THEME PAPER

DIGITAL PRESERVATION AND MANAGEMENT: AN OVERVIEW

Jagdish Arora

Abstract

Digital preservation is not a new concept for libraries. The libraries have been migrating and refreshing their OPAC records as well as their databases developed in-house ever since automation in libraries started in mid 1980s. With availability of products and services in digital forms, libraries are committing larger portions of their budgetary allocation for either procuring or accessing digital contents. Preservation and archiving of digital contents has become a serious concern of libraries for collection either acquired through subscription, purchased in digital media or converted in-house. The article deliberates upon need, relevance and major challenges of digital preservation. It enumerates on dimensions and manifestation of digital preservation and describes traditional preservation tenets as applicable to the digital preservation. The article describes various digital preservation strategies with a caution that appropriate strategies may be adopted depending upon data types, situations, or institutions. The article touches upon digital preservation metadata as a subset of metadata that describes attributes of digital resources essential for its long-term accessibility and describes OAIS Reference Model as well as other major preservation metadata initiatives taken up by the OCLC and ARL. Considering the fact that short life of storage media, is one of the major crucial threat to digital preservation, the article briefly describes storage management as applicable to digital preservation repositories. Lastly, the article touches upon microfilming and digitization as hybrid solution for reliable preservation.

Keywords : Digital Library; Digital Preservation; Digital Contents; IPR; OAI

1. Introduction

The librarians have been concerned about the digital preservation ever since the first computer was introduced and its products and services found its way into the libraries. The libraries have been migrating and refreshing their OPAC records ever since automation in libraries started. Since mid 1980s, the libraries in India also started building their in-house databases and began subscribing electronic resources such as Current Contents on Disc (CCOD) as well as other computer-based services that were delivered on 5¼ inch floppy discs. Several books in 1980s and 1990s had accompanied floppy discs. 5¼ inch floppies are already obsolete and floppy drives that were used for reading them have completely disappeared. CD ROMs, once respected for its longitivity, are known to dysfunction much faster than expected. Moreover, in time to come, the CD ROM may completely be phased out in favour of its more evolved avatar, i.e. DVD ROM with greater storage capacity. Institutions such as national archives, data archives, and other cultural institutions with

preservation as one of their main mandate, have established digital preservation programmes as early as the late 1960s. These programmes addressed the issues of preservation of technology and digital contents that existed at that time (paper tapes, punch cards, etc).

Libraries acquire digital materials through different channels that include buying digital contents from publishers or aggregators, and licensing access to online databases and journals. Moreover, libraries and institutions around the world are taking projects to convert their analogue collections into digital form with an aim to increase their access thus far confined to the four-walls of their libraries, many a times, without ensuring their long-term accessibility. The crucial issue of moving a digitization pilot to a fully operational system with elements of preservation and sustainability built-in has not been given serious consideration that it deserve. Undoubtedly, digital resources have several advantages over its print counter part, however, preservation is definitively not one of them. The fact that the risk of loss of data in digital form is much greater than any other physical form has not been addressed to full extent. Digital preservation addresses the issue of adapting concepts of preservation to manage risk in the midst of rapid technological advancements.

With rapid developments in imaging, storage and communication technology, it is believed that while the quality of image produced would improve manifold, the cost and space requirement for image would reduce drastically with proportionate increase in capacity of networks to transmit high-quality images. However, the issues of digital preservation that are relevant now would remain relevant.

Digital documents are vulnerable to loss because of decay and obsolescence of the media on which they are stored, and they become inaccessible and unreadable when the software needed to interpret them, or the hardware on which that software runs, becomes obsolete and is lost. Preserving digital documents may require substantial new investments, since the scope of this problem extends beyond the traditional library domain, affecting digital records such as government records, environmental and scientific data, data on nucleic acid sequences, human genome, documentation of toxic waste disposal, medical records, corporate data, and electronic-commerce transactions. The digital preservation involves a variety of issues and challenges including policy issues, institutional commitments, legal and IPR issues and metadata. The article examines these issues with a librarians perspective.

2. Definition

The term "digital preservation" refers to preservation of materials that are created originally in digital form and never existed in print or analogue form (also called "born-digital") as well as those converted from legacy documents and artefacts (printed documents, pictures, photographs or physical objects) into images using scanners, digital cameras, or other imaging technologies for access and preservation purposes.

Digital preservation refers to a series of managed activities designed to ensure continuing access to all kinds of records in digital formats for as long as necessary and to protect them from media failure, physical loss and obsolescence (Cornell University Library, 2005). The Wikipedia (Wikipedia, 2006) defines "digital preservation 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",

where "retrieval" means obtaining required digital files from the long-term, error-free digital storage, without corrupting the error-free stored digital files and "interpretation" means that the retrieved digital files, which may be texts, charts, images or sounds, are decoded and transformed into usable representations for access to human.

Digital Preservation Coalition (2006) defines digital preservation as "all activities that are required to maintain access to digital materials beyond the limits of media failure or technological change. Those materials may be digital records created during the day-to-day business of an organisation, i.e. "born-digital" materials created for a specific purpose (e.g. teaching resources), or the products of digitisation projects".

Digital objects are the basic unit of both access and digital preservation and one that contains all of the relevant pieces of information required to reproduce the document including metadata, bit-streams, and special scripts that govern dynamic behaviour. This data is encapsulated in the digital object and should be managed as a whole (Jantz and Giarlo, 2005).

3. Why Digital Preservation ?

The traditional libraries are increasingly getting transformed into digital libraries, atleast partially. The availabilility of web-based digital information products are exerting ever-increasing pressure on the traditional libraries, which, in turn, are committing larger portions of their budgetary allocation for either procuring or accessing web-based online or full-text search services, CD ROM products, online databases, multi-media products, etc. The availability of digital information products and services, in turn, has trigerred a major shifts in the traditional practices and policies from buying and storing information services to accessing them. Besides, acquiring and buying access to digital collections, libraries are exerting efforts on initiating digital library projects in their respective institutions to build their own digital collections (Arora, 2002). The libraries are increasingly converting their existing print collections into digital formats or are increasingly capturing collections that are "born digital". Preservation and archiving of digital contents is one of the most serious concerns of libraries, whether acquired through subscription, purchased in digital media or converted in-house. Moreover, the academic community looks upon libraries to preserve materials that was ever accessible to them on Internet at least in an offline digital format, such as CD-ROM. While access to digital collection has definite advantages over its paper-based or microform-based counter-part in terms of convenience of usage, accessibility and functionality, however, long-term preservation of digital information is plagued by short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites, said Chen (Chen, 2001). The exponential growth in digital information and its ephemeral nature, as well as considerable challenges associated with ensuring its continued access, necessitate that concerted efforts be made to overcome these challenges. There are enough evidences to suggest that many potentially valuable digital materials have already been lost and it incurs substantial costs to recover these digital contents as observed in the following examples:

- The Census Bureau saved the 1960 Census on Univac paper tapes that could be read only with a UNIVAC type II-A tape drive. By the mid-seventies, these paper tape drives were obsolete. When it was decided to archive the information on computer tapes containing the raw data from the 1960 federal census, there were only two machines in the world capable of reading those tapes: one in Japan and the other already deposited in the Smithsonian as a relic.

- NASA/NSF/NOAA rescued valuable 20-year-long TOVS/AVHRR satellite data documenting global warming.
- In the late 1960s, the New York State Department of Commerce and Cornell University undertook the Land Use and Natural Resources Inventory Project (LUNR). The LUNR project produced a computerized map of New York State depicting patterns of land usage and identifying natural resources. It created a primitive geographic information system by superimposing a matrix over aerial photographs of the entire state and coding each cell according to its predominant features. The data were used for several comprehensive studies of land use patterns that informed urban planning, economic development, and environmental policy. In the mid-1980s, the New York State Archives obtained copies of the tapes containing the data from the LUNR inventory along with the original aerial photographs and several thousand transparencies. Staff at the State Archives attempted to preserve the LUNR tapes, but the problems proved insurmountable. The LUNR project had depended on customized software programs to represent and analyze the data, and these programs were not saved with the data. Even if the software had been retained, the hardware and operating system needed to run the software were no longer available.

4. Challenges for Preserving Digital Contents

Although, the digital technology offers several advantages over their print counter part, it along with other associated Internet and web technologies are in a continuous flux of change. New standards and protocols are being defined on a regular basis for file formats, compression techniques, hardware components, network interfaces, storage media and devices, etc. The digital contents face the constant threat of "techno-obsolescence" and transitory standards. Magnetic and optical discs as a physical media are being re-engineered on continuingly to store more and more data. There is a constant threat of backward compatibility for products, including software, hardware and associated standards and protocols that were used in the past. The challenges in maintaining access to digital resources over time are related to notable differences between digital and paper-based material. Some of the important challenges for preserving digital contents are as follows:

4.1 Dynamic Nature of Digital Contents

The initial problem with digital preservation is the content itself (Chen, 2001). Preservation in analogue world involves static objects like printed documents, manuscripts and other artefacts, collecting and storing these items in some form is simple and straightforward process. Preserving digital contents requires reconsideration in terms of meaning and purpose of preservation. Digital information exists in several forms and type. There are several digital documents that are true replica of their print counterpart, such as books, reports, correspondences, etc. However, there are other types of digital material that varies greatly from their tradition forms. There are yet another types of digital material, which cannot be replicated in traditional hard-copy or analogue media, for example, interactive Web pages, geographic information systems, and virtual reality models. For example, web sites have links that not only change but point to dynamically changing sites. As the object grows and changes over time, new questions emerge about what it means to preserve a digital object. Internet users are all familiar with the link failure syndrome that plagues the Web. Spinellis (2002) indicates that approximately 28% of the URLs referenced in Computer and Communications of the ACM articles between 1995 and 1999 were no longer accessible in 2000 and the figure increased to 41% in 2002.

4.2 Machine Dependency

Digital contents are machine-dependent. It may not be possible to access the information unless there is appropriate hardware, and associated software, which will make it intelligible. Access to digital contents may require specific hardware and software that were used for creating them. Since computer and storage technologies are in a continuous flux of change, the timeframe available for migrating digital contents to new software / hardware is generally very short, typically 3 to 5 years, as opposed to decades or even centuries that may be available for preserving traditional materials. Techno-obsolescence is considered as the greatest technical threat to ensuring continued access to digital contents. Digital contents stored on 5¼ inch floppy disk, for example, can not be accessed since it has been superseded by 3½ inch floppy disks along with drives to access data from it.

4.3 Fragility of the Media

The storage media used for storing digital contents are inherently unstable and highly fragile because of problems inherent to magnetic and optical media that deteriorate rapidly and can fail suddenly because of exposure to heat, humidity, airborne contaminants, or faulty reading and writing devices (Hedstrom and Montgomery, 1998). Magnetic storage media is highly sensitive to dust, heat, humidity and other climatic conditions. Most storage devices, without suitable storage conditions and proper management, may deteriorate very quickly without displaying any physical characteristics of external damage. Deterioration of storage media may lead to corrupted digital files in such a fashion that it may not be easy to identify the corrupted portion of digital contents. Moreover, unless digital contents receive preservation treatment at an early stage, it is likely that it would be rendered unusable in near future.

Besides unintentional corruptions, digital contents are amenable to intentional corruption and abuse. The ease with which digital contents can be altered and amended, necessitates that digital preservation also addresses the issues associated with ensuring the continued integrity, authenticity and history of digital contents.

4.4 Technological Obsolescence

Unlike the situation that applies to books, digital archiving requires relatively frequent investments to overcome rapid obsolescence introduced by galloping technological change (Feeney, 1999) Technological obsolescence can affect hardware (including storage media and devices to read them), software and file format. Not only computers are continuingly superseded with their faster and more powerful versions, the media used to store digital contents also become obsolete in two to three years before they are replaced by newer and denser versions of that medium, or by new types of media that is smaller, denser, faster, and easier to read. The digital materials stored on older media could be lost because the hardware or software to read them may become obsolete. Although the media may physically survive for years, the technology to read and interpret it may exist for only a brief period of time. As a result, even if the storage media is retained in the best condition, it may still not be possible to access the information it contains.

Obsolescence also affects software that is used to create, manage, or access digital contents since the software are being superseded by newer versions or newer generations with more capabilities. There is a constant threat of backward compatibility for digital contents that were created using older versions of software. Similarly, file formats are being superseded with newer versions, and the newer versions of software may not read files in older formats. Although some file formats are largely independent of specific software (for example ASCII and Unicode), most are tied to individual or related groups of software. Proprietary software with associated file formats represents some of the most enduring and successful software in use. Commercial software developers regularly release new versions of their software and associated file formats with added features and functionality in order to entice users to upgrade.

It may be noted that digital contents created on Word Star can no longer be accessed unless the software is still available. Likewise, thousands of software programs common in the early 1990s are now extinct and unavailable. Given the fact that technological changes are inevitable, it is considered as one of the greatest threat to successful digital preservation.

4.5 Shorter Life Span of Digital Media

The greatest concern of digital preservation is relatively short life span of digital media and higher rate of obsolescence of the hardware and software used for accessing the digital records. Rapid change in the IT industry and the move from science-based development to commercial development of software and hardware systems, has resulted into media becoming inaccessible at a faster pace. Magnetic tapes, disks and optical storage disks (e.g. CDs and DVDs) are manufactured for short-term storage of digital objects, and, therefore, cannot be used for long-term archival retention.

4.6 Formats and Styles

Information contents that were earlier confined to traditional formats like books, maps, photographs, and sound recordings are getting increasingly available in diversity of digital formats. New formats have emerged, such as hypertext, multimedia, dynamic pages, geographic information systems and interactive video. Each format or style poses distinct challenges relating to its encoding and compression for digital preservation.

4.7 Copyright and Intellectual Property Rights (IPR) Issues

Copyright and intellectual property rights (IPR) have a substantial impact on digital preservation. The IPR issues for digital contents are much more complex than for printed material. IPR issues in digital environment have implications not only on digital contents but also to any associated software. Long-term preservation and access may require migration of digital material into new forms or emulation of the original operating environment which may not be possible without appropriate legal permissions from the original rights owners of the content and underlying software. Moreover, simply refreshing digital materials onto another medium, encapsulating content and software for emulation, or migrating content to new hardware and software, may lead to infringement of IPR unless statutory exemptions exist or specific permissions have been obtained from the rights holders. Furthermore, since migration and emulation may involve manipulation and changing presentation and functionality to some extent, it is important that these issues are addressed to with the copyright holder of the contents during negotiations ensuring preservation of selected items.

Some of the additional complexity in IPR issues relates to the fact that digital materials can be copied and distributed easily. Rights holders are, therefore, concerned with controlling access and potential infringements of copyright. Technology developed to address these concerns can also inhibit or prevent actions needed for preservation. These concerns over access and infringement and preservation need to be understood by organisations preserving digital materials and addressed by both parties in negotiating rights and procedures for preservation.

5. Dimensions of Digital Preservation

Digital preservation activities can broadly be divided into two components, i.e. i) activities that promote the long-term maintenance of digital image; and ii) activities that provide continued accessibility of its contents. Standards and formats, software technology, upgrade path, staff and technological resources necessary to manage the digital objects would depend on the life-span of a digital object as mentioned below, determines preservation strategies:

- Long-term preservation: Continued access to digital materials, or at least to the information contained in them, indefinitely.
- Medium-term preservation: Continued access to digital materials beyond changes in technology for a defined period of time but not indefinitely.
- Short-term preservation: Access to digital materials either for a defined period of time while use is predicted but which does not extend beyond the foreseeable future and/or until it becomes inaccessible because of changes in technology.

The concept of digital preservation has come to have at least three distinct meanings:

5.1 Make Use Possible

For a very small subset of valuable but deteriorated documents, digital imaging technology is a viable, and possibly the only, cost-effective mechanism for facilitating its use for researchers. A recent experiment involving digitizing oversize colour maps (Gertz, 1995) demonstrated that the only way to really use the maps, which have faded badly and are very brittle, is to view them on a large colour monitor after they have been digitized and enhanced. Similarly, the managers of the Andrew Wyeth estate have found that reproductions of the artist's work are most faithfully represented in digital form (Mintzer and McFall, 1991).

5.2 Protect Original Items

Digital image technology can be used to create a high-quality copy of an original item. By limiting direct physical access to valuable documents, digital imaging becomes a "preservation application" as distinct from an "access application". The original order of the collection, or book, is "frozen", much like microfilm sets images in a linear sequence. Sophisticated indexing schemes facilitate browsing and minimize the potential for damage or disruption to a collection caused by "fishing expeditions" through the published or unpublished record. Preservation via digital copying has been the most compelling force motivating archives and libraries to experiment with hardware and software capabilities.

5.3 Maintain Digital Objects

Once digital conversion of the original document has been completed, the challenge of protecting the digit contents from corruption or destruction becomes the preservation focus. This facet, called "digital preservation", typically centers on the choice of interim storage media, the life expectancy of a digital imaging system, and the concern for migrating the digital files to future systems as a way of ensuring future access (Preserving Digital Information, 1995).

5.4 Ease Access to Digital Objects

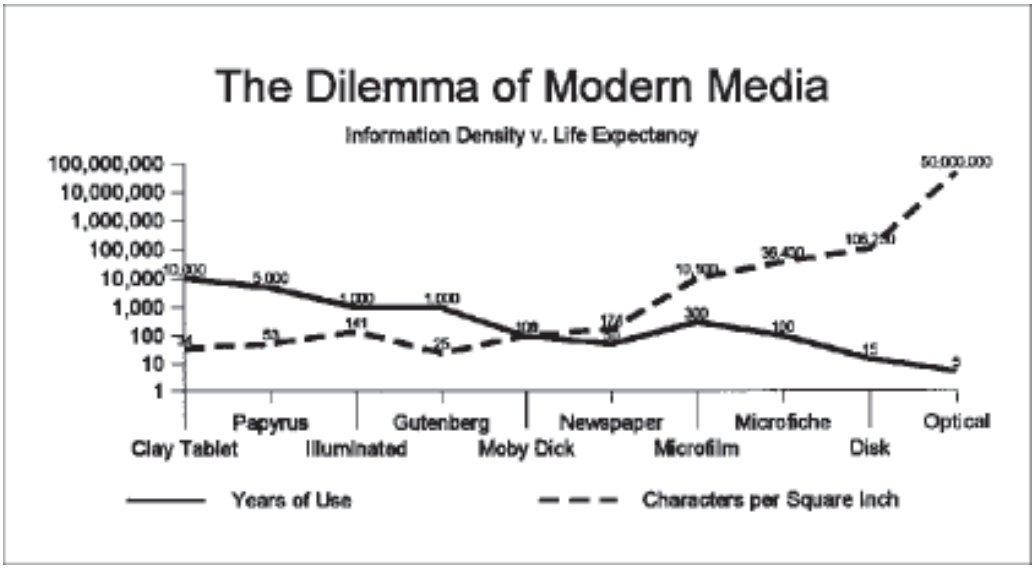
Digital technologies present a preservation solution for the documents in the libraries with increased access to them over the data networks. Digital preservation activities, therefore, are not confined to the simply act of preservation of contents. The goal of digital preservation is to extend ease of access over the data networks. The access to digital contents and its structure and integrity, occupies the central stage in preservation process and the ability of a machine to transport and display this information object becomes an assumed end result of preservation action rather than its primary goal. Digital preservation activities, therefore, include creation of descriptive metadata of digital contents being preserved so as to facilitate ease of access.

6. Principles of Preservation as Applied to Digital Preservation

The basic principles of preservation that are being practiced for preservation of analogue media are also applicable to preservation in the digital world. In essence, digital preservation defines priorities for extending the life of digital information resources. Convey (Convey, 1996) identified five principles, i.e. longevity, choice, quality, integrity, and accessibility that are being practiced for preservation of analogue media and can be extended to digital preservation.

6.1 Longitivity

Density of media to record information has increased exponentially over time while its longevity to store the information has decreased proportionately. The graph given below (Convey, 1996) plots ten "writing" media on "X" axis in chronological order with their corresponding capacity to write information on "Y" axis on a logarithmic scale. It can be observed that the capacity to write information increases at each level by a factor of ten.



The solid line in the graph represents the life-expectancy in years of each recording medium which is declining through the years. Papyrus fragments of Egyptian writing from 4,500 years ago, while quite fragile today, are still legible. Similarly, manuscripts and other documents from Medieval times are quite able to withstand centuries of climatic conditions. A similar situation prevails with early modern book printing technologies. Books published on acidic paper in 1850s do present a challenge to preservation but are still readable. During the twentieth century, the permanence, durability, and stamina of newer recording media have continued to decline, with the exception of microfilm (Sebera 1990). Magnetic tape may be unreadable just thirty years after manufacture (Van Bogart 1995, p. 11). The newest recording medium—optical disk—may indeed have a longer life than the digital recording surfaces that have gone before. It is likely, however, that today's optical storage media may long outlast the life of the computer system that created the information in the first place. In order to achieve the kind of information density that is common today, we must depend on machines that rapidly reach obsolescence to create information and then make it readable and intelligible (Dollar, 1992).

The longitivity of digital contents dependents on the life expectancy of the access system, including hardware and software. Digital storage media should be handled with care, however, storage media is likely to have longer life span in comparison to computer systems that is used to retrieve and interpret the data stored on them. The libraries must always be prepared to migrate valuable digital contents, indexes, and software to future generations of the computer and storage devices. Migration of digital contents would remain a continuing activity to ensuring perpetual availability of digital information. The libraries must ensure continuing institutional commitment to support long-term migration strategies.

6.2 Selection

Selection of digital material for preservation is an ongoing process intimately connected to the active use of the digital files. The process of selection and value judgment is involved every time a decision is to be made to convert documents from paper or digital image and migrate it from one storage media and access system to another so as to continue preserving the information. Rare collection of digital files can only justify the cost of a comprehensive migration strategy. (Conway, 1996).

Selection of digital contents for preservation should reflect the broader institutional mission. Moreover, as with analogue documents, the main criteria in the selection of digital contents for preservation should be their authenticity, significance and lasting cultural value in reflecting subject matter, form and style, the identity and diversity of a given people, place or period of time.

6.3 Quality

Quality in the digital world is concerned with usefulness and usability of digital contents, and is essentially govern by the limitations of capture and display technology. Imaging technology, for example, facilitates scanning at resolution of 1500 dpi, however, the printing and display technology has its limitation, since it can only faithfully display images at maximum of 600 dpi. Image scan at higher resolution may occupy much more disc space but it does not make any qualitative difference on the output resolution. Moreover, digital conversion places more emphasis on getting the best representation of the original in digital form rather than obtaining a faithful reproduction of the original. The primary goal of preservation quality is to capture as much intellectual and visual contents as is technically possible and then display that content to users in ways most appropriate to their needs.

Quality of the digital object, including the richness of both the image and the associated indexes, is the heart and soul of preservation in the digital world. This means maximizing the amount of data captured in the digital scanning process, documenting image enhancement techniques, and specifying file compression routines that do not result in the loss of data during telecommunication. (Convey, 1996)

6.4 Integrity

Digital preservation is concerned with physical as well as intellectual integrity of digital contents. In terms of digital preservation, the physical integrity of a digital image file is determined in terms of loss of information that occurs when a file is created in the process of scanning, and compressed mathematically for storage or transmission across the networks. The metadata (descriptive or structural) that describes intellectual contents of an image file or its organization is an integral part of the digital file, which must be preserved along with the digital image files themselves. The preservation of intellectual integrity also involves authentication procedures to make sure files are not altered intentionally or accidentally (Lynch, 1994).

Librarians can exercise control over the integrity of digital image files by authenticating access procedures and documenting successive modifications to a given digital file. They can also create

and maintain structural indexes and bibliographic linkages within well-developed and well-understood database standards. Librarians are acknowledged as experts in organizing information and, therefore, have a vital role to play in influencing the development of metadata interchange standards, including the tools and techniques that will allow structured, documented, and standardized information about data files and databases to be shared across platforms, systems, and international boundaries.

6.5 Access

Digital technologies present a preservation solution for the documents in the libraries with increased access to them over the data networks. The access to digital contents, therefore, occupies the central stage in preservation process in digital world. Preservation in the digital world is not simply the act of preserving access but also includes a descriptive metadata of digital contents being preserved. Acquisition of non-proprietary hardware and software components can ensure perpetual access to digital image files. The librarians and archivist should encourage vendors for adoption open system architectures and non-proprietary hardware. Vendors and manufacturers should also be convinced to develop new systems that are "backwardly compatible" to ensure continuing accessibility of digital contents. This capability assists image file system migration created with earlier versions to the present version.

7. Digital Preservation Strategies

Many digital preservation strategies have been proposed, but no single strategy is appropriate for all data types, situations, or institutions. Tristram (2002) describes the following options available for digital preservation:

7.1 Bit-stream Copying

Bit-stream copying, commonly known as "backing up data" refers to the process of making an exact duplicate of a digital object. Though a necessary component of all digital preservation strategies, bit-stream copying in itself is not a long-term maintenance technique, since it deals only with the question of data loss due to hardware and media failure, whether resulting from normal malfunction and decay, malicious destruction or natural disaster. Bit-stream copying is often combined with remote storage so that the original and the copy are not victims of the same disastrous event. Bit-stream copying should be considered the minimum maintenance strategy for even the most lightly valued, ephemeral data.

7.2 Refreshing

Refreshing essentially means copying digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bit-stream (e.g. from a decaying 4mm DAT tape to a new 4mm DAT tape, or from an older CD-RW to a new CD-RW). "Modified refreshing" is the copying to another medium of a similar type with no change in the bit-pattern that is of concern to the application and operating system using the data, e.g. from a QIC tape to a 4mm tape; or from a 100 MB Zip disk to a 750 MB Zip disk. Refreshing is a necessary component of any successful digital preservation project. It potentially addresses both decay and obsolescence issues related to the storage media.

Durable / persistent Media (e.g., Gold CDs)—may reduce the need for refreshing digital contents, and help diminish losses from media deterioration. However, durable media has no impact on any other potential source of loss, including catastrophic physical loss, media obsolescence, as well as obsolescence of encoding and formatting schemes. Durable media has the potential for endangering content by providing a false sense of security.

Copying from medium to medium, however, also suffers limitations as a means of digital preservation. Refreshing digital information by copying will work as an effective preservation technique only as long as the information is encoded in a format that is independent of the particular hardware and software needed to use it and as long as there exists software to manipulate the format in current use. Otherwise, copying depends either on the compatibility of present and past versions of software and generations of hardware or the ability of competing hardware and software product lines to interoperate. In respect of these factors, backward compatibility and interoperability, the rate of technological change poses a serious threat to longevity of digital information.

7.3 Technology Preservation

Technological preservation is based on preserving the technical environment that runs the system, including operating systems, original application software, media drives, etc. It is sometimes called the "computer museum" solution. Technology preservation is more of a disaster recovery strategy for use on digital objects that have not been subjected to a proper digital preservation strategy. It offers the potential of coping with media obsolescence, assuming the media has not decayed beyond readability. It can extend the access for obsolete media and file formats, but is ultimately a dead end, since no obsolete technology can be kept functional indefinitely. This is not a strategy that an individual institution can implement. Maintaining obsolete technology in usable form requires a considerable investment in equipment and personnel.

7.4 Digital Archaeology

Digital archaeology includes methods and procedures to rescue content from damaged media or from obsolete or damaged hardware and software environments. Digital archaeology is explicitly an emergency recovery strategy and usually involves specialized techniques to recover bit-streams from media that has been rendered unreadable, either due to physical damage or hardware failure such as head crashes or magnetic tape crinkling. Digital archaeology is generally carried out by for-profit data recovery companies that maintain a variety of storage hardware (including obsolete types) plus special facilities such as clean rooms for dismantling hard disk drives. Given enough resources, readable bit-streams can often be recovered even from heavily damaged media (especially magnetic media), but if the content is old enough, it may not be possible to make it renderable and /or understandable.

7.5 Analogue Backups

Analogue backups combine the conversion of digital objects into analogue form with the use of durable analogue media, e.g., taking high-quality printouts or the creation of silver halide microfilm from digital images. An analogue copy of a digital object can, in some respects, preserve its content and protect it from obsolescence, without sacrificing any digital qualities, including sharability and

lossless transferability. Text and monochromatic still images are the most amenable to this kind of transfer. Given the cost and limitations of analogue backups, and their relevance to only certain classes of documents, the technique only makes sense for documents whose contents merit the highest level of redundancy and protection from loss.

7.6 Migration

Migration is a broader and richer concept than "refreshing" for identifying the range of options for digital preservation. Migration is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware / software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. Migration includes refreshing as a means of digital preservation but differs from it in the sense that it is not always possible to make an exact digital copy or replica of a database or other information object as hardware and software change and still maintain the compatibility of the object with the new generation of technology.

Migration theoretically goes beyond addressing viability by including the conversion of data to avoid obsolescence not only of the physical storage medium, but of the encoding and format of the data. However, the impact of migrating complex file formats has not been widely tested. Digital objects will have to be constantly migrated and converted to new formats, computing devices, storage media and software to ensure that valuable digital objects are not left behind on obsolete system which will eventually breakdown rendering data inaccessible. The initial conversion of printed-text into digital objects is not only expensive, it would also necessitate diversion of manpower and resources into constant re-invention of wheel in terms of migration of digital resources (Conway, 1996).

7.7 Replication

Replication is used to represent multiple digital preservation strategies. Bit-stream copying is a form of replication. LOCKSS (Lots of Copies Keeps Stuff Safe) is a consortial form of replication, while peer-to-peer data trading is an open, free-market form of replication. In each case, the intention is to enhance the longevity of digital documents while maintaining their authenticity and integrity through copying and the use of multiple storage locations.

7.8 Reliance on Standards

Reliance to standards is mean to "harden" the encoding and formatting of digital objects by adhering to well-recognized standards and discarding proprietary or less-supported standards. It assumes in part that such standards will endure and that problems of compatibility resulting from the evolution of the computing environment (applications software, operating systems) will be handled by the continuing need to accommodate the standard within the new environment. For example, if JPEG2000 becomes a widely adopted standard, the sheer volume of users will guarantee that software to encode, decode, and render JPEG2000 images will be upgraded to meet the demands of new operating systems, CPUs, etc. Like many of the strategies described here, reliance on standards may lessen the immediate threat to a digital document from obsolescence, but it is no more a permanent preservation solution than the use of gold CDs or stone tablets.

7.9 Normalization

Normalization is a formalized implementation of reliance on standards. Within an archival repository, all digital objects of a particular type (e.g., colour images, structured text) are converted into a single chosen file format that is thought to embody the best overall compromise amongst characteristics such as functionality, longevity, and preservability. The advantages and disadvantages of reliance on standards also apply to normalization.

7.10 Canonicalization

Canonicalization is a technique designed to allow determination of whether the essential characteristics of a document have remained intact through a conversion from one format to another. Canonicalization relies on the creation of a representation of a type of digital object that conveys all its key aspects in a highly deterministic manner. Once created, this form could be used to algorithmically verify that a converted file has not lost any of its essence. Canonicalization has been postulated as an aid to integrity testing of file migration, but it has not been implemented.

7.11 Emulation

Emulation uses a special type of software, called an emulator, to translate instructions from original software to execute on new platforms. The old software is said to run "in emulation" on newer platforms. This method attempts to simplify digital preservation by eliminating the need to keep old hardware working. Emulation combines software and hardware to reproduce in all essential characteristics the performance of another computer of a different design, allowing programs or media designed for a particular environment to operate in a different, usually newer environment. Emulation requires the creation of emulator programs that translate code and instructions from one computing environment so it can be properly executed in another.

A widely-known, general purpose emulator is the one built into recent versions of the Apple Macintosh operating system that allows the continued use of programs based on an earlier series of CPUs no longer used in Apple computers. However, most emulators available today were written to allow computer games written for obsolete hardware to run on modern computers.

The emulation concept has been tested in several projects, with generally promising results. However, widespread use of emulation as a long-term digital preservation strategy will require the creation of consortia to perform the technical steps necessary to create functioning emulators as well as the administrative work to assemble specifications and documentation of systems to be emulated and obtain the intellectual property rights of relevant hardware and software.

7.12 Encapsulation

Encapsulation may be seen as a technique of grouping together a digital object and metadata necessary to provide access to that object. Ostensibly, the grouping process lessens the likelihood that any critical component necessary to decode and render a digital object will be lost. Appropriate types of metadata to encapsulate with a digital object include reference, representation, provenance, fixity and context information. Encapsulation is considered a key element of emulation.

7.13 Universal Virtual Computer

Universal Virtual Computer is a form of emulation. It requires the development of a computer program independent of any existing hardware or software that could simulate the basic architecture of every computer since the beginning, including memory, a sequence of registers, and rules for how to move information among them. Users could create and save digital files using the application software of their choice, but all files would also be backed up in a way that could be read by the universal computer. To read the file in the future would require only a single emulation layer—between the universal virtual computer and the computer of that time.

8. Uniform Resource Characteristics (URC) or Metadata

The digital contents, with their increasing size and complexity, need to be identified, described, stored, organized and disseminated to its end users. Uniform and structured meta information can effectively be employed to achieve this goal. Stored in digital repositories, digital objects must have their unique identifications or names that can be used for their retrieval. Uniform Resource Characteristics (URC) or metadata, as more popularly known, provide metadata or meta information about an object, and is analogous to bibliographic records. In other words, metadata is information about information available on the web. The following three types of metadata are associated with the digital objects:

- Descriptive Metadata: Include content or bibliographic description consisting of keywords and subject descriptors.
- Administrative or technical Metadata: Incorporates details on original source, date of creation, version of digital object, file format used, compression technology used, object relationship, etc. Administrative data may reside within or outside the digital object and is required for long-term collection management to ensure longevity of digital collection.
- Structural Metadata: Elements within digital objects that facilitate navigation, e.g. table of contents, index at issue level or volume level, page turning in an electronic book, etc.

Since virtually any metadata element can be seen as having value for preservation purposes, preservation metadata is a separate category, but an amalgamation of all types of metadata. However, preservation metadata may include unique elements and /or finer detail than metadata used for other purposes.

8.1 Digital Preservation Metadata

The digital preservation metadata is a subset of metadata that describes attributes of digital resources essential for its long-term accessibility. Preservation metadata provides structured ways to describe and record information needed to manage the preservation of digital resources. In contrast to descriptive metadata schemas (e.g. MARC, Dublin Core), which are used in the discovery and identification of digital objects, preservation metadata is sometimes considered as a subset of administrative metadata design to assist in the management of technical metadata for assisting continuing access to the digital content. Preservation metadata is intended to store technical details on the format, structure and use of the digital content, the history of all actions performed on the resource including changes and decisions, the authenticity information such as technical features or

custody history, and the responsibilities and rights information applicable to preservation actions. The scope and depth of the preservation metadata required for a given digital preservation activity will vary according to numerous factors, such as the "intensity" of preservation, the length of archival retention, or even the knowledge base of the intended user community.

8.1.1 Open Archival Information System (OAIS)

The OAIS Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS). It is a framework for understanding and applying concepts needed for long-term digital information preservation. It is also a starting point for a model addressing non-digital information. The model establishes terminology and concepts relevant to digital archiving, identifies the key components and processes endemic to most digital archiving activity, and proposes an information model for digital objects and their associated metadata. The reference model does not specify an implementation, and is therefore neutral on digital object types or technological issues. For example, the model can be applied at a broad level to archives handling digital image files, "born-digital" objects, or even physical objects, and no assumptions are imposed concerning the specific implementation of the preservation strategy, for example, migration or emulation. (Sayer, 2001). OAIS has now been adopted as an ISO standard –OAIS is an ISO standard (ISO 14721:2003)

Over past several years, a number of institutions and projects like CEDARS, NEDLIB, the National Library of Australia, Harvard University have released preservation metadata element sets, reflecting a wide range of assumptions, purposes and approaches. The OCLC/RLG Preservation Metadata Framework Working consisting of representatives from leading institutions compared, analysed and consolidated all existing recommendations and expertise. The recommendations of Working Group culminated in June 2002 with production of a framework for implementing preservation metadata documented in "Trusted Digital Repositories: Attributes and Responsibilities (TDR)" The TDR embraces OAIS and demonstrates what adhering to Reference Model for an Open Archival Information System (OAIS) will mean for an institution. The OAIS reference model is being used by many initiatives for developing preservation metadata sets. The OAIS framework enjoys the status of a de facto standard in digital preservation. The OAIS reference model provides a high-level overview of the types of information needed to support digital preservation that can broadly be grouped under two major umbrella terms called i) Preservation Description Information (PDI); and ii) Representation and Descriptive Information.

- Preservation Description Information

The preservation description information consists of four major types of metadata elements, namely reference information, provenance information, context information and fixity information as mentioned below:

- Reference Information: enumerates and describes identifiers assigned to the content information such that it can be referred to unambiguously, both internally and externally to the archive (e.g., ISBN, URN).
- Provenance Information: Documents the history of the content information (e.g., its origins, chain of custody, preservation actions and effects) and helps to support claims of authenticity and integrity.

- Context Information: documents the relationship of the content information to its environment (e.g., why it was created, relationships to other content information).
- Fixity Information: documents authentication mechanisms used to ensure that the content information has not been altered in an undocumented manner (e.g., checksum, digital signature).

- Representation and Descriptive Information

Representation information facilitates proper rendering, understanding, and interpretation of a digital object's content. At the most fundamental level, representation information imparts meaning to an object's bit-stream. For example, it may indicate that a sequence of bits represents text encoded as ASCII characters and furthermore, that the text is in French. The depth of the representation information required depends on the designated community for whom the content is intended. Descriptive Information metadata contains more ephemeral metadata, the information used to aid searching, ordering, and retrieval of the objects.

8.2 PREMIS (PREservation Metadata: Implementation Strategies)

The OAIS Framework prompted interest in moving it toward a more implementable status. In response to this, OCLC and RLG sponsored a second working group called PREMIS (PREservation Metadata: Implementation Strategies). Composed of more than thirty international experts in preservation metadata, PREMIS sought to: i) define a core set of implementable, broadly applicable preservation metadata elements, supported by a data dictionary; and ii) identify and evaluate alternative strategies for encoding, storing, managing, and exchanging preservation metadata in digital archiving systems. In September 2004, PREMIS released a survey report describing current practice and emerging trends associated with the management and use of preservation metadata to support repository functions and policies. The final report of the PREMIS Working Group was released in May 2005. The PREMIS Data Dictionary is a comprehensive, practical resource for implementing preservation metadata in digital archiving systems. It defines implementable, core preservation metadata, along with guidelines and recommendations for management and use. A maintenance activity has been set up to manage the Data Dictionary and coordinate future revisions.

9. Storage Management for Digital Preservation

One of the crucial threats to digital preservation is short life of storage media, obsolete hardware and software, and slow read times of old media. While the selection and installation of software components are crucial to building a digital repository, the core of the repository is the storage infrastructure. The basic tenets of digital preservation extend much beyond storage media life. Devices used for reading storage media rapidly become obsolete, various formats (and their changing versions) of digital documents and images introduce additional complications. The storage operation in digital archives primarily addresses to the media level formatting of information objects. Primary considerations for digital preservation include levels of hierarchy and redundancy. A digital archive may have multiple levels of storage depending upon the levels of expected use and expected retrieval performance. Digital repositories that are too large to store on a single disk can use hierarchical storage mechanisms (HSM). In an HSM, the most frequently used data is kept on fast disks while less frequently used data is kept in nearline such as an automated (robotic) tape library. An HSM can

automatically migrate data from tape to disk and vice-versa as required. Digital material in a distributed network may be stored online in multiple locations. Besides offline and online storage, near-line storage may be adopted wherein information objects may be stored on optical or tape media and loaded in a jukebox. Retrieval time in near-line storage systems is higher in comparison to online storage, but is considerably more responsive to user demand than off-line storage. A digital archive may use any or all of these methods. The most sophisticated systems combine the resources so that objects in use or recent use are stored online and, as they age from the time of most recent use, they move to near-line storage and then eventually to off-line storage.

Redundancy is another important storage consideration. In a system that is completely dependent on the interaction of various kinds and levels of hardware and software, failure in any one of the subsystems could mean the loss or corruption of the information object. Effective storage management thus means providing for redundant copies of the archived objects to ensure availability of documents in case of loss. A number of RAID (Redundant Array of Inexpensive Disks) models are now available for greater security and performance. The RAID technology distributes the data across a number of disks in a way that even if one or more disks fail, the system would still function while the failed component is replaced. Digital archives may also choose to make backup copies on their own or to make arrangements for other sites to serve as backup.

Although harddisc (fixed and removable) solutions are increasingly available at an affordable cost, optical storage devices including WORM, CD-R, CD ROM, DVD ROM or opto-magnetic devices in standalone or networked mode, are attractive alternatives for long-term storage of digital information. Optical drives record information by writing data onto the disc with a laser beam. The media offer enormous storage capabilities. Some of the important features of storage infrastructure for satisfying requirements of digital preservation are as follows:

- Increased scalability: The storage media should be scalable depending on the requirement of a digital archive.
- Availability of storage devices to multiple servers: The storage system should be a sharable device that can be accessible from multiple servers. Increased availability and sharing among storage devices allows for effective load balancing and redundancy. Intelligent storage networks and snap-servers are now available in which the physical storage devices are intelligently controlled and made available to a number of servers.
- High-speed throughput: The storage device should utilize Fibre Channel, for carrying traffic between devices at high speed.
- Separation from the LAN: The storage system attached to a digital repository should only be accessible via devices physically connected to it so that the storage system remains unaffected by traffic on the user LAN and vice versa.

10. Microfilming and Digital Preservation: A Hybrid Solution

Microfilming is a tried and tested technology for preservation of documents with proven longitivity. The life expectancy of microfilm is in the 500+ year range. Microfilm master, if stored properly, is quite simply the most stable reformatting method available. Don Willis (1992), is a report published by the Commission on Preservation and Access, argued convincingly for the creation of both microfilm for preservation and digital images for access. The proposed hybrid solution suggests microfilming

of document as first step and then digitized from the film master. It is argued that for a computer image to match the resolution of high-resolution microfilm, the item would need to be scanned at over 5,000 dots per inch, which is practically impossible with prevailing scanning technology as it would require incredible scanning time and storage space. Moreover, neither the scanners are designed to scan at such a high resolution nor the documents scanned at such a high resolution can be displayed using present day technology. The hybrid solution provides the best of both worlds. The high-resolution microfilm masters can be safely archived, and retrieved when needed to generate new high-use, highly accessible digital version. The process also serves to circumvent the problems with digital technology, i.e. constant migration. New digital files in successive software generations could be created as required from the microfilm master (Davis, 1997).

11. Conclusion

Preservation in the digital world is a challenging task for librarians and archivists. However, protocols, strategies and technologies involved in digital preservation have now been well defined and understood. Digital preservation is a cost-intensive activity of continuing nature. Library, archives, or museum cannot make a decision to adopt digitization with long-term preservation and storage of research collections without deep and continuing commitment to preservation by the parent institution. The preservation in digital world is no more a prerogative of the libraries, but has become the mandate of the parent institution. The necessary financial and technological commitments to maintain digital contents and to migrate it to future generations must be an organizational commitment. Failure to address to the well-defined digital preservation problems and strategies may result in loss of valuable digital data and may contribute to cultural and intellectual loss resulting in exorbitant costs for recovery, if at all possible. Librarians are compelled to meet the research challenge to resolve the conflict between the creation context and the use context to facilitate digital information preservation.

Digital resources, undoubtedly, have several advantages over its analogue counter part, however, preservation is definitively not one of them. The fact that the risk of loss of data in digital form is much greater than any other physical form is well understood and addressed to. Long-term preservation of digital information is plagued by short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites (Chen, 2001).

References

1. Arora, Jagdish. Integrating network-enabled digitized collection with traditional library and information services: Brewing a heady cocktail at the IIT Delhi. In: IT and Digital Library Development (ed. Ching-chih Chen). West Newton, MicroUse Information, p. 7-16, 1999.
2. Conway, Paul. Preservation in digital world. Microform and Imaging Review, 25(4), 156-171,1997. Also available online (http://www.clir.org/pubs/reports/conway2/) (last visited on 4th Oct., 2006)
3. Cornell University Library. Tutorial on digital preservation management: Implementing short-term strategies for long-term problems. 2005 (http://www.library.cornell.edu/iris/tutorial/dpm/index.html) (last visited on 4th Oct., 2006)
4. Davis, Eric T. An overview of the access and preservation capabilities in digital technology. 1997. (http://www.iwaynet.net/~lsci/diglib/digpapff.html)
5. Digital Preservation Coalition. (http://www.dpconline.org/) (last visited on 4th Oct., 2006)

6. Dollar, Charles M. Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods. Macerata: University of Macerata Press, 1992.
7. Feeney, M. (ed). Digital Culture: Maximising the Nation's Investment. London, The National Preservation Office, 1999. p.11.
8. Gertz, Janet, et al. Oversize Color Images Project, 1994-1995: Final Report of Phase I. Washington, D.C.: Commission on Preservation and Access, 1995.
9. Hedstrom, M. and Montgomery, S. Digital Preservation Needs and Requirements in RLG Member Institutions. Mountain View, CA: RLG., 1998. (http://www.rlg.org/preserv/digpres.html) (last visited on 4th Oct., 2006)
10. Jantz, Ronald and Giarlo, Michael J. Architecture and Technology for Trusted Digital Repositories. D-Lib Magazine, 11 (6), 2005.
11. Lynch, Clifford. The integrity of digital information: Mechanics and definitional issues. Journal of the American Society for Information Science, 45, 737-44, 1994.
12. Mintzer, Fred, and John D. McFall. Organization of a system for managing the text and images that describe an art collection. SPIE Image Handling and Reproduction Systems Integration,1460, 1991.
13. Preserving digital information: Draft Report of the Task Force on Archiving of Digital Information. Version 1.0 August 23, 1995. Research Libraries Group and Commission on Preservation and Access. URL: http://www.oclc.org:5046/~weibel/archtf.html. (last visited on 4th Oct., 2006)
14. Sayer, Donald, et al (2001). The Open Archival Information System (OAIS) Reference Model and its usage.
15. http://public.ccsds.org/publications/documents/SO2002/SPACEOPS02_P_T5_39.PDF (last visited on 4th Oct., 2006)
16. Sebera, Donald. The effects of strengthening and deacidification on paper permanence: Some fundamental considerations. Book & Paper Group Annual, 9. Washington, D.C., American Institute for Conservation, pp. 65-117, 1990.
17. Spinellis, D. The decay and failures of web references. Communications of the ACM, 46, (1), 71 – 77, 2002.
18. Tristram, Claire. Data Extinction, MIT Technology Review, October 2002, p.42.
19. Van Bogart, John W. Magnetic tape storage and handling: A guide for libraries and archives. Washington, D.C.: Commission on Preservation and Access, 1995.
20. Willis, Don. A hybrid systems approach to preservation of printed materials. Washington, D.C., Commission on Preservation and Access, 1992.
21. Wikipedia, the Free encyclopedia, 2006 (http://en.wikipedia.org/wiki/) (last visited on 4th Oct., 2006)