Digital Collections: Preservation and Problems

Mange Ram J K Mishra

Abstract

In educational and research institutions, the quantity of digital content created is huge, and libraries at these institutions have a liability to bring this digital material under guardian control in order to manage and preserve it over time. These all have presented challenges and problems to libraries because of the need for special equipment to display items in these formats, obsolescence of this equipment and/or the formats, and the need to preserve the information contained on sometimes fragile storage media. Now the Libraries are working with the DSpace, Greenstone etc (open-source digital repository platform) to explore the problem of capturing research and teaching material in any digital format and preserving it over time. It is a daunting task with few confirmed models, requiring new technology, policies, procedures, core staff competencies, and cost models Problems are not ending; now burning problem is what we should be selected for Preservation for future as this task is not trouble-free. Today the problem of harvesting material on the web and in information centres for preservation and search is a major concern.

Keywords: Digital Preservation, Digitisation

1. Introduction

Now the library materials was extended to include information stored on other physical carriers such as microfilm, film of various types, audio cassette tapes, video tapes, computer disks, CD-ROMS, and DVDs. These have all presented challenges and problems to libraries because of the need for special equipment to display items in these formats, obsolescence of this equipment and/or the formats themselves, and the need to preserve the information contained on sometimes fragile storage media. With the development of the World Wide Web in 1993, which opened up online publishing as an easily available, ubiquitous, and relatively inexpensive means of creating and distributing information, national and other deposit libraries accepted that, once again, they must expand their roles to encompass this new form of publishing and all that its collection, description, storage, management, preservation, and provision of access entailed. There are additional challenges to face over and above those inherent in the formats that they already collected. The volume of online publishing is huge. Almost anyone can set themselves up as a publisher, meaning that issues of quality and authority of information need to be addressed, as well as a wide range of competence (or otherwise) in using publishing software and compliance in applying standards. In addition, many of these items are complex Web objects—for instance, Web sites that contain a number of different file formats - and this makes strategies for preservation particularly difficult to formulate and undertake. What Should Be Collected and Preserved?

2. Definitions of Key Terms

- A record is a document made or received and set aside in the course of a practical activity. Thus a record is not merely private information. As one of our key informants put it, a record has a "fixed content" that can be "re-presented in the structure or form in which it was born."
- An electronic record is a record that is created (made or received and set aside) in electronic form.
- An authentic record is a record that is what it purports to be and that is free from tampering or corruption.
- A digital record is one that now exists in electronic form, though it may or may not have been created in electronic form. A digital record may have been created on paper and digitized later. Subsequent digitization may change its "recordness."
- A digital component is a stored digital object that is necessary to reproduce an electronic record (or other digital asset).
- Digital assets and digital content refer to all types of information, text, graphic, image, and multimedia.

3. Digitization

Digitization is the process of representing an object, an image, or a signal (usually an analog signal) by a discrete set of its points or samples. The result is called "digital representation" or, more specifically, a "digital image", for the object, and "digital form", for the signal.

Digitization refers to the conversion of non-digital material to digital form (i.e. a form which uses a binary numerical code to represent variables). A wide variety of materials as diverse as maps, manuscripts, moving images and sound may be digitized.

In the past few years, procedures for digitizing books at high speed and comparatively low cost have improved considerably with the result that it is now possible to plan the digitization of millions of books per year for creating digital libraries [1].

4. Why Active Preservation is Necessary

Society's heritage has been presented on many different materials, including stone, bamboo, silk, paper and etc. Now a large quantity of information exists in digital forms, including emails, blogs, social networking websites, national elections websites, web photo albums, and sites which change their content over time. According to a report by the US Library of Congress, 44% of the sites available on the internet in 1998 had vanished one year later.

The unique characteristic of digital forms makes it easy to create content and keep it up-to-date, but at the same time brings many difficulties in the preservation of this content. Margaret Hedstrom points out that "digital preservation raises challenges of a fundamentally different nature which are added to the problems of preserving traditional format materials."[2]

5. Strategies

In 2006, the Online Computer Library Center (OCLC) [3] developed a four-point strategy for the long-term preservation of digital objects that consisted of:

- Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications.
- Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied.
- Determining the appropriate metadata needed for each object type and how it is associated with the objects.
- Providing access to the content.

There are several additional strategies that individuals and organizations may use to actively combat the loss of digital information.

5.1 Refreshing

Refreshing is the transfer of data between two types of the same storage medium so there are no bitrates' changes or alteration of data [4]. This strategy may need to be combined with migration when the software or hardware required to read the data is no longer available or is unable to understand the format of the data. Refreshing will likely always is necessary due to the deterioration of physical media.

5.2 Migration

Migration is the transferring of data to newer system environments (Garrett et al., 1996)[5]. This may include conversion of resources from one format to another (e.g., conversion of Microsoft Word to PDF or Open Document), from one operating system to another (e.g., Solaris to Linux) or from one programming language to another (e.g., C to Java) so the resource remains fully accessible and functional.

5.3 Replication

Creating duplicate copies of data on one or more systems is called replication. Data that exists as a single copy in only one location is highly vulnerable to software or hardware failure, intentional or accidental alteration, and environmental catastrophes like fire, flooding, etc. Digital data is more likely to survive if it is replicated in several locations. Replicated data may introduce difficulties in refreshing, migration, versioning, and access control since the data is located in multiple places.

5.4 Emulation

Emulation is the replicating of functionality of an obsolete system (Rothenberg, 1998)[6]. For example, emulating an Atari 2600 on a Windows system or emulating WordPerfect 1.0 on a Macintosh.

Emulators may be built for applications, operating systems, or hardware platforms. Emulation has been a popular strategy for retaining the functionality of old video game systems.

5.5. Metadata Attachment

Metadata is data on a digital file that includes information on creation, access rights, restrictions, preservation history, and rights management [7]. Metadata attached to digital files may be affected by file format obsolescence. ASCII is considered to be the most durable format for metadata because it is widespread, backwards compatible when used with Unicode, and utilizes human-readable characters, not numeric codes. It retains information, but not the structure information is presented in. Now a days many libraries use Dublin Core a metadata standard.

6. Selection of the DL Software

The software selection based on set parameters is an uphill task, as the technology itself was still emerging only. In general, what is desirable is a system that is flexible enough to fit the current digital information system as above and to accommodate future migration. That the system should be in a position to embrace almost all predominant and emerging digital object formats and capable of supporting the standard library technology platforms, should be the major focus. There are many digital library softwares available, proprietary as well as open source, and most of them conform to international standards. As mentioned earlier, VTLS and ACADO are the commercial ones available and popular in the Indian market. Some of the popular Open Source Softwares for digital libraries, which are in use internationally, are 'DSpace', 'Dienst', 'Eprints', 'Fedora', 'Greenstone' etc.

6.1 Greenstone

The Greenstone (http://www.greenstone.org) Digital Library Software (GSDL) is a top of the line and internationally renowned Open Source Software system for developing digital libraries, promoted by the New Zealand Digital Library project research group at the University of Waikato, headed by Dr. Ian H. Witten, and is sponsored by the UNESCO. The software is issued under the terms of GNU General Public License. Greenstone provides a way of building, maintaining and distributing digital library collections, opening up new possibilities for organizing information and making it available over the Internet or on CD-ROM. One of the pioneering and laudable efforts of Greenstone is its commitment to lower the bar for construction of practical digital libraries, yet at the same time leave a great deal of flexibility in the hands of the user. Greenstone is being used by lot of libraries and institutions across the world.

6.2 EPrints

The GNU EPrints is free software which creates online archives. The default configuration is a repository of the research output of an academic institution. An EPrint archive can be adapted for many more purposes. It has been developed at the University of Southampton in relation to a variety

of projects. The latest recommended version is 'eprints-2.3.11'. In India NCSI has been conducting workshops and training programmes for installing and customizing EPrints software.

6.3 DSpace

DSpace (http://www.dspace.org) is a digital repository platform jointly developed by Hewlett-Packard and MIT Libraries collaborating over two years. DSpace provides the basic functionality required to operate an institutional digital repository, and is intended to serve as a base for future development to address long term preservation and access issues. On November 2002, the system was launched as a live service hosted by MIT Libraries, and the source code made publicly available according to the terms of open source license, with the intention of encouraging the formation of an open source community around DSpace Initial developments in this area have been very promising.[8]

7. Problems/ Challenges

7.1 Digital Collection Harvesting Problem

Today is the problem of harvesting material on web and in information centres for preservation like a search / selection of imperative pearl (valuable information) in the sea (WWW. There may be varied approach of selection according to needs. Some material may be very classical, important and some may be dynamic (text books etc). Some approaches are the followings[9]:

7.11 Selective Archiving Approach

Each item in the archive is quality assessed and functional to the fullest extent legally recognized by current technical capabilities. A gathering schedule can be individually adapted for each selected title, taking into account its publication schedule or the frequency with which the Web site changes, thus enabling the content gathered to be as complete as possible. Each item in the records can be fully catalogued and therefore can become part of the national bibliography. Each item in the archive can be made accessible via the Web to readers immediately because permission to do so can be negotiated with publishers. The "significant properties" of individual resources and classes of resources within the archive can be analyzed and determined. (These are the attributes that convey the full meaning and intellectual content of an item and enable it to be experienced as the creator intended.) This enhances our knowledge of preservation requirements and enables risk assessments and preservation strategies to be put in place.

Disadvantages

- Though we believe that we are selecting titles based on sound professional experience and judgement, do we really know what will be important for future researchers? Selection is largely based on a traditional understanding of the concept of "publication." Perhaps in the future this will not be as relevant, or, perhaps more likely, something in addition to this traditional approach will also be required.

- The selective approach is very labour-intensive, and the unit cost per item is therefore high.
- The selective approach takes a resource out of context and often does not include other resources to which it is linked. Appropriate meaning is therefore lost, and this will be more critical for some resources and research requirements than others. The value of "sampling" is as yet unproven. Will this approach satisfy the majority of research needs for these kinds of resources in the future?

Australia is the only known country with an established program for archiving dynamic as well as static publications and Web sites on a selective basis, once again with a high degree of intellectual input from library staff.

7.12 Whole Domain Harvesting

This involves harvesting not only all the resources in the specific country domain but also identifying those of country origin or subject matter in.com and other generic domains. In theory, the obvious advantage of the whole domain harvesting approach is that the whole domain is captured automatically at periodic intervals with minimal human intervention and therefore comparatively low staff cost per item gathered. The whole domain is available to future researchers, and resources can be seen in their broader context, with links to other documents retained.

Disadvantages

In practice this ideal is a long way from being the reality. Because whole domain harvests are demanding in relation to computer time and storage space, they are usually run at intervals of at least a couple of months. Any publications that come into being and disappear in the interim are missed. Any changes to existing sites that are made and overwritten in that period will also be missed. Because of the huge volume of publications involved, quality control checks cannot be made on more than a very small sample of titles. Nationally significant material is likely to be missing, and the archive administration will not be aware of it.

The National Libraries of Sweden, Finland, Iceland, Norway, and more recently Austria have been pursuing this approach

7.13 Hybrid Approaches

All of the approaches discussed so far have disadvantages the selective approach misses material that may be of future value, the whole domain model is not as comprehensive as its name would suggest, and collaborative agreements with publishers to date exclude the majority of publishers and a lot of freely available resources. A multi-pronged approach that combines a periodic snapshot of a country's domain, supplemented by selective archiving of nationally significant, authoritative publications of long-term research value and provision for deposit of publications by agreement with specific publishers, would be ideal. Funding is an issue, however, with each approach having its own technical infrastructure and staff support costs.

7.2 Challenges

While digitization offers great advantages for access, allowing users to find, retrieve, study and manipulate material, reliance on digitization as a preservation strategy could place much material at risk. Rapid obsolescence of digital technologies and media instability render the digitized object vulnerable to loss. The first challenge digital preservation faces is that the media on which digital contents stand are more vulnerable to deterioration and catastrophic loss. While acid paper are prone to deterioration in terms of brittleness and yellowness, the deterioration does not become apparent in at least six decades; and when the deterioration really happens, it happens over decades too. It is also highly possible to retrieve all information without loss *after* deterioration is spotted. The recording media for digital data deteriorate at a much more rapid pace, and once the deterioration starts, in most cases there is already data loss. This characteristic of digital forms leaves a very short time frame for preservation decisions and actions.

7.21 Digital Obsolescence

This challenge is exacerbated by the lack of established standards, protocols, and proven methods for preserving digital information[10]. We used to save copies of data on tapes, but media standards for tapes have changed considerably over the last five to ten years, and there is no guarantee that tapes will be readable in the future[11]. Hedstrom further explained that almost all digital library researches have been focused on "architectures and systems for information organization and retrieval, presentation and visualization, and administration of intellectual property rights" and that "digital preservation remains largely experimental and replete with the risks associated with untested methods". While the rapid advance of technology threats access of digital contents in length, the lack of digitizing standards affects the issue in width.

7.22 Cost

There is always a cost in its creation, its production, and its dissemination. Digital libraries introduce new and uncertain economic realities and relationships into libraries. Where the costs of accessing information were once hidden to patrons, the digital era is likely to require customers who will be required to pay fees for access to digital services and collections. The major obstacle is digitization. Digitization is very cost intensive. Especially when one goes single handed toward digitization.

7.23 Organizational

For building and working with Digital Library the long-term organizational, fiscal, and institutional commitments will be necessary. Management of the technical infrastructure for "digital library" services will be a significant obstacle for most libraries, especially as budgets continue to shrink and the costs of developing and maintaining collections increases. Administration of the digital collections locally, is harder and more expensive than managing a comparable print collection.

7.24 Intellectual Property Right

An intellectual property right is a big barrier for preserving the digital documents. Copyright protects the owner's creative or intellectual work. Digitization of documents are involved with complex method for resolving the legal and practical questions of migrating intellectual property, that includes the creators and owners of intellectual property, managers of digital archives, and actual and potential users of intellectual property[12].

7.25. Lack of Expertise

Digital library are considered by many to be a challenging area. The development of an infrastructure for the networked resource discovery and retrieval of highly distributed, autonomously created, and diverse electronic information is required. Above all, this infrastructure will need to be managed by professionals who understand information needs and uses.

8. Conclusion

A strategy with defined selection priorities for digitization is critical and should be informed by a convergence in the consideration for both preservation and access. The focus should be based on traditional preservation decisions such as the value of materials; the condition of materials; use of materials; and material characteristics ensuring a high level of success. For the Library of Congress, items of national interest are prime candidates and digitizing these objects improves access while reducing the wear and tear on the originals[13].

In India, libraries are facing many problems initially in digital preservation as shortage of fund provided them, intellectual property right issues, less interest of parent institutes and staff.

References

1. Digital Libraries: Principles and Practice in a Global Environment, Ariadne April 2005.
2. Hedstrom, M. (2007). Digital preservation: a time bomb for Digital Libraries. Retrieved on December 4th, 2007, from http://www.uky.edu/~kiernan/DL/hedstrom.html.
3. OCLC Digital Archive Preservation Policy and Supporting Documentation, p. 5.
4. Cornell University Library. (2005) Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems.
5. Garrett, J., D. Waters, H. Gladney, P. Andre, H. Besser, N. Elkington, H. Gladney, M. Hedstrom, P. Hirtle, K. Hunter, R. Kelly, D. Kresh, M. Lesk, M. Levering, W. Lougee, C. Lynch, C. Mandel, S. Mooney, A. Okerson, J. Neal, S. Rosenblatt, and S. Weibe (1996). "Preserving digital information: Report of the task force on archiving of digital information". Commission on Preservation and Access and the Research Libraries Group
6. Rothenberg, Jeff (1998). Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. Washington, DC, USA: Council on Library and Information

Resources. ISBN 1-887334-63-7.
7. NISO Framework Advisory Group. (2004). A Framework of Guidance for Building Good Digital Collections, 2nd edition, p. 27,
8. http://dspace.iimk.ac.in/bitstream/2259/252/1/05-mgs-ps-paper.pdf
9. LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 57–71
10. Levy, D. M. & Marshall, C. C. (1995). Going digital: a look at assumptions underlying digital libraries," Communications of the ACM, 58, No. 4: 77-84.
11. Flugstad, Myron. (2007). Website Archiving: the Long-Term Preservation of Local Born Digital Resources. Arkansas Libraries v. 64 no. 3 (Fall 2007) p. 5-7.
12. Mange Ram (2005) 'Digital Preservation : A Challenge to Libraries' in Library Progress (International) Vol. 25, No. 1(January-June), 2005. p. 43-48
13. Gertz, J. (2000 April). Selection for preservation in the digital age: An overview. Library Resources & Technical Services , 44(2), 97-104.

About Authors

Mr. Mange Ram, In-charge, Central Library, Dayalbagh Educational Institute (Deemed University), Dayalbagh, Agra - 282005 (U.P.)
E-mail: m_ram72@yahoo.co.in

Dr J K Mishra, Reader, Department of Library & Information Science, Dr Hari Singh Gour University, Sagar (M.P.)
E-mail: jmishr@gmail.com