 1.

Figure 1. A Digital Object

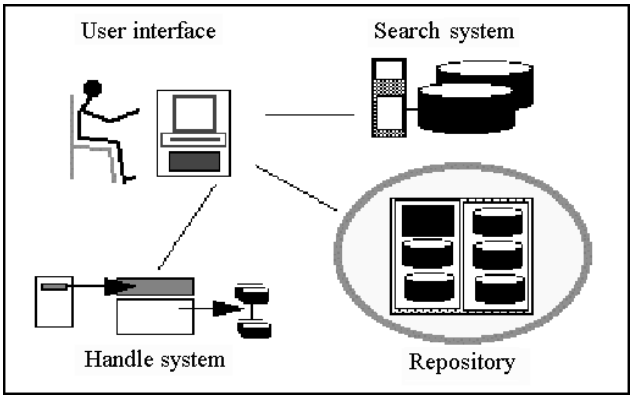


Arms, Blanchi, and Overly (1997) recommend a set of guidelines a developer might want to follow when developing this architecture. The first one is that all data should be given an explicit data type. The data type would specify that the data has a certain format, should be processed in a specific way, or has a specific organization. Another rule is that all the metadata should be encoded explicitly. The next rule is that handles should be given to each individual data of intellectual property, so that all metadata required to manage the data or provide access to the data is coded explicitly, since some data contains small pieces of data that could be used on its own. This rule indicates that smaller data should have its own metadata information. For example, if the data is an article from a magazine that contained a chart or illustration, the chart or illustration should have its own data type and metadata. This rule allows for greater flexibility and long-term control. Another rule is that

meta-objects should be used to aggregate digital objects. This is useful when a piece of data exists in several places within a repository, so that the meta-object for each data contains links to all versions of data and to all the structural metadata. The final rule is that handles should be used to identify items listed in meta-objects.

The key components of a digital library system should be interfaces, repositories, handle system, and search system, as seen in figure 2. The user interface is what enables the user to perform various processes on the digital library.

Figure 2. Major Digital Library System Components



Source: Arms, Blanchi, and Overly (1997)

The repositories are used to store and manage the data. The handles are identifiers that can be used to identify the data over long periods of time and to manage the data stored in the repositories. These could also be called metadata. Thus a handle system provides a distributed directory service for the handles of the data. It could be used to return the location of data in the repository when provided with the handle for the data. The search system is a method used by various indexes and catalogs in the repositories to find information.

Since most digital libraries contain information requiring large amounts of memory that might reach into the terabytes, developing a good architecture for the digital library improves its manageability. A data warehouse could be used to store the vast amount of data in a digital library. A data warehouse can be defined as a "subject oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process (Inmon and Hackathorn, 1994). The first characteristic of a data warehouse, its subject orientation, is defined by the major subjects of a library that make up most of its database. Some of the major subjects for a digital library may include users, librarians, publishers, and distributors. The second characteristic, cross-reference integration, implies consistent naming conventions, consistent measurement of variables, consistent encoding structures, and consistent physical attributes of data. Time variance is another characteristic of a data warehouse, because information in a digital library is accurate as of some moment in time. The final characteristic of a data warehouse, non-volatility, relates to the actual placement of data into the data warehouse and its accessibility.

Developing Efficient Multimedia Search and Retrieval Techniques

One of the major issues of digital libraries is search and retrieval of information. Most digital libraries would have a variety of sources of information. After performing a search on a large database, the user might get back a very large number of potential selections/hits. The user would then have to go through and examine these hits to determine whether or not that is what he was looking for (Paepcke, 1996). An example might be if a person was trying to find all information related to a topic like paper feed design.

One way of helping the user perform his search is by having an effective user interface (Byrd, Croft, and Shneiderman, 1997). The user might be able to narrow down his search if it gave the location where his source might be found. For example, instead of looking in the databases of every library on campus, a student may choose a specific library to search, such as a general, business, or engineering library. Another way to help narrow down the search might be by providing multiple search fields. If a person was looking for a phone number, the search could have a separate field for first name, last name, middle name, city, and state. Also if the user is not sure of the correct spelling or does not know the complete name of subject he is searching for, there should be variants that could help out with these problems.

Another issue with searching and retrieval is support for heterogeneous systems. If the search coding is not the same for both systems, the user might not be able to perform the search in the other systems. The main way of overcoming this problem would be if all systems used a standard protocol to develop their search coding. Fortunately, the International Organization for Standardization has approved a standard called ISO 23950 (Moen, 1998). This was taken from an American National Standard Interface (ANSI) protocol known as ANSI/NISO Z39.50-1995. This standard may be used to enable uniform access to textual and non-textual digital collections.

Besides having text materials, most digital libraries have information in multimedia format, such as graphics, images, audio and video. Searching for multimedia information would require that the user search by indexes of captions. For example, if the user wanted to find an image he had in mind, he would need to know the caption that the image was stored under. This would be a problem if the user had an idea of what the image looks like but

did not know the name of the caption the image is stored under.

One way to overcome this problem is to use image analysis (Stix, 1997). The user scans in an image and asks the system to search for any images in its digital library that are similar to it. An advantage of this approach is that the user retrieves images regardless of the caption of the image. This type of search would be narrowed down to find items in the digital library such as images with acronyms like GIF or JPEG files. The user may want to narrow down a search by using words to describe the image. If the user wanted to find images of a black cat, for instance, he or she would scan in an image of a cat, then describe what color it should be. Another way to help with the search is by using low-level visual features of the image as a handle for the image. The system would then search for a match between what the user scanned in and the handle of the low-level visual features of the images. A handle could also be made to describe their spatial and temporal attributes. A prototype system called WebSeek has been developed to try to do this (Chang et al., 1997). WebSeek analyzes images and videos in two separate automatic features. In the first one the visual features are extracted and then indexed offline. Then the associated text is parsed, and utilized to classify the images into subject classes in a customized image category.

Developing Interoperability of the Digital Library

When designing a digital library, one must address the key issue of interoperability. Most organizations develop their digital libraries separately from other organizations. This isolation tends to make digital libraries structurally different from one another. As a consequence, users might not be able to retrieve data from digital libraries other than the one in use. In order to prevent this situation, digital libraries should be interoperable with each other (Chang et al., 1997).

Interoperability allows the user to subscribe to a publication and be connected to the publisher's digital library. The library does not have to convert physical data into digital. Another good point about this technique is that the digital library gets to use the information as soon as the publisher puts it in. However, in order to do so, the developer needs to make his system compatible with that of the publisher. A likely situation is that the library would need to be connected with many different publishers who might be using different standards. One solution to this problem is to create specifically dedicated hardware for viewing files and a uniform method of text encoding. Such a strategy was adopted with the generation of e-books that came on the market in early 1999.

However, a potentially more useful solution might be to employ the international ISO 23950 standard set by the International Organization for Standardization. This is an open standard that allows for communications between systems that have heterogeneous hardware and software (Turner, 1997). This standard also helps overcome the problem associated with multiple database searching such as knowing the unique menus and search procedures of each system being accessed, by allowing users to use the familiar interface of their own system to search multiple databases.

Conclusion

The development of a digital library is a complex task. However, for the digital library to be most beneficial to the user it needs to be managed on an ongoing basis after it has been built. The information that an organization decides to place into the digital library should be of the kind that would most help its users in their studies and research. This is also a difficult task, since they would have to determine what is being used and what is not. When an organization is about to develop a digital library it usually has a large amount of information in both hard copy and in digital copy. After the digital library is built, the organization will have to decide on the publications and/or which part of the publication they need in a digital format. Once the organization has selected the books, journals, magazines and other sources of information that it wants to put in to the database, it needs to select the best technology to scan the written words on paper to digital code on the computer, and store it using magnetic or optical storage technology. Indexing and other access software must be developed that will allow some of the process of archiving and retrieving to be automated at all stages of organization and production. Last, but not least, the administration of the digital library must be carefully considered. Changing hardware and software, along with new emerging standards, may require long-range strategies for migration to different platforms and forms of encoding. Unlike printed materials, digital information can be rendered inaccessible if coding conventions change or hardware becomes obsolete. In earlier times one could depend on the unchanging character of the human eye; digital information requires hardware as an intermediary between eye or ear and the data itself. The manager of a digital library must be ready to anticipate rapid changes in methods of storage and be prepared to transform large databases so that they are usable decade after decade.

Digital libraries may not save us money in the long run. After all, costs are part of the system and have a tendency toward self-preservation. What benefits they do promise, however, are a vast improvement over older modes of accessing, storing, and transporting documents. As our collective habits change, we will find ourselves compelled to move in this direction whether we like it or not. The models presented here are meant to suggest guidelines for making the inevitable transition smoother.

References

Arms, W., Blanchi, C., and Overly. "An Architecture in Digital Libraries," *D-Lib Magazine*, www.dlib.org/dlib/february97/cnri/02arms1.htm, (Current February 1997).

Bush, Vannevar. "As We May Think," *Atlantic Monthly*, July, 1945, pp. 101-108.

Byrd, D., Croft, W. and Shneiderman, B. "Clarifying Search: A User Interface for Text Searches," *D-Lib Magazine,* www.dlib.org/dlib/january97/retrievel/01shneiderman.html, (Current January 1997).

Chang, C., Molina,H. Paepcke, A., and Winograd. "Interoperability for Digital Libraries WorldWide," *Communications of the ACM* (41:4), 1998, pp. 33-43.

Chang, S., Meng, H., Smith, J., Wang, H., and Zhong, D. "Finding Images/Videos in Large Archives," *D-Lib Magazine*, http://www.dlib.org/dlib/february97/columbia/02chang.html, (Current February 1997).

Elliot, M. and Kling, R. "Digital Library Design for Usability," http:csdi.tamu.edu/csdi/DL94/paper/kling.htm, (Current 1994).

Fox, E. and Marchionini,G. "Toward a Worldwide Digital Library," *Communications of the ACM* (41:4), 1998, pp. 29-32.

Inmon, W. and Hackathorn, R. *Using the Data Warehouse*, John Wiley & Sons, Inc., New York,.NY, 1994.

Lesk, M. "Going Digital," *Scientific American*, http://www.sciam.com /0397lesk.html, (Current March 1997).

Johnson, Elmer D. *History of Libraries in the Western World,* 2nd ed., The Scarecrow Press, Metuchen, N. J., 1970.

McLuhan, M. H. and McLuhan, E. *Laws of Media: The New Science*, Univ. of Toronto Press, Toronto, 1998.

Moen, W. "Accessing Distributed Cultural Heritage Information," *Communications of the ACM* (41:4), 1998, pp. 45-48.

Nurenberg, P., Furuta, R., Leggett, J., Marshall, C., and Shipman, F. "Data Libraries: Issues and Architectures," http://csdi.tamu.edu/csdi/DL95/papers/ nuernberg/nuernberg.htm, (Current 1995).

Paepcke, A. "Digital Libraries: Searching Is Not Enough," *D-Lib Magazine*, http://www.dlib.org/dlib/may96/stanford/05paepcke.html, (Current May 1996).

Stix, G. "Finding Pictures on the Web," *Scientific American*, http://www.sciam.com/0397/lynchbox1.html, (Current March 1997).

Turner, F. :An Overview of the Z39.50 Information Retrieval Standard," http://www.ifla.org/VI/5/op/udtop3/udtop3.htm, (Current 1997).