COLLECTION DEVELOPMENT AND PRESERVATION POLICIES IN DIGITAL ENVIRONMENT

By

M Natarajan**
G N Narasimhan*

ABSTRACT

This article describes the concept of digital library along with the resources available in it. It deals with the collection development and preservation policies for digital environment. The need for collection development and preservation are explained in detail. The digital preservation, its requirement and proposed strategies are given in detail. The preservation modes are explained with economic and licensing issues. The policies and procedures for accessing Internet are discussed in detail.

** Scientist-in-charge, INSDOC, Chennai, CSIR Madras Complex, Tharamani, Chennai 600 113.
* Scientist, INSDOC, Chennai, CSIR Madras Complex, Tharamani, Chennai 600 113.

0. Introduction

Libraries are organized to facilitate access to controlled collections of information. Traditional libraries (TL's) possess three organizational characteristics that, together, provide a basis for such access. These are:

- The organization of information into physical information objects (IO's) such as books;
- The physical organization of the collections of IO's according to various attributes, such as subject matter and author;
- An organized information environment that facilitates direct access to the IO's based on such attributes as author, title, and subject matter, as well as a limited degree of indirect access to the information contained in the IO's.

This last characteristic of a TL typically involves multiple sources of information to support access, such as librarians, catalogs, and the manner in which the collections are organized physically. Since it involves information about information, which can be termed as the meta-information environment of a library.

1. Digital library (DL)

A librarian, a computer scientist, an educator, a journal publisher, or a Web master will each have a different perception of what a digital library is from their point of view. For the majority of participants in the technical construction of the infrastructure that provides access to information on the Internet or within the World Wide Web, and for most of the users of that world, a digital library is simply a collection of information stored in electronic format.
Definitions:

- Terence R Smith, defined digital library as " Controlled collections of Information Bearing Objects (IBOs) that are in digital form and that may be organized, accessed, evaluated and used by means of heterogeneous and extensible set of distributed services that are supported by digital technology".

- Clifford Lynch, a well-known expert on Internet and web technologies, has defined digital library as " system providing a community of users with coherent access to a large, organized repository of information and knowledge. The digital library is not just one entity, but multiple sources seamlessly integrated".

- Digital libraries are electronic libraries in which large number of geographically distributed users can access the contents of large and diverse repositories of electronic objects. Electronic objects include networked text, images, maps, sounds, videos, catalogues and scientific, business and government datasets.

- A digital library is understood to have the information stored predominantly in electronic or digital medium. It is expected to provide access to the digital information collection.

- Digital libraries are libraries in which the controlled collections are in digital form and access to the information in the collections is based almost entirely on digital technology.

From a user's point of view, digital technology changes the three organizational characteristics of TL's:

- The organization of information into physical IO's is replaceable with a more flexible organization into logical IO's.

- The single physical organization of a collection of IO's is replaceable with multiple logical organizations of IO's.

- The most significant changes, however, occur in the meta-information environment of a library. In terms of advantages, having the IO's in digital form permits the use of digital technology in extracting information from the IO's.

The extracted information may satisfy a user's ultimate need for information or it may be employed by "digital librarians" in characterizing the IO's in the collection. In the latter

case, this meta-information may be employed in providing access to the information encoded in the IO's. In terms of disadvantages, important interactions between librarians and users that occur in the meta-information environments of TL's may be lost with the near-automation of information access in DL's. This article describes the collection development and preservation policies in the digital environment and its implications.

2. Collection development

Libraries strive to develop collections, resources, and services that meet the cultural, informational, educational, and recreational needs of our community. The Internet, as an information resource, enables libraries to provide information beyond the confines of its own collection. It is in this context, that libraries offer access to the Internet.

2.1 Additional Collecting Levels

Print-based collecting level designations are still useful within the digital realm, but for digital collections even more information is required. It is proposed to keep four levels of collecting, which may also include designation of preservation commitment.

- Archived - The material is hosted here, and the library intends to keep the intellectual content of the material available on a permanent basis.
- Served - The material resides here, but the library has not (yet) made the level of commitment to keeping it available that it has for "archived" materials.
- Mirrored - A copy of material residing elsewhere is hosted here, and the library makes no commitment to archiving. Also, an institution other than the library has primary responsibility for the content and its maintenance.
- Linked -The material is hosted elsewhere and the library points to it at that location. Therefore the library has no control over the information.

Material in any category except archived may be re-designated from one level to another as required to meet changing information needs, remote server accessibility or responsiveness, local resource demands, etc. Material that receives the archived designation cannot be downgraded to a lower status.

3. Digital preservation

"Digital preservation" or "digital archiving" means taking steps to ensure the longevity of electronic documents. It applies to documents that are either "born digital" and stored on-line (or on CD-ROM, diskettes or other physical carriers) or to the products of analog-to-digital conversion, if long-term access is intended.

4. Collection development policy

A collection policy is a standard library practice for publicly declaring a library's intent for breadth and depth of the material it will collect within certain subject areas, genres, or physical formats. Such declarations are useful tools that scholars can use to determine the

relative utility of a collection for their purposes, as well as to assist in cooperative collection development with other libraries. Typical collection development categories for print collections include such as:

- ? Comprehensive - A collection in which a library endeavors, so far as is reasonably possible, to include all significant works of recorded knowledge (publications, manuscripts, other forms), in all applicable languages, for a necessarily defined and limited field.
- ? Research - A collection, which includes the major published source materials required for dissertations and independent research, including materials containing research reporting, new findings, scientific experimental results, and other information useful to researchers. It also aims to include all important reference works and a wide selection of specialized monographs, as well as a very extensive collection of journals and major indexing and abstracting services in the field.
- ? Study - A collection which is adequate to support undergraduate and most graduate course work; that is, which is adequate to maintain knowledge of a subject required for limited or generalized purposes, of less than research intensity.
- ? Basic - A highly selective collection which serves to introduce and define the subject and to indicate the varieties of information available elsewhere.
- ? Minimal - A subject in which few selections are made beyond very specific works.

5. Preservation Requirements

Preservation measures ensure that a document is accessible in a usable form over time. Maintaining the accessibility of digital media, is much more complex than with such non-digital media as paper. For example, when a book is preserved in its original format, all aspects of the book are preserved, its format, its layout, and its content. It is practically impossible to extract individual elements (e.g., content without layout) because they are inextricably linked. Even reformatting to paper or microfilm does not completely divorce content from layout as page sequences and physical appearance, for instance, can still be captured. Digital objects, in contrast, are easily decomposed into individual elements, and significantly more effort must be made to preserve them as a "whole." For example, one can retain the content of an electronic document, while losing the layout. Further, one can keep its physical presence (i.e., a file), but fail to preserve its readability.

In the digital world, the first task is to identify the multiple aspects of a work that must be preserved. Next, to succeed in the preservation of digital objects, preservation measures must ensure that as many of these aspects as possible persist over time. In preserving a digital object, aim at the following:

- ✍ Fix the object as a discrete whole. The boundaries of digital objects are less clear, especially if they are compound objects created by assembling different media or by linking to resources from around a network.

- ✍ Preserve the physical presence. The physical presence refers to the computer file and does not mean that the object will remain accessible.
- ✍ Preserve content. Refers to maintaining the ability to access the content at its lowest level, such as ASCII text, without the embellishments of font variations and layout features.
- ✍ Preserve the presentation. Content is typically rendered in some presentation, format or layout. In many types of digital documents, the layout specifications are separate from the content. To retain the original look of a document, these layout specifications must also be preserved, especially when they contribute significantly to the understanding and interpretation of the content.
- ✍ Preserve functionality. Digital objects can contain multimedia components (i.e., text, graphics, audio, and video). Special efforts must be made to preserve the functionality.
- ✍ Preserve authenticity. An individual accessing the object must be able to verify that it is what the individual wanted and that the transformations to keep it accessible have preserved its original form.
- ✍ Locate and refer to the digital object over time. Digital objects can be readily altered, copied or moved. An individual must be able to match a citation to a digital object, and to distinguish it from other versions or editions.
- ✍ Preserve provenance. Provenance is an archival concept that asserts the origin and chain of custody of an object and contributes to defining it as a whole. Establishing an object's origin and history help confirm that the work is authentic and its content is intact.
- ✍ Preserve context. Digital objects are partly defined by their hardware and software dependencies, their mode of distribution and linkages to other digital objects. Preserving context is a particular challenge.

5.1 Preservation modes

The electronic repository must be preserved. Preservation of information needs to be looked at from at least three points of view:

- ✍ Medium preservation
- ✍ Technology preservation and
- ✍ Intellectual preservation.

The electronic information must now be dealt with separately from its medium. This can be illustrated by an analogy, one which is very oversimplified, as readers will be aware: if a book is placed on a closet shelf, and the closet door is closed for 500 years, then at the end of that time one can, still open that door and read the book. With an electronic resource one does not have that confidence after ten years due to several reasons.

5.1.1 Medium preservation

Medium preservation is the concern for preserving the medium on which information is stored, such as tapes, disks, optical disks, CD-ROMs and the like. Backup is appropriate, as is copying to other devices of the same kind, a technique that is known

as "refreshing". Refreshing a tape means by copying its contents to another similar tape. In the current climate of protection of intellectual property rights, copyright concerns must be recognized.

5.1.2 Technology preservation

More problematic than medium decay are the rapid changes in the means of recording, in the storage formats and in the software that allows electronic information to be of use. One has to be aware of technology obsolescence as even more of a problem than medium decay, and undertake steps of technology preservation. Rather than simply refreshing, the migration of information forward through technology stages as they become available and as the old technologies cease being supported by vendors and the user community.

5.1.3 Intellectual preservation

Intellectual preservation addresses the integrity and authenticity of the information as originally recorded. Preservation of the media and of the software technologies will serve only part of the need if the information content has been corrupted from its original form, whether by accident or design. The need for intellectual preservation arises because the great asset of digital information is also its great liability; the ease with which an identical copy can be made, quickly and flawlessly, is paralleled by the ease with which an undetectable change may be made.

5.2 Proposed preservation strategies

Several strategies attempt to address the primary digital preservation problem of technological obsolescence. These include migrating information through successive generations of technology; using software to emulate the behavior of older machines; preserving original hardware and software to run obsolete programs, and creating hard copies (paper or microform) of digital objects. Each of these strategies meets certain preservation goals.

5.2.1 Migration

Migration is the primary strategy articulated by most organizations that plan to preserve digital objects. It covers a range of activities to periodically copy, convert or transfer digital information from one generation of technology to subsequent ones. Migration may involve copying digital information from a medium that is becoming obsolete or physically deteriorating to a newer one (e.g., floppy disk to CD-ROM), and/or converting from one format to another (e.g., Microsoft Word to ASCII), and/or moving documents from one platform to another (e.g., VAX to UNIX). Migration certainly preserves the physical presence and the content of a digital object. However, it may not preserve presentation, functionality and context.

5.2.2 Emulation

Emulation refers to creating new software that mimics the operations of older hardware or software in order to reproduce its performance. Thus, not only are physical presence and content preserved, but digital objects could display original features and functionality available with the older software. Emulation has recently attracted attention

as a potential strategy to assist preservation, recognizing that some electronic material that is highly dependent on particular hardware and software will not lend itself to migration. However, emulation for preserving digital objects over the long term has not been widely tested or priced.

5.2.3 Output to permanent paper or microfilm

Outputting a hard copy of a digital file is a "low tech" solution that can result in a well-standardized product with a life expectancy of several hundred years. Certainly, this strategy could fix the object as a whole and preserve content and to some extent layout.

5.3 Problems in digital preservation

The fundamental problems are:

- Accessible only by using combinations of computer hardware and software.
- Present hardware and software can become obsolete after sometime.
- Requires currency with technology changes.
- To move digital objects from obsolete to current file formats,storage media, operating systems and so on..
- The rapidly increasing number of digital objects and proliferation of document standards and formats.
- The increasing complexity of digital objects (incorporating text, images, audio, video in various formats) and their increasing software dependence (e.g., storage in databases).
- The lack of planning to incorporate preservation needs in systems and lack of availability of off-the-shelf products supporting preservation needs.
- Copyright/intellectual property rights that may interfere with the ability to preserve digital objects through systematic copying.

6. Economic and licensing issues

Some materials in the library will be openly available. Others will be commercial products. Most core educational materials are created commercially as business ventures.

Copyright has been used as the mechanism by which materials are prevented from being misused. At present there is intensive debate about the form that copyright should take in digital libraries. One opinion is that this is fundamentally an economic debate. Whatever legal framework develops will enable the owners of educational materials to control their use, set terms and conditions, and price them, as the market will bear.
Materials are paid for in three different ways. The first is by the student directly, through purchasing books, photocopies, computer software, lab fees, etc. The second is by the educational establishment, through its library, computing, and media budgets: The third is by the producer of the materials, such as by creating Web sites.

1. Controls on access to materials. A change in the balance between these three methods of payment is happening particularly with the growth of freely available scientific research and other resources over the Internet. Thus one can expect that large amounts of good material will not require payment, but the library must be built around a framework that permits control of access to materials if required by the owner.
2. Controls on accuracy. The principal reason that authors and publishers wish to control educational materials is the desire to make money. A secondary reason is the wish to control the content, in particular to ensure accurate representation of the ideas and concepts, with appropriate attribution. One approach to this issue is to register each item as it is added to the collection with a unique identifier and a digital signature, which can be used to verify that an item has not changed

Policies and Procedures

To introduce and reiterate, the following policies and procedures are to be followed for Internet access:

- ✍ All networked personal computers will be distinguished and equipped with a blocking software application that will limit exposure to websites.
- ✍ Students will be asked to close a website that is depicting or discussing information that is "harmful to others".
- ✍ Staff should not attempt to limit Internet access to information, especially when it is made available in other formats in the library. Rated materials are equally available to students.
- ✍ Patrons may print all information within reason (text or graphics) that is needed as long as the information is not obscene.
- ✍ Use of the Internet access computers is on a first-come, first-served basis. Users agree to limit their total time to 30 minutes if someone is waiting.
- ✍ The cost for printing can be fixed at nominal cost or download files to a disk
- ✍ Information provided on the Internet is frequently protected by copyright. The user is responsible for the appropriate use of material.
- ✍ Sometimes users cannot always gain access to places on the Internet. Reasons?
    - o Host computer closed.
    - o Site is licensed for subscribers.
    - o Technical difficulties on the Internet or the Library connection.
    - o Too many Internet visitors.
    - o Host computer has changed address.

The intent of these policies and practices is to protect both the freedom and rights of the individual students in their pursuit of education and information and at the same time seem to protect the legal position and practice of the library. The library recognizes that these policies and procedures do not solve or end the issue of inappropriate Internet use. They do serve as guidelines for user and staff alike.

7. Conclusion

The collection development and preservation environment of a library is the aspect of library structure that is likely to be most affected by DL's technology. It is important to design information environment for DL's that simultaneously compensate for the loss of

many of the services of librarians and take advantage of the ability to apply digital processing to information objects in the collection of DL's. Such an environment is probably best implemented within a distributed object framework. The collection and preservation policies given in this article may not be exhaustive. The digital collection development and preservation techniques for these types of materials go on changing. Librarians do not have control over the web based resources and the authenticity of them. However librarians should take it as a challenge and move with information technology applications.

8. References

1. Arms, William Y - A National Library for Undergraduate Science, Mathematics, Engineering, and Technology Education - Needs, Options, and Feasibility.
2. Arora, Jagdish - Digital Libraries: An Overview. Paper presented in the " National Seminar on Knowledge Networking in Engineering and Technology Education and Research, Delhi" during Dec 01-02, 2000 at IIT, Delhi.
3. Bullock, Alison - Preservation of digital information: issues & current status Network Notes #60 , IT series, National Library of Canada, April 22,1999
4. Information Technology & Libraries: Federal Roles Library and Information Technology Association, American Library Association, Chicago, IL, May 1991.
5. Linch, Clifford and Hector Garcia-Molina, IITA Digital Library Workshop, Reston, VA, May 18-19,1995.
6. Terence R Smith - The Meta information environment of Digital Libraries,1997.
7. www.cclib.org/main.htm
8. www.nlc-bnc.ca/publications/netnotes/notes60.htm
9. www.dlib.org