DIGITAL PRESERVATION IN THE TWENTY FIRST CENTURY: CONCEPT, NEEDS, PROBLEMS AND SOLUTIONS

Samir Chattopadhyay

Abstract

More and more information is being created in digital form. This is done either through converting existing materials to digital form or increasingly born digital. There is no other format but digitise the original. There are increasing expectations in all spheres of life the information we all need will be available on the Internet or at least in an off-line digital format, such as CD-ROM. Digital access has many advantages over paper-based or microform access in terms of convenience and functionality. The present paper deals with the concept of preservation and digital preservation. Difference between digital and paper-based material are discussed. The paper highlights the objective and benefits of digital preservation. Preservation by document type like electronic journals, theses and dissertation, scientific data sets etc. and standard by document type like Text, images, numeric data, and video and audio are briefly discussed. The paper throws light on the problem and solution of digital preservation.

Keywords: Digital Preservation; Digital Objects

1. Introduction

The Preservation of digital information has become a major challenge for society. The availability of electronic information is today taken for granted. With the rapid growth of the Internet and the world wide web, millions of people have grown accustomed to using these tools as resource to acquire information. Digital preservation is rapidly becoming a principal medium to create, distribute and store content, from text to motion pictures to recorded sound. Increasingly, digital information embodies much of the nation's intellectual, social and cultural history.

Traditional preservation, as responsible custody, works best when evidence of human activity can be touched and sensed directly. When the value of the evidence exceeds the cost of the keeping it, and when the creators, keepers, and users of the evidence works cooperatively towards the same end. While this evidence is traditionally seen as documents on paper, increasingly archivists must preserve a variety of formats, including still and moving images and sound recordings. All these media can now reproduce in digital form. Such digital information cannot be used without the help of a machine. Word-wide access through the Internet raises many questions regarding the owning of digital information, who has the right to profit from other's work, and who has responsibility for guaranteeing or regarding access to valuable information. The pace of technological change is at once blinding and stubbornly inadequate. Preservation of digital information is not so much about protecting physical objects as about specifying the creation and maintenance of tangible electronic files whose intellectual integrity is their primary characteristic.

2.   What is Preservation ?

Preservation is the conservation of knowledge. The area of conservation is the creation of text extends to publication, hence to acquisition and then storage in the library for access and use. The term conservation has two parts: preservation and restoration. In other words preservation is the maintenance of objects close to their original conditions as far as possible or until they are no longer needed. Preservation highlights three facts – papers, books and digital technologies. It is too much inseparable a drive to retain recorded knowledge from the basic urge to record knowledge. The evolution of libraries and archives show this drive of mankind to retain information and the development in information technology exhibit a promising means for the prosperity.

3.   Difference between Digital and Paper-based Material

There are some notable differences between digital and paper-based material, which are as follows:

- Machine Dependency: Digital materials all require specific hardware and software in order to access them.

- The speed of changes in technology means that the timeframe during which action must be taken is very much shorter than for paper. Timeframes during which action needs to be taken is measured in a few years, perhaps only two to five, as opposed to decades or even centuries we associate with the preservation of traditional materials. Technology obsolescence is generally regarded as the greatest technical threat to ensure continued access to digital material.

- Fragility of the media. The media digital materials are stored on is inherently unstable and without suitable storage conditions and management can deteriorate very quickly even though it may not appear to be damaged externally.

- The ease with which changes can be made and the need to make some changes in order to manage the material means that there are challenges associated with ensuring the continued integrity, authenticity, and history of digital materials.

- The implications of allocating priorities are much more severe than for paper. A digital resource, which is not selected for active preservation treatment at an early stage, will very likely be lost or unusable in the near future.

- The nature of the technology requires a life-cycle management approach to be taken to its maintenance. A continual programme of active management is needed from the design and creation stage if preservation is to be successful. This in turn leads to much more involvement both within and between institutions and changing roles.

4.   What is Digital Preservation ?

DP refers to the management of digital information over time. Unlike the preservation of paper or microfilm, the preservation of digital information demands ongoing attention. This constant input of effort, time and money to handle rapid technological and organisational advancement is considered as main obstacle for preserving digital information beyond a couple of years. Indeed, while we are still able to read our written heritage from several thousand years ago, the digital information created merely a decade ago is in serious danger of being lost. DP can therefore be seen as the set

of processes and activities that ensure the continued access to information and all kinds of records, scientific and cultural heritage existing in digital formats.

DP is defined as long-term, error-free storage of digital information, with means for retrieval and interpretation, for all the time span that the information is required for. "Retrieval" means obtaining needed digital files from the long-term, error-free digital storage, without possibility of corrupting the continued error-free storage of the digital files. "Interpretation" means that the retrieved digital files that are of texts, charts, images or sounds, are decoded and transformed into usable representations. This is often interpreted as "rendering", i.e. making available to a human to access. However, in many cases it will able to be processed by computational means.

4.1 Objective of Digital Preservation

- **To Protect originals:** General application of digital technologies in an archive or library is digital copies that can be used for ready reference instead of causal browsing through the original sources. Physical access to the original documents are limited, so objective of preservation are met.

- **To Present originals:** To represent the detailed information content of the original sources, a digital system could be built for which that system can be used to fulfill most (if not all) of the research and learning potential of the original documents. High resolution system of this intermediate level of quality open new avenues of research effect on the use could have a transformative effect on the missions of those who create the products.

- **To transcend originals:** Very small but increasing number of applications, digital imaging holds the promise of generating a product that can be used for purposes that are impossible to achieve with the original source. This category also include digital imaging products that incorporate searchable full text (marked up raw). Additionally, digital products that draw together, organize, and enhance access to widely dispersed research materials may have transcendental impact on the people who use them.

4.2 Benefits of Digital Preservation

- It helps to preserve rare and fragile objects without denying access to those who wish to study them.

- Once books are converted to digital, users can retrieve them in seconds by searching for words, combination of words, phrases or ideas. Readers can choose whether to view, or store on a computer or take prints. Several people can simultaneously read the same books or view the same picture.

- Electronic copies occupy millimeters of space rather than meters on shelf and the problems of space vanishes when libraries opt to digital medium.

5. Preservation of Library Document

There are many document types or genres that are important in scientific communication. These include journal articles, books, theses and dissertations, conference proceedings and papers, and project documentation. These document types may be presented as Web sites and they may also qualify as electronic records. These genres may include multiple format types. For example, electronic journals may require supplemental files such as spreadsheets, videos, or software.

5.1   Preservation by Document Type

- **Electronic Journals:** Electronic journals have been at the forefront of preservation discussions because of their critical role in scientific communication and the commercial interests involved. The practices for preserving electronic journals show an increased maturity as evidenced by more formalized procedures such as a DTD for journals.

- **Theses and Dissertations:** Many institutional and national library preservation efforts involve theses and dissertations, since these institutions often have responsibility for providing this genre to their respective national library for incorporation in the national bibliography.

- **Scientific Data Sets:** Data was the earliest digital output of science to be archived. Data sets range from simple numeric data streams of simple structure but large size, to large collections of still and moving images.

- **Technical Reports:** Technical reports and other gray literature are key mechanisms for the dissemination of research and development results especially in industry and government. Many government and institutional archives are focused on technical reports, since libraries may not routinely collect them.

- **Conferences, Meetings and Lectures:** Significant scientific information is first presented at conferences, meetings, lectures, colloquia, etc. Many disciplines, such as biotechnology, rely heavily on this method of communication rather than the formal publications. Therefore, the ability to preserve and access this type of information into the future is important.

- **E-Records:** The extent to which government-produced scientific and technical information is treated as an electronic record depends on the practice of the particular government or institution. Of course, e-records can include any or all of the above document types. However, there are significant e-records efforts underway within the governments that will have an impact on the overall digital preservation landscape.

5.2   Standards by Document Type

The best format for long term preservation remains elusive, perhaps because there is no single answer to the question. Instead, it depends on the format type of the original object, the characteristics of the original that the preserving organization considers to be most important to preserve, and the expected use/re-use of the object in the future (distance education versus legal evidence). Most experts agree that the best format for preservation is that which is least proprietary while conveying significant aspects of the original.

- **Text:** The most common formats for storing text were XML (ASCII, with or without Unicode), PDF and TIFF. Each of these formats has its place in the preservation strategy.

- **Images:** There are a variety of image formats that archives may receive including JPEG and GIF. However, in general the majority of the institutions interviewed who are truly doing preservation convert these formats to TIFF Group IV or V. The rationale is to preserve the best image in a format that is the most standardized and not subject to loss or compression.

- **Numeric Data:** Numeric data is similar to text but it generally has more structure. Whenever possible, the preferred form for data is an ASCII delimited file or an XML tagged file. However, many datasets, particularly those stored in local laboratories and by individual researchers or research groups, are stored in proprietary database formats. Accessing and reusing the data

when it is stored in these formats over the long term, especially if there has been a disruption in the migration from one version of software to another, becomes problematic. However, organizations have also noted that good documentation, particularly retention of the data dictionary (as long as it is not in a proprietary database or CASE product) is necessary as well.

■ **Video and Audio:** Video and audio is used in the sciences to record experiments, supplement human field observations, record engineering and laboratory tests, capture knowledge and lessons learned from researchers, and to teach science and engineering in distance learning environments. However, the standards for preserving video and audio formats remain an issue as the systems grow more complex and proprietary interests dominate.

6. Problems of Digital Preservation

The basic problems of preserving electronic documents or digital objects stems from the nature of the objects themselves. Unlike non-digital formats such as books, magazines, manuscripts, or microfilms, digital objects are accessible only by using combinations of computer hardware and software. Market competitions means that this hardware and software can become obsolete in cycle of less than three years. Ensuring ongoing access, therefore, requires currency with technology changes, and moving digital objects from absolute to current file formats, storage media, operating systems and so on.

6.1 Some other technical, social and legal problems are

■ Rapidly increasing number of digital objects and proliferation of documents standards and formats;

■ Increasing complexity of digital objects (increasing text, images, audio, video in various formats) and their increasing software dependence (e.g., storage in database);

■ Lack of planning to incorporate preservation needs in systems and lack of availability of off-the-shelf products supporting preservation needs;

■ Lack of consideration of long-term access requirement when creating digital products;

■ Lack of awareness for digital presentation;

■ Absence of widely accepted standards which will assure access over time;

■ Copyright / intellectual property rights that may interfere with the ability to preserve digital objects through systematic copying;

■ Unstable storage media (e.g., diskettes) whose life span is limited;

■ Emphasis on the creation and/or acquisition of digital material in an era of diminishing resources, rather than ongoing preservation and access to existing electronic holdings.

■ Lack of technical expertise in collection managers and preservation experts.

6.2 Solution of Digital Preservation

■ A vast amount of the world's information is now produced digitally, and most of this exists in digital form only. The WWW functions as a resource for information and communication as well as a cultural space where a diversity of materials are produced that cannot easily be classified in well-known categories. Much of these digital materials is potentially of lasting cultural value, and new, pro-active strategies need to be developed to ensure it is saved for posterity.

- Preservation of digital heritage is a continuous process that requires commitment and involvement, not only from heritage institutions, but also from governments, policy makers, producers of information, and the software industry. Solutions depend on large-scale cooperation and the creation of a lasting infrastructure.

- A clear division of tasks and responsibilities, based on existing roles and expertise, needs to be established. To come to an infrastructure of distributed archives, it has to be determined that which organization should meet to function as trusted digital repositories. It should be established how tasks can be shared between national heritage institutions and discipline-oriented organizations working (internationally) for specific communities.

- Existing legislation should be adopted to support national heritage institutions in the preservation of digital materials. Deposit legislation should extend to all materials regarded as publications, and legal frameworks for archives should include everything that constitutes a record, in whatever format it is produced. Additional procedures will have to be developed for materials that fall outside these categories (such as research data).

- Awareness of preservation issues should be raised with producers of digital information. They should realize the importance of the use of standards and open source software and of adequate description and documentation. Outreach strategies of heritage institutions are needed to provide guidance and establish strong cooperation with the creators of materials.

- The Information and Communication Technology (ICT) industry should be made aware of the need to take preservation requirements into account. The value of standards and open source software should be promoted among software developers. They should be encouraged to make detailed and complete specifications of their products publicly available, especially for (versions of) programmes no longer in the market. Initiatives should be developed to build sustained repositories of specifications, documentation and related software.

- Copyright legislation should not act as an impediment for preservation of digital heritage. Owners of rights, of content as well as software, should be convinced of the need to allow heritage institutions to take actions necessary for preservation of materials. It should be possible to carry out such actions in the framework of general agreements specifying conditions for access and use.

- The leadership role in digital preservation of a number of heritage institutions worldwide should be acknowledged. Their pioneering work in exploring legal, organizational, technical and economic aspects can provide the basis for defining best practices which should be strongly promoted in the whole community.

- Further research to develop promising models and technology should be widely supported in order to come to fully operational systems for preservation of digital heritage as quickly as possible. As the digital world moves ahead at a rapid pace, the risk that materials will be left behind and irretrievably lost is very serious. With so many resources being invested in the creation of digital materials, it is crucial to stimulate efforts aimed at keeping them accessible also in the future.

- Extensive training programmes are needed to disseminate the expertise and experience gathered so far widely among management and staff of heritage institutions. Preservation of digital heritage requires new organizational structures, new approaches and new ways of thinking. Programmes will have to focus, not only on technical aspects, but also on training staff to deal with a changing environment and new directions.

7.   Conclusion

Preservation of digital information has to deal not only with maintenance of the files themselves but also with ways of keeping them accessible. This means that either the programmes have to be preserved as well and some how kept running on new platforms, or the files have to be converted to another format that can be interpreted by new programmes. As the digital world moves on all the time, this is a continuous process if materials need to be kept accessible for decades (or even forever). In many cases this will result in loss of information functionality and/or appearance, especially with complex multimedia materials that combine a variety of file formats and applications. Preservation of digital information is unknown area to most institutions, when they take on responsibility in this area, they will have to adopt organizational structures and redefine tasks of staff. Cooperation and exchange of experience will be essential to avoid expensive mistakes, and training programmes for staff are priority for all institutions facing digital challenge. Cooperation, guidance, leadership and sharing of tasks are all key elements of programmes for preservation of digital information.

References

1.   Alison, Bullock. Preservation of Digital Information: Issues and current status. Network Notes#60

2.   Brian, Lavoie and Loran, Dempsey. Thirteen ways of looking at… Digital Preservation. D-Lib Magazine, 10 (7/8), 2004.

3.   Commission on Preservation and Access. Website: http://www.clir.org

4.   Desh Pande, J N and Pange, B M. Digitization of library materials. University News 37 (38), 1999, 9-12p.

5.   Gail Hodge, Gail and Frangaki, Evelyn. Digital Preservation and permanent access to scientific information: the state of the practice, CENDI AND ICSTI, 2006.

6.   Hanid, sarah. Constructing a global time capsule: challenges in the digital preservation of society's culture memory archive. Information Technology and Libraries. September 1998, 207-211p.

7.   Headstrom, Margaret. Digital Preservation: a time bomb for digital libraries. Website: http://www.uky.edu

8.   Hussain, Akhtar and Kumar, Krishna. Digital Preservation System: a new approach. In: NICMDil, IIT, Kharagpur, Aug 2-4, 2006.

9.   Lusenet, Yola de. Preservation of digital heritage. Washington D.C: Commission on Preservation and Access, March, 2002

10.   Misra, Joyti, and Satyanarayan, N R. Perspective of digital Preserva In: NICMDil, IIT, Kharagpur, Aug 2-4, 2006.

11.   Mulla, K R and Chandrashekara, M . Preservation of Digital information system: concept and concern. In: NICMDil, IIT, Kharagpur, Aug 2-4, 2006.

12.   Taskforce on archiving of digital information. Website: http://www.rlg.org/Arch TF/tfadi

13.   Website: http://www.gscis.ucla.edu/-howard/copyright/

14.   website: http://www.knaw.nl/ecpa/PUBL/unesco.html

15.   Website: http://www.sunsite.berkeley.edu/longevity/