Digital Preservation: 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.

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.

* *This paper is reproduced from Proceeding of PLANNER 2006, published by INFLIBNET Centre.*

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 way back in 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, is generally not given serious consideration that it deserve. The fact that the risk of loss of data in digital form is much greater than any other physical form is required to be addressed in greater details.

Digital documents are vulnerable to loss because of decay and obsolescence of the media on which they are stored, 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. Digital preservation addresses the issue of adapting concepts of preservation to manage risk of loss of digital contents due to rapid technological advancements. 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 librarian's 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".

3. Why Digital Preservation?

Traditional libraries are increasingly getting transformed into digital libraries, atleast partially. The availability 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-parts 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 continuously 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 contents 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 for a brief period of time. As a result, even if the storage media is retained in the best condition, it may still be not 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 web 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

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

The primary objective of digitisation is to improve access to high quality information. Digitization offers dual benefits, it preserves rare and fragile objects as well as it provides enhanced access to multiple numbers of users simultaneously at remote locations. The concept of digital preservation is not restricted to "preservation" only, it has the following 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. While "digital copies" can be used for providing access to the users, direct access to physical document can be restricted to few. Digital imaging can thus become a "preservation application" as oppose to "access application". The original print-based collection is preserved in a linear sequence. Digital images are made more accessible using sophisticated indexing schemes to facilitate browsing and searching of digital collections. Preservation via digital copying has been the most compelling force motivating archives and libraries to experiment with hardware and software capabilities for preservation purposes.

5.3. Quality Preservation

The digital information has potential for qualitative preservation of information. The preservation-quality images can be scanned at high resolution and bit depth for best possible quality. The quality remains the same inspite of multiple usage by several users. However, caution need to be exercised while choosing digitised information as preservation media.

5.4. Maintain Digital Objects

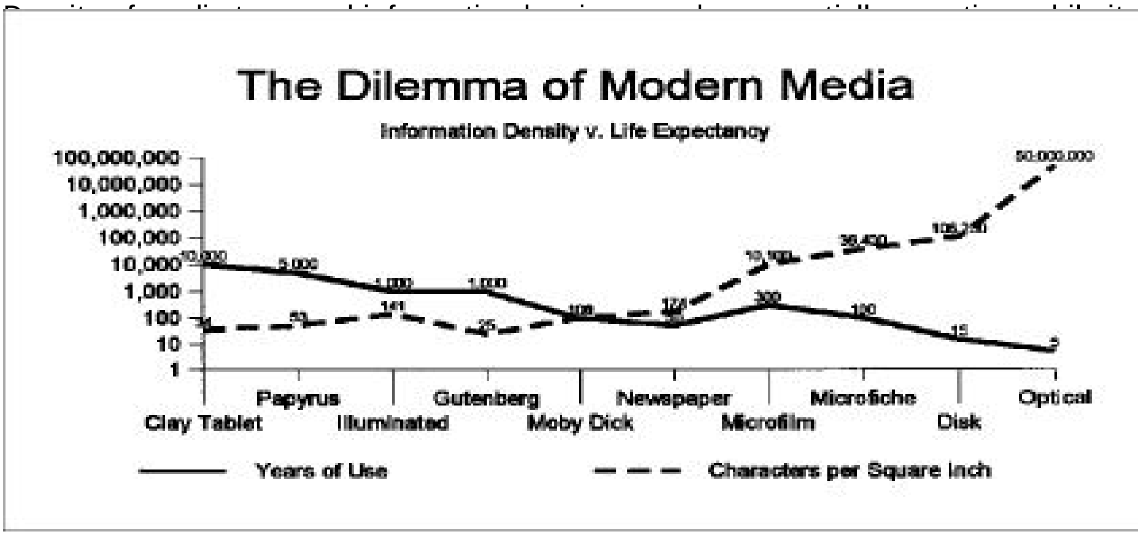
Once digital conversion of the original document has been completed, the challenge of protecting the digit contents from corruption or destruction becomes the preservation focus. This facet, called “digital preservation”, typically centers on the choice of interim storage media, the life expectancy of 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).

5.5. Enhanced Access to Digital Objects

Digital technologies present a preservation solution for the documents in the libraries with enhanced 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 retrieval and 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.



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.

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

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. Several strategies have been proposed but it is unlikely to find a single solution that is appropriate for all data types, situations, or institutions (Tristram, 2002). A set of digital preservation strategies can be applied depending on the life-span of a digital object as mentioned below:

- 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.

UNESCO’s Guidelines for the Preservation of Digital Heritage (2003) group these strategies under the following four categories:

- 7.1. Short-term Strategies
  - 7.11. Bit-stream Copying
  - 7.12. Refreshing
  - 7.13. Replication
  - 7.14. Technology Preservation or Computer Museum
  - 7.15. Backwards Compatibility and Version Migration
- 7.2. Medium- to Long-term Strategies
  - 7.21. Migration
  - 7.22. Viewers and Migration at the Point of Access Emulation
  - 7.23. Canonicalization
  - 7.24. Emulation
- 7.3. Investment Strategies
  - 7.31. Restricting Range of Formats and Standards
  - 7.32. Reliance on Standards
  - 7.33. Data Abstraction and Structuring
  - 7.34. Encapsulation
  - 7.35. Software Re-engineering
  - 7.36. Universal Virtual Computer
- 7.4. Alternative strategies
  - 7.41. Analogue Backups
  - 7.42. Digital Archaeology or Data Recovery
- 7.5. Combinations

These strategies have demonstrated to work in certain circumstances over limited periods of time. None of them have proven themselves against unknown threats over centuries of change. Most of these strategies are, however, being used in the management of data, and it is likely that combinations of these strategies will continue to be researched and proposed for large-scale, long-term preservation. It is, therefore, reasonable for preservation programmes to look for multiple strategies, especially if they are responsible for a range of materials over extended periods.

7.1. Short-term Strategies

Short-term digital preservation strategies are likely to work for a short period of time only. These strategies include bit-stream copying, refreshing, replication, technology preservation or computer museum, backwards compatibility and version migration.

7.11. 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.

7.12. 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.

7.13. 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. LOCKSS uses low-cost tools to crawl the Web to cache "redundant, distributed, decentralized" e-journal presentation files for which a library has a subscription or license. LOCKSS supports the traditional model whereby individual libraries build and maintain local collections of journals, and work is underway to develop a user interface for local collection management of e-journals cached using the LOCKSS system. A LOCKSS Alliance of participating libraries has been formed and the system is currently in beta test mode.

The intention of replication is to enhance the longevity of digital documents while maintaining their authenticity and integrity through copying and the use of multiple storage locations.

7.14. Technology Preservation

Technological preservation is based on keeping and maintaining the technical environment that is used for creation of contents including operating systems, original application software, media drives, etc. It is sometimes called the "computer museum" solution. In other words, technological preservation becomes applicable to digital materials that are left on obsolete storage media and hardware and software required to access them are discarded. 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. The strategy requires management and maintenance of a wide range of equipment and software, along with ancillary materials such as manuals and licenses, which may be difficult and expensive to achieve. Maintaining obsolete technology in usable form requires a considerable investment in equipment and personnel.

7.15. Backwards Compatibility and Version Migration

This strategy relies on the ability of current versions of software to interpret and present digital material created with previous versions of the same software and to save them in current format. In the case of backwards compatibility, the presentation may be limited to temporary viewing, whereas version migration permanently converts documents into a format that can be presented by the current version of the software. For example, most web browsers are capable of interpreting and displaying material written using earlier versions of the HTML standard. MS Word, Excel and Access applications, usually allow previous versions of their file formats to be transformed and resaved in a new version, as part of application upgrade paths.

The option to convert into current version may not be available for all types of objects. Internet Browser and PDF Readers, for examples, only present the older version of document without giving option to save them into current version. Even those applications that offer backward compatibility with provision to migrate to current version, it is unlikely that backward compatibility will be retained over many generations of the software. Moreover, the process of migration is likely to introduce unwanted changes to a document incrementally if used over many generations.

7.2. Medium to Long-term Preservation Strategies

Strategies proposed for medium and long-term preservation are likely to work for a long period of time. Such strategies should be used for digital materials that are likely to be of value for a long period of time. Medium and long-term preservation strategies include:

7.21. Migration

Migration is a broader and richer concept of digital preservation than “refreshing”. 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).

7.22. Viewers and Migration at the Point of Access

Migration or providing viewing facility at the point of access has been proposed as an alternative to recurring and incremental migration. The process involves use of appropriate viewers, software tools or transformation methods that provide accessibility at the time of access, using the original data stream. For example:

i) The “migration on request” approach has been developed in CEDARS and CAMiLEON projects that uses a software tool to record method of access. Access to digital object is provided “on the fly” depending on the method of access. As technology changes, the software is updated to reflect changes in the method of access (Cedars, 2002; Mellor, Sergeant and Wheatley, 2003).

ii) The TOMS (Typed Object Model Server) approach provides transformation methods for common document and data types, allowing a server to choose a suitable transformation path for a range of object types. (Thibodeau, 2002)

iii) The VERS strategy converts documents to a PDF format on the basis that third-party viewers for PDF may be constructed from the format specification.

Limitations of this approach includes i) viewers may not be available for all formats such as executable files; ii) Viewers may be able to represent some, but not all, elements of digital materials; iii) the gap between the original format and the prevailing technologies at the time of access may be too great for the tools or methods to cope with; and iv) Viewers, tools or methods, and corresponding metadata must also be maintained or adjusted as technologies change.

7.23. 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.

7.24. 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.

Limitation of this strategy are i) it is technically complex, cost-intensive, labour-intensive and requires a high degree of expertise; ii) emulation is still in the research stage; iii) Effective emulation could be frustrated by inadequate documentation of software, or by non-standard use of file formats; iv) As systems become more complex, so will the requirements for emulation, which may need to include multiple components. Moreover, emulation of all aspects of a system or application may not be possible; and v) As technology and platforms change over time, emulators themselves will either have to migrate to, or have their host systems emulated on, the new platform, potentially leading to layers upon layers of emulators.

7.3. Investment Strategies

Investment preservation strategies involve investment of efforts at the time of archiving digital materials. Such strategies include: Restricting Formats and Standards, Reliance on Standards, Data Abstraction and Structuring, Encapsulation, Software Re-engineering and Universal Virtual Computer.

7.31. Restricting Formats and Standards

Preservation programmes may decide to only store data in a limited range of formats and standards. This can be achieved either by only accepting material in specified formats or by converting material from other formats before storage. All digital objects within an archival repository of a particular type (e.g., colour images, structured text) can be converted into a single chosen file format that is thought to embody the best overall compromise amongst characteristics such as functionality, longevity, and preservability. For, example most of the textual and graphical information can be converted into PDF format. The UK Archaeology Data Service (ADS), for example, specifies a preferred (but not exclusive) range of formats for deposit and provides guidelines for depositors on creating or preparing materials for submission.

The strategy does not necessarily solve the access problem unless the obsolescence of formats and standards used are handled effective through some other strategy. This strategy imposes serious restrictions on the range of materials that a preservation programme can accept. Moreover, the process of conversion from original format may cause some loss of essential elements.

7.32. Reliance on Standards

This preservation strategy involves the use of open, widely available and supported standards and file formats that are likely to stable for a longer period of time discarding proprietary or less-supported standards. Such standards or formats may either be formally agreed or may be de facto standard formats that have been widely adopted by industry. For example, majority of digitisation programmes choose TIFF (Tagged Image File Format) as an open, stable and widely supported standard for creation of preservation master images. Similarly, most publisher use PDF as de facto standard for electronic distribution of their research articles, due to the availability of PDF readers for all platforms. Reliance on standards may lessen the immediate threat to a digital document from obsolescence, but it is not a permanent preservation solution.

7.33. Data Abstraction and Structuring

Data abstraction, sometimes also called normalization, involves analyzing and tagging data so that the functions, relationships and structure of specific elements can be described. Using data abstraction, the representation of content can be liberated from specific software applications, the digital contents can, however, be read using different applications as technology changes. Data abstraction makes a document application-independence and simplifies the transport of data between platforms and over generations of technology. The technique, however, has its limitation, it requires extensive development of tools and methods for analysis and processing in order to correctly represent and tag each type of data. Moreover, technology eventually used for presentation may still limit what functions can be represented.

The San Diego Supercomputer Center, for example, have used custom algorithms to apply XML tags to a collection of one million emails (Moore et al, 2000).

7.34. Encapsulation

Encapsulation may be seen as a technique of grouping together digital objects and metadata necessary to describe and provide access to that object. The grouping process lessens the likelihood that any critical component necessary to decode and render a digital object will be lost. Encapsulation is considered a key element of emulation.

Encapsulation may also bundle metadata that describe or provide link to the software applications and platform used for original contents considering the fact that it is impractical and unnecessary to encapsulate the software. Open Archival Information System (OAIS) Reference Model, for example, describes incorporating data objects and their associated metadata into Archival Information Packages (AIPs).

7.35. Software Re-engineering

Digital materials are mostly tied to the application software used for creating them. The application software, in turn, are dependent on a specific system or platform in order to function. Application software get most affected by changes in technology. Moreover, they are also usually unsuited for preservation strategies, including regular migration. Software reengineering may offer a number of strategies for transforming software as technologies change, similar to transformation of data formats. Some possibilities include:

i) Adjustment and re-compiling of source code for a new platform;

ii) Reverse-engineering of compiled code into higher level code and porting that to the new platform;

iii) Re-coding of the software from scratch, or re-coding in another programming language; and

iv) Translation of compiled binary instructions for one platform directly into binary instructions for another platform.

Reengineering application would require source code, which may not be available except for open source programmes and software that are developed in-house. Even when source code is available, porting to other platforms is not a trivial job, it requires considerable time and effort per object. Moreover, compilers or interpreters are required for the new platform for the code language.

7.36. Universal Virtual Computer

Universal Virtual Computer is a form of emulation. It requires 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.

This approach requires substantial investments both at the time of archiving while developing encoding methods or UVC-native interpretive programmes for each data type as well as at the time of restoration in developing a UVC emulator and restore programmes. Moreover, if original data objects are abstracted or transformed for encoding purposes, such transformation may discard essential characteristics.

The proof-of-concept prototype for the UVC approach (Lorie, 2002) has been used to produce a logical schema, decoder programme and representation mechanism for PDF documents, such that the document content can be represented using a UVC interpreter and restore programme.

7.4. Alternative Strategies

Alternative strategies to digital preservation include taking analogue backup of document (print or microfilm) or recovering data from obsolete digital media.

7.41. 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.

Limitation of this strategy includes i) advantages offered by digital technology such as convenience of use, storage efficiency, search and navigation possibility is lost; ii) the strategy does not completely remove the threat of technological obsolescence; and iii) long-term stability of analogue material may depend on expensive storage environments that prove to be less reliable than well-managed computer systems based on high levels of redundancy.

7.42. 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. Combination Strategies

As mentioned before, no single strategy is appropriate for all data types, situations, or institutions. A number of strategies may, therefore, be necessary to cover the range of objects and characteristics to be preserved. Preservation programmes should also consider the potential benefits of redundancy in pursuing more than one strategy. It may be noted that even with good planning, a single strategy may fail leaving the programme with no means of access. Several digital preservation projects use more than one approach, for example:

i) Standards such as TIFF for image collections are often chosen in preparation for eventual migration to other standard formats over the long-term;

ii) The VERS strategy couples the use of standards (PDF, XML) to the future use of viewers and the likely migration of XML encoded metadata in the future;

iii) Persistent archives (Moore, 2001) use data abstraction with the view to eventual migration – migration of the data, the mark up system and the supporting software, and upgrading of hardware;

iv) The Universal Virtual Computer (UVC) approach combines data abstraction with rules for migration of data objects at the point of access, and an emulation approach for software objects. The “durable encoding” approach adds the use of fundamental standards for encoding data, including encoding that could be understood by the UVC.

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 deployed 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 four types of metadata are associated with the digital objects:

- **Descriptive Metadata**: Descriptive metadata is used to describe textual / non-textual contents of a digital object. It includes content or bibliographic description consisting of keywords and subject descriptors that may be assigned using controlled vocabulary or thesaurus like Medical Subject Headings (MESH), INSPEC Thesaurus, Library of Congress Subject Headings (LCSH).
- **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 metadata 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.
- **Identification Metadata**: Used for tracking different versions and editions of same digital work, i.e. pdf, HTML, PostScript, MS Word, etc. and TIFF, JPG, BMP, etc. in case of images.

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 details 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.

8.11. Open Archival Information System (OAIS)

The OAIS Reference Model was developed by the Consultative Committee for Space Data Systems (CCSDS) as a conceptual framework describing the environment, functional components and information objects associated with a system responsible for the long-term preservation of digital materials. The model establishes terminology and concepts relevant to digital archiving, identifies the key components and processes endemic to

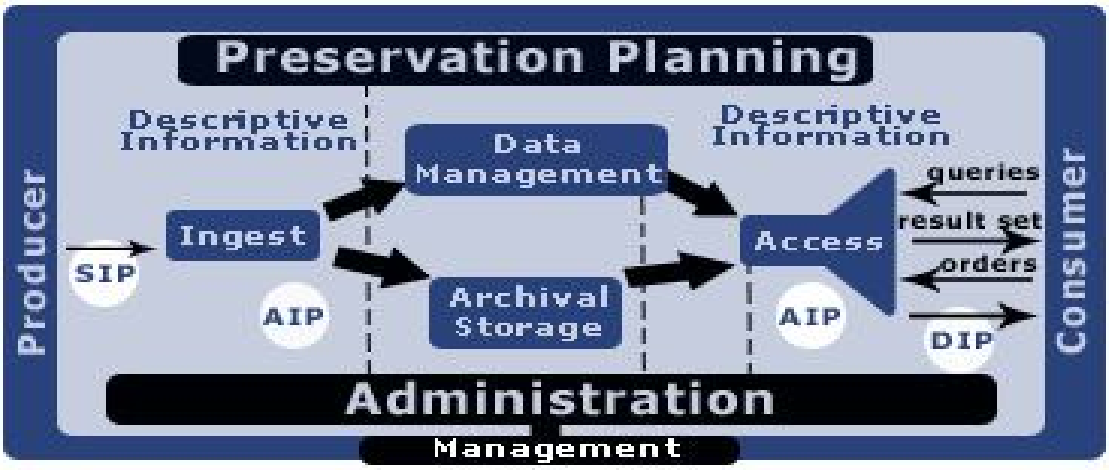
ıg

ed

d

d

y



Open Archival Information System (OAIS) Model

Figure given below provides thematic diagram of OAIS Reference Model.

The metadata in OAIS Model plays an essential role in preserving digital content and supporting its use over the long-term. The OAIS information model implicitly establishes the link between metadata and digital preservation – i.e., preservation metadata. 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 the following four major types of metadata elements:

i) 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).

ii) 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.

iii) Context Information: documents the relationship of the content information to its environment (e.g., why it was created, relationships to other content information).

iv) 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.

The reference model does not specify an implementation, and is therefore neutral on digital object types or technological issues (Sayer, 2001). OAIS has now been adopted as an ISO standard –OAIS is an ISO standard (ISO 14721:2003)

Using conceptual framework of OAIS Model, a number of institutions and projects like CEDARS, NEDLIB, the National Library of Australia and 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.

8.12. PREMIS (PREservation Metadata: Implementation Strategies)

The OAIS Framework prompted interest in moving it toward a more implementable status. To achieve this objective, 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. PREMIS also developed a set of XML schema to support use of the Data Dictionary by institutions managing and exchanging PREMIS conformant preservation metadata.

8.13. METS: A Standard for Packaging Metadata and Content together in Digital Repository System

METS (Metadata Encoding and Transmission Standard) is designed to organize and link different types of metadata to its associated content in the OAIS Reference Model. METS is an XML schema designed specifically as an overall framework within which all the metadata associated with a digital object can be stored. A METS file comprises of the following four major constituent sections:

i) A file inventory for all the files associated with the digital object including still image files, text, video or audio files;

ii) A section for administrative metadata;

iii) A section for descriptive metadata; and

iv) A structural map, which indicates in a hierarchical manner how the various components of the item relate to each other, so allowing its constituent elements to be navigated by the user.

These four sections are linked to each other by means of identifiers. An item in the structural map corresponding to a page in a digitized material will have pointers to the files in the file inventory which contain the scanned image of that page or a marked-up version of its constituent text, another pointer to the part of the descriptive metadata section which contains a full description of its intellectual content, and another to the part of the administrative metadata section which contains technical and rights information necessary to deliver the images or text.

METS allows two approaches to the storage of the metadata and data associated with a digital object: both may be either stored internally within the METS file, or held externally

and referenced from within METS. The flexibility of METS implies that its practical implementation can be very flexible as well. Any system capable of handling XML documents can be used to create, store and deliver METS-based metadata. (Lavoie, 2005).

9. Major Functions of a Preservation Programmes

The UNESCO's Guidelines for the Preservation of Digital Heritage (2003), describes the following functions that a full-fledged digital preservation programmes should perform:

9.1. Creation of a Safe Place

Preservation programmes must identify or create a safe place for storing and managing digital materials. Organizations may either set-up its own infrastructure or outsource its digital preservation activity to a reliable third party. However, it is the responsibility of the concern organization to ensure long-term availability of their digital contents.

9.2. Ingest

The process of loading a digital file into a digital repository along with its descriptive metadata for subsequent retrieval is referred to as ingest. The steps involved in ingest include:

i) Applying collection policies and selection criteria to ascertain whether material can be accepted on submission or not;

ii) Checking the quality of the material submitted, including its completeness, authenticity and ascertain that the material has duly been scanned for viruses;

iii) Assigning unique identifiers to digital objects;

iv) Assigning and managing copyright of digital material;

v) Assessing the elements that must be maintained, and assigning preservation objectives;

vi) Setting retention and review periods for the digital material as deemed appropriate;

vii) Checking and upgrading the documentation that describes the material, including the technical and preservation metadata;

viii) Checking the file format(s) and converting them into another format if so desired to comply with the policy of digital preservation programme;

ix) Saving digital objects and associated metadata after verification to the archival storage system

9.3. Archival Storage

A digital preservation programme should provide archival storage that maintains, protects and verifies the integrity of the stored digital objects and associated metadata, whether stored as a single data stream or as separate but linked data streams.

The archival storage system must include practices to ensure protection of data stream from unintended change, damage or loss which can be achieved by regular copying of the data stream to fresh media, or to new media types, when necessary. Storage practices should also includes regular checks such as: checking of data stream against corruption, system security; backup regimes that place copies at remote sites and disaster recovery plans that address contingencies such as complete loss of the system's operating infrastructure.

9.4. Preservation Planning

The basic function of preservation planning is to monitor threats to accessibility to digital material and to specify action required to counter such threats. While archival storage offers data protection, its continuing access has to be ensured. The technology changes that affect accessibility should be monitored. The remedial action may involve migrating or upgrading of the digital object into different format or encoding or changing the metadata that describes the means of access and links to current access tools.

9.5. Implementing Preservation Strategy

Preservation strategies discussed above differ substantially in their method of preserving digital records. However, the process of implementation is quite similar. Steps involved in implementing preservation strategies adopted by the National Archives of Australia (2007) are given below: Identify Materials Requiring Preservation**:**

i) Identify and select digital materials that require preservation treatments.

ii) Research Appropriate Preservation Strategy: Investigate the hardware and software technologies required to successfully implement the preferred preservation approach. Different preservation strategies may have different pre-requisites. For example, in the case of emulation, it may involve the development of specialised software capable of re-creating the source records within a new computer environment. In the case of migration, this may involve identifying suitable migration paths (i.e. software applications with sufficient backward compatibility to transfer source records from an outmoded data format to a current data format). In the case of encapsulation, this may involve software with the ability to embed metadata or 'package' it with the record.

iii) Test Proposed Solution: Before a preservation approach is fully implemented, comprehensive testing of the technical processes must be conducted. Testing should be performed on duplicates of source records.

iv) Back up Records Identified for Preservation: Prior to implementation, all digital records identified for preservation treatment should be backed up with its integrity duly verified. These duplicate source records should not be subjected to a preservation process and will serve as master copies should the selected preservation treatment be unsuccessful.

v) Apply the Preservation Treatment: After successful testing, the treatment should be applied to all digital records identified for preservation treatment. This treatment would vary from one preservation strategy to another, For example, for migration and encapsulation techniques, it would entail applying preservation treatments to the source records, thereby altering their format. For an emulation-based technique, the records identified for preservation would be transferred to the new environment – without altering the records themselves.

V) Apply the Preservation Treatment The preserved records should be subjected to rigorous testing to ensure that there is no loss of content and change in its structure or format. The integrity of all relevant metadata associated with the preserved records should be verified. Metadata should also be updated to record the preservation treatment.

vii) Destroy Source Records where Appropriate: Once the preservation process has been completed and the integrity of the preserved records has been verified, duplicate source records may be destroyed.

viii) **Establish Monitoring Regimes:** The integrity of the preserved records, their functionality, structure, content and context, and associated metadata, should be monitored periodically following preservation to ensure the stability of the preserved records and to identify when subsequent preservation treatments are required.

9.6. Data Management

Managing digital materials in the archive generates its own data about what material is stored, what can be accessed, and about the management of the archive. This data must be managed to support use of the archive, and to support its effective administration.

9.7. Access

This function provides a user interface to the archive, allowing users to browse, search and discover its holdings, to request for material and receive its copies. Access to archives may either be restricted or it may be made available to all potential users. The access function may well require mechanisms to control access.

9.8. Liaison and Advocacy

The preservation programme and libraries must advocate good practices among producers of digital contents with an aim to facilitate long-term availability of the material for which the programme will be responsible. There is also a need to understand who would be the likely users of the material, so that preservation and access arrangements can be tailored to their needs and expectations.

9.9. Management, Administration and Support Functions

A digital preservation programme must be managed professionally. It involves evelopment of policy frameworks and standards covering all areas of operations. The responsibilities and functions mentioned above are described in detail in the Reference Model for Open Archival Information Systems (OAIS) released in 2002.

10. 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 storage of digital materials 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 Network Attached Storage (NAS) 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.

11. 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 display 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).

12. 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. **Cedars Project.** Cedars guide to digital preservation strategies. Cedars, University of Leeds, 2002.(http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html)

3. **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)

4. **Cornell University Library.** Tutorial on digital preservation management: Implementing short-term strategies for long-term problems. 2005

5. (http://www.library.cornell.edu/iris/tutorial/dpm/index.html) (last visited on 4th Oct., 2006)

6. **Davis, Eric T.** An overview of the access and preservation capabilities in digital technology. 1997. (*http://www.iwaynet.net/~lsci/diglib/digpapff.html*)

7. **Digital Preservation Coalition.** (http://www.dpconline.org/) (last visited on 4th Oct., 2006)

8. **Dollar, Charles M.** Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods. Macerata: University of Macerata Press, 1992.

9. **Feeney, M.** (ed). Digital Culture: Maximising the Nation's Investment. London, The National Preservation Office, 1999. p.11.

10. **Gertz, Janet,** et al. Oversize Color Images Project, 1994-1995: Final Report of Phase I. Washington, D.C.: Commission on Preservation and Access, 1995.

11. **Hedstrom, M. and Montgomery, S.** Digital Preservation Needs and Requirements in RLG Member Institutions. Mountain View, CA: RLG., 1998.

12. (http://www.rlg.org/preserv/digpres.html) (last visited on 4th Oct., 2006)

13. **Jantz, Ronald and Giarlo, Michael J**. Architecture and Technology for Trusted Digital Repositories. D-Lib Magazine, 11 (6), 2005.

14. **Lavoie, Brian and Gartner, Richard.** Preservation metadata. Digital Preservation Coalition, 2005. (DPC Technology Watch Series Report 05-01).

15. **Lorie R.** The UVC: A method for preserving digital documents - Proof of concept. Amsterdam, IBM Netherlands, 2002. (http://www.kb.nl/kb/ict/dea/ltp/reports/4-uvc.pdf)

16. **Lynch, Clifford.** The integrity of digital information: Mechanics and definitional issues. *Journal of the American Society for Information Science*, 45, 737-44, 1994.

17. **Mellor, P., Sergeant, D., Wheatley, P.** Migration on request: A practical technique for preservation. CAMiLEON Project, University of Michigan, 2002. (http://www.si.umich.edu/CAMILEON/reports/migreq.pdf)

18. **Moore, R.** et al. Collection-based persistent digital archives – Part 1. D-Lib Magazine, 6(3), 2000. (http://www.dlib.org/dlib/march00/moore/03moore-pt1.html)

19. **Moore R.** et al. Collection-based persistent digital archives – Part 2. D-Lib Magazine, 6(4), 2000. (http://www.dlib.org/dlib/april00/moore/04moore-pt2.html)

20. **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.

21. National Archives of Australia. UNESCO's Guidelines for the Preservation of Digital Heritage. Paris, UNESCO, 2003. (http://www.unesco.org/webworld/mdm). Last visited on 3rd May, 2007.

22. National Archives of Australia. Digital record keeping guidelines: Preserving digital records for the long term. (http://www.naa.gov.au/recordkeeping/er/guidelines/10-preservation.html). Last visited on 3rd May, 2007.

23. 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)

24. **Sayer, Donald,** et al (2001). The Open Archival Information System (OAIS) Reference Model and its usage. (http://public.ccsds.org/publications/documents/SO2002/SPACEOPS02_P_T5_39.PDF) (last visited on 4th Oct., 2006)

25. **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.

26. **Spinellis, D.** The decay and failures of web references. Communications of the ACM, 46, (1), 71 – 77, 2002.

27. **Thibodeau K.** Overview of technological approaches to digital preservation and challenges in coming years. In: The State of Digital Preservation: An International Perspective – Conference Proceedings, Documentation Abstracts, Inc., Institutes for Information Science, Washington, D.C., April 24025, 2002. Council on Library and Information Resources, Washington, D.C., 2002.(http://www.clir.org/pubs/reports/pub107/thibodeau.html)

28. **Tristram, Claire.** Data Extinction, MIT Technology Review, October 2002, p.42.

29. **Van Bogart, John W.** *Magnetic tape storage and handling: A guide for libraries and archives.* Washington, D.C.: Commission on Preservation and Access, 1995.

30. **Willis, Don.** A hybrid systems approach to preservation of printed materials. Washington, D.C., Commission on Preservation and Access, 1992.

31. Wikipedia, the Free Encyclopaedia, 2006 (http://en.wikipedia.org/wiki/) (last visited on 4th Oct., 2006)

About Author

Dr. Jagdish Arora, Director, Information and Library Network Centre, Ahmedabad.

E-mail : director@inflibnet.ac.in