PRESERVATION METHODS FOR DIGITAL LIBRARY

By

L. RAJENDRAN
Senior Librarian
Tagore Engineering College
Chennai - 600 048
India

M. VENKATESAN
Assistant Librarian
Tagore Engineering College
Chennai - 600 048
India

Dr. S. KANTHIMATHI
Librarian (SG)
Rani Anna Govt. College for Women
Tirunelveli
India

ABSTRACT

Going digital is the way to minimize handling of damaged materials, but the imaging process is demanding and must be done with oversight by preservation staff and with a high enough level of quality to ensure the reusability of the archival electronic file for as long as possible. This paper focuses on the scope and needs of digital preservation, and various types of available preservation methods. The remainder of this paper explores some of the approaches and technological issues facing our profession.

I. INTRODUCTION

Information and communication technology has dramatically altered the process of teaching and scholarly research. The paradigmatic shifts rutting from the introduction of new and evolving technologies will almost certainly continue well into the 21st century. Digital information and resources through scholarships are in so many different ways that often we struggle to clearly identify the impact and articulate the implications. Libraries as information service providers have come to rely increasingly on digital information both as supplements to and parallels of print materials. Libraries are also encountering new resources, which are "born digital" and have no print or analogue equivalent and they exist only in digital form. The relative ease with which digital resources can be created has also meant, "our ability to create, amass and store digital materials" that far exceeds our current capacity to preserve even small levels with continuing value".

II. WHAT IS DIGITAL PRESERVATION?

Digital preservation is concerned with ensuring that records which are created electronically using today's computer systems and applications, will remain available, usable, and authentic in ten to one hundred years time, when the applications and systems which were used to create and interpret the records will, more likely than not, no longer be available. Digital preservation consists of preserving more than just the record's bit stream. We must also be able to interpret the boot stream in order for the record to survive. Without interpretation, the bit stream is nothing more than a meaningless series of 0's and 1's. During preservation, questions of record, context, content, structure, appearance and behaviour must also be taken into account. Appearance and behaviour are aspects that are peculiar to digital records. These may therefore require the most attention to authentically preserve the records over a long term.

There are wide ranges of digital formats available and to make matters more complicated, different digital objects have different preservation requirements. These can depend on the reason for which the record is being preserved, how long it needs to be preserved, the context and history of the record, and its original format. Digital preservation does not mean the same thing for each digital object. Whilst it is often considered that digital

preservation means preserving the object so that it is identical to its original format, is not always required. It is not always necessary to preserve every aspect of a digital record, and thus research is underway to define the essential aspects of records and their authenticity requirements. In all cases, however, the record must be preserved so that it retains it's integrity and is authentic and usable.

III. Types of Preservation

When considering digital materials, there are three types of "preservation" one can refer to:

- **The preservation of the storage medium:** Tapes, hard drives, and floppy discs have a very short life span when considered in terms of obsolescence. The data on them can be refreshed; keeping the bits valid, but refreshing is only effective as long as the media are still current. The media used to store digital materials become obsolete in anywhere from two to five years before they are replaced by better technology. Over the long term, materials stored on older media could be lost because they will no longer have the hardware or software to read them. Thus, libraries will have to keep on moving digital information from one storage medium to the other.
- **The preservation of access to content:** This form of preservation involves preserving access to the content of documents, regardless of their format. While files can be moved from one physical storage medium to another, what happens to the formats (e.g., Adobe Acrobat PDF) containing the information? This is a problem perhaps bigger than that of obsolete storage technologies. One solution is to do data migration--that is, translate data from one format to another preserving the ability of users to retrieve and display the information content. However, there are difficulties here too-data migration is costly, there are as yet no standards for data migration, and distortion or information loss is inevitably introduced every time, data is migrated from one format to the other.
- **The preservation of fixed-media materials through digital technology:** This slant on the issue involves the use of digital technology as a replacement for current preservation media, such as microforms. Again, there are, as yet, no common standards for the use of digital media as a preservation medium and it is unclear whether digital media are as yet up to the task of long-term preservation. Digital preservation standards will be required to consistently store and share materials preserved digitally (Chepesuik, 1997).

IV. Benefits

The natural forces of deterioration threaten virtually every medium used to record information. The magnitude of preservation needs in a library is determined by the interplay of many factors; chief among them are the age, scope, and composition of the collections. Research library collections include a multiplicity of formats (e.g., monographs, journals, newspapers, maps, manuscripts, photographs, digital images) and media (e.g., paper, vellum, photographs, films, magnetic tapes, various types of disks). Among these diverse resources there is tremendous variation in life expectancy. Paper made from cotton fibre has lasted for more than a thousand years, preservation microfilm can have a life expectancy of hundreds of years, wood pulp newspaper pages deteriorate within decades, and some types of computer disks show loss of information after a few years.

The important benefits of digital library are listed below

- **Cost Savings:** Create digital images and microfilms more economically and efficiently than possible with in-house operations.
- **Longevity:** Microfilm your collections with OCLC Preservation Service Centers for proven longevity.

- **Searchable repositories:** Enhance access to your newspaper collections by creating searchable repositories.
- **Protection for your investment:** Store your valuable microfilm in our secure, climate-controlled vault.
- **Loss prevention:** Duplicate deteriorating acetate film onto today's stable polyester film.
- **Access:** Digitize your collections to provide greater access for your patrons and researchers worldwide.
- **Improved Search ability:** Use metadata to make your digital collections more searchable.
- **Quality:** Rely on the centers of technical expertise and rigorous QA procedures.

V. Various Approaches for digital preservation

- **Science Data**

A number of initiatives to preserve digital materials have been ongoing for some time. In scientific and scholarly research, computerized data have been created and used for decades. The space and earth observation communities, using massive amounts of data that need to be studied over a long period of time, have been very active in developing a reference model for archiving data that is being widely adapted. Data archives, especially in the social sciences and the humanities, have for years been collecting data sets created in research projects so that they are maintained and can be re-used.

- **Library Initiatives**

National libraries generally approach the digital environment from the angle of deposit legislation. Deposit of offline digital products, such as CD-ROMs, in several countries is already a legal requirement. Online electronic journals are treated as an extension of a long tradition of print publishing, which libraries have always collected and preserved. To ensure continued access to the whole of the scientific electronic journal environment, including live links, data and multimedia presentations, libraries are now trying to come to arrangements with publishers about deposit, as yet often on a voluntary basis. Several libraries have developed strategies for actively selecting and preserving websites on the basis of a concept of 'publication'. 'Publication' is defined in broad terms: anything on the Internet is regarded as a publication, only organizational records are explicitly excluded.

- **Initiatives by archival institutions**

Some national archives, as for instance the Public Record Office, have extended policies for electronic record management to include websites of government agencies (public sites as well as intranet sites) and developed guidelines describing best practices. The Public Record Office warns that materials on websites are not always recognized as records. They are often 'very different in nature from the traditional image of a "record". So much so that it can tend to give the impression that no records are present. This can be highly misleading'. For, on the contrary, 'rigorous records management' is required for websites. Responsibilities and procedures for identifying records and managing them remain valid in the Internet world.

- **Harvesting the web**

Apart from these selective approaches for preserving web content, there are also examples of comprehensive approaches, which collect enormous numbers of WebPages without any selection for content. The Internet Archive, started in 1996 as a private, nonprofit enterprise, 'is working to prevent the Internet a new medium with major historical significance and other "born-digital" materials from disappearing into the past'. It collects freely available WebPages worldwide and now comprises over 10 billion WebPages or 100 terabytes of data (5 times

the size of all the materials held by the Library of Congress). The Internet Archive launched a 'Way back Machine' in October 2001 to provide free access to the archive over the web. At the moment, the main aim of these initiatives is to save web materials that would otherwise in any case have been lost forever. They give us 'both the record of the origins and evolution of the Internet, as well as snapshots of our society as a whole around the turn of the century'. However, rendition of captured sites is as yet incomplete, for capturing online information, and is extremely complex. Links to external sites will in many cases be broken and interactive navigation cannot be always retained. More and more WebPages are dynamic, generated 'on the fly' by databases hidden behind the static front end of the site. It is estimated that the databases behind websites, together called the 'deep web', contain many more times the amount of information accessible on the surface. The information in those databases cannot be captured by copying the website, as it is not available in ready-made pages at the surface. Moreover, capturing web content is only the first step in the preservation process. After five years of archiving, there is no saying yet how it can be ensured that these materials will still be available after 25 or 50 years. In spite of many uncertainties, the initiatives taken by memory institutions are valuable explorations of the legal, organizational, economic and technical frameworks required for preservation of on- and offline materials. The experience gained by the pioneers in this area will be of huge benefit to the whole cultural sector and will contribute considerably to the development of infrastructure and policies for preservation.

VI. Technological Issues

Most digital materials cannot meaningfully exist outside the digital environment as they rely on software for their interpretation and functionality. Printing the information out on paper to preserve it would only work for a small category of straight text files. Generally, in order to use the material at some future moment, as it is meant to be used, both content and functionality need to be preserved. In many cases, the 'look and feel' of the material is an aspect that cannot be ignored either. Preservation of digital materials is therefore a complex technological task that has to deal with several aspects simultaneously.

VII. How Materials become inaccessible

Basically, there are three ways in which digital materials can become inaccessible:

1. Degradation of the media on which they are stored
2. Obsolescence of software making it impossible to read digital files
3. Introduction of new computer systems and peripherals that cannot handle older materials.

Tapes and disks are all subject to physical decay and none of them have a lifespan that is comparable to that of preservation-standard microfilm or acid-free paper. They need to be stored under controlled conditions, but even so materials should be copied onto new media at regular intervals to prevent loss through deterioration of carriers. 'Refreshment' of materials, i.e. transferring those to new media, often becomes necessary because a specific type of disk or tape can no longer be used in current computer systems. The disappearance of the 5 1/4 disk and the accompanying disk drives is a case in point. Refreshment is a recurring activity in any preservation programme. In fact, the media on which information is stored are transitory carriers that serve their function only for a limited period of time, and preservation has to take account of other aspects as well. Obsolescence of software and hardware leads to (partial) loss of information or functionality of files in their original format. Successive versions of programmes may be compatible, but software producers do not usually support compatibility over a long

period. Programmes also disappear from the market or can no longer be used on a new platform. The combination of dependence on older versions of programmes that used to run on older platforms of outdated computer systems inevitably spells digital death.

VIII. Technological approaches

For the short term, it may be possible to keep the original environment (hardware and software) functioning. There is, however, wide agreement that this will not work in the long run, as it will result in an ever-growing collection of outdated computers and peripherals that is very hard to maintain over time. Such computer museums may still have a role for exceptional cases. Different approaches have been suggested to combat obsolescence of software and hardware. One method is to convert files to new platforms or different programmes. This is especially attractive if they can be converted to a standard, nonproprietary format, as this facilitates maintenance over time. However, conversions may lead to unacceptable loss of functionality, especially with complex databases or multimedia materials. Even with relatively simple materials it is hard to predict the cumulative effect of successive conversions over time. Other approaches aim to recreate superseded versions of operating systems and programmes in new environments, so that the files can be kept in their original format and read with the software in which they were first created. This is certainly a way to bridge one or two generations of platforms, but over time, as new systems keep on being introduced, one may be faced with a Chinese boxes effect that becomes complex to handle. Another disadvantage may be that functionality is kept at the level of outdated systems, which may not be very satisfactory for future users.

IX. Standards and Documentation

These approaches are not mutually exclusive but should be combined in an institutional preservation policy. It is as yet uncertain what will prove to be feasible and successful, and many institutions are doing research, creating test beds and pilots to gain more experience with potential solutions. In the meantime, a better appreciation of the risks and complexities among producers of digital materials could make all the difference for institutions engaged in developing preservation systems. Producers can facilitate preservation efforts by using standards. Emerging standards like XML and TIFF are promising because they are open standards not dependent on a specific platform; others, like PDF, are so widely used that this offers some hope that they will be supported over a long time. The use of proprietary software complicates matters not only because it is protected, but also because it is often inadequately documented. Even when programmes are taken off the market, source codes are not usually brought into the public domain. Adaptations made during the life of the software are not always documented, so that one cannot predict the outcome of a conversion in every detail.

X. Conclusion

Libraries around the world have been working on this daunting set of challenges for several years now. They have created many digital library initiatives and projects, and have formed various national schemes to develop excellent preservation methods. Librarians have discovered that, with a few exceptions, making a business case for digitization and investments in digital technology is more difficult than first envisioned, especially given the technical and legal constraints that must be first overcome. As with most other technical developments in libraries over the years, we will have to move forward in small, manageable, evolutionary steps, rather than in a rapid revolutionary manner.

References

Chepesuik, R. (1997). *The future is here: America's libraries go digital. American Libraries, 2(1), 47-49.*

Waters, D.J. (1998). *What are digital libraries? CLIR Issues, July/August. URL: http://www.clir.org/pubs/issues/issues04.*

Graham, P.S. (1995b). *Long-term intellectual preservation. URL: http://aultnis.rutgers.edu/texts/dps.html*

Hendley, Tony. *Comparison of Methods and Costs of Digital Preservation. Joint Information Systems Committee: 1998. Http://palimpsest.stanford.edu/byorg/nara/nistsum.html*

Hedstrom, M and Montgomery, S. *Digital Preservation Needs and Requirements in RLG Member Institutions: a study commissioned by the Research Libraries Group. RLG: 1998.http://www.as400.ibm.com/as400/three.html*