Workshop on
Digital Libraries: Theory and Practice
March, 2003
DRTC, Bangalore

Paper: J

Digital Information Preservation

Jaba Das

Documentation and Research Training Centre
Indian Statistical Institute
Bangalore-560 059
email: *jabadas@yahoo.co.in*

Abstract

Digital technologies present a preservation solution for the documents in the libraries with increased access to digitized documents over the electronic networks. Digital technology as well as all other associated Internet and Web technologies is in a continuous flux of change. The digital librarian is threatened by "techno obsolescence" and transitory standards. In recent decades, many major libraries and archives have established formal preservation programs for traditional materials which include regular allocation of resources for preservation, preventive measures to arrest deterioration of materials, remedial measures to restore the usability of selected materials, and the incorporation of preservation needs and requirements into overall program planning. This paper represents the challenges of digital preservation and the strategies for solving the preservation problem.

1. INTRODUCTION

Digital technology and high-speed networks are leading to sweeping changes throughout the society. In the past few years, significant progress has been made to define the terms and outline a research agenda for preserving digital information that was either originally digital or transformed to digital from traditional sources. Everyday information is increasing and so is the user expectation. So to save the valuable information and fulfill the user's requirement for posterity, digital documents and resources should be managed and preserved at a tremendous rate.

2. DEFINITIONS

According to a recent statement from the Council on Library and Information Resources "Digital preservation refers to the various methods of keeping digital materials alive into the future."(1) Digital preservation typically centers on the choice of interim storage media, the life expectancy of a digital imaging system, and the expectation to migrate the digital files to future systems while maintaining both the full functionality and the integrity of the original digital system. Digital materials can include everything from electronic publications on CD-ROM to online databases and collections of experimental data in digital format

The goal of digital preservation is to maintain the ability to display, retrieve, and use digital collections in the face of rapidly changing technological and organizational infrastructures and elements. In the digital world, "preservation is the creation of digital products worth maintaining over time." (1)

Each of these words carries weight.

- **IS** - Preservation is a reality and not merely a metaphor for or symbol of access.
- **CREATION** - The time to be concerned about the long-term persistence of digital products is when a system is designed and before digital conversion has begun.
- **PRODUCTS** - A digital product has its own identity and exists within a market economy of end-users
- **WORTH** - The work to design and create a digital product adds value to the information contained in the documents that serve as sources.
- **MAINTAINING** - The persistence of digital products requires careful attention to the maintenance of content (the bits and bytes) functionality (how the bits work in a system).
- **OVER TIME** - Preservation in the digital world is not absolute, but depends instead on the continuing transformative impact of the digital product on the information work of end-users.

3. NEED FOR DIGITAL PRESERVATION

In order to preserve digital materials on a scale commensurate with mass storage capabilities and in formats that are accessible and usable, it is necessary to articulate some basic requirements. These are:

- **Users' perspective:** User's expectation is always changing, yet users, specially research scholars need both traditional documents and electronic documents or old information and current information.
- **Institution's responsibility:** Libraries, archives and other custodians have responsibility for their any properties. So institution should plan for digital materials including their maintenance, preservation, and distribution.

- **Missions of parent institution:** First object of libraries, archives and other custodians is to satisfy the user's expectation and user's requirements. They should preserve all materials in all formats.
- **Storage media:** Storage media is having different formats such as text, data, graphics, video, and sound, different storage capacity like floppy disk, CD-ROM, VCD, etc. and different durability 2years,5 years or 10 years. Handling system of digital materials also is different kinds.
  It is to be noted that Digital storage is not for only print documents, it is also a requirement for oral, cultural and those information are out of print.

4. THE CHALLENGES OF DIGITAL PRESERVATION

Digital preservation raises challenges of a fundamentally different nature which are distinct compared to the problems of preserving traditional format materials. Some of them are in the areas of planning, resource allocation, and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable. The challenges are:

4.1. Longevity Problem

Changing the technological environment, documents, data, records, and informational and cultural artifacts of all kinds are rapidly being converted to digital form. In this context digitized form makes perfect copies of digital materials and it can be published as a web document for remote use. Even it can be reformatted and converted into the alternative formats that can be easily accessible for the end users. Yet the longevity of digital content is problematic for a number of complex and interrelated reasons. Because most of the documents exist only in encoded form, specific software is required for handling them. Even digital component like hardware or software is often changing their versions or processing capacity. Technology used is becoming obsolete and digital documents depend on them. So digital document will be unreadable.

Considering other aspects of this problem, there are many criteria such as administrative, procedural, organizational, and policy issues surrounding the management of digital material. According to technological environment digital documents are different from traditional documents by the way they are generated, captured, transmitted, stored, maintained, accessed, and managed. But longevity of digital information will be reduced without some form of active preservation Nontechnical issues such as management, funding, staffing, maintenance for digital documents, and updated the development of policies for standard techniques and practices to prevent the loss of digital information, also need to be planned.

4.2. Form of Material

Digital materials are very short life comparatively in traditional format. Recording media for digital materials are vulnerable to deterioration and catastrophic loss. Librarian and archivists are trying to preserve acid-based papers, thermo-fax, nitrate film, and other fragile media for decades. In this situation magnetic and optical media are qualitatively different. They are the reusable media.

4.3. Media Problem

There are two types of problems for storage media.

i) Physical lifetime and
ii) Media obsolescence.

Physical lifetime and media obsolescence are correlated to each other. Electronic market makes older storage media obsolete. More insidious and challenging than media deterioration is the problem of obsolescence in retrieval and playback technologies. Innovation in the computer

hardware, storage, and software industries media obsolescence is a very common fact. When greater storage and processing capacities are available in market at lower cost, slowly old product's market will be down. So Devices, processes, and software for recording and storing information are being replaced with new products and methods on a regular three- to five-year cycle. Even those documents are created as a digital form, all are equally vulnerable to technological obsolescence. For example, the short lifetimes media are eight-inch floppy disks, tape cartridges and reels, hard-sectored disks, and seven-track tapes those storage formats are inaccessible and more durable storage media are CD_ROM and optical WORM. The following table shows some estimates made by Ken Bates of DEC (2):

Expected Media Lifetime

Media	Life (years)
9-track tape	1-2 years
8mm tape	5-10 years
4mm tape	10 years
3480 cartridges	15 years
DLT cartridges	20 years
magneto-optical	30 years
WORM	100 years

4.4. Software-dependent Problem

Another problem is that the digital documents are in general dependent on application software to make them accessible and meaningful. But software is also developing and changing versions. Copying media correctly at best ensures that the original bit stream of a digital document will be preserved. But a stream of bits cannot be made self-explanatory. A bit stream can represent anything as a symbol. It is not just text but also data, imagery, audio, video, animated graphics, and any other form or format, current or future, singly or combined in a hypermedia lattice of pointers whose formats themselves may be arbitrarily complex and idiosyncratic. Every software is having different kinds of encoding. So every computer needs some specific software to active the digital documents. A bit stream can be made intelligible only by running the software that created it, or some closely related software that understands it.

4.5. Standard

Another challenge is the absence of established standards, protocols, and proven methods for preserving digital information. Generally, objectives of digital libraries are organizing the information; maintain the intellectual property rights and presentation, retrieval and visualization of digital materials. The main role of digital libraries and archives are the future accessibility of information and preservation of the valuable materials. Digital preservation remains largely experimental and replete with the risks associated with untested methods. Moreover, digital preservation requirements have not been factored into the architecture, resource allocation, or planning for digital libraries.

4.6. Legal and Organizational Issues

Converting digital information also poses the problems in the legal and organizational environment. Intellectual property rights is a big barrier for preserving the digital documents. Digitization of documents are involved with complex method for resolving the legal and practical questions of migrating intellectual property, that includes the creators and owners of intellectual property, managers of digital archives, and actual and potential users of intellectual property. In addition, the parties who represent, different kinds of intellectual property such as text and other

document-like objects, photographs, film, software, multimedia objects, impose their rights in different ways.

5. WHY IS DIGITAL PRESERVATION SO CHALLENGING? DIGITAL PRESERVATION CHALLENGES ARE MAINLY TWO CATEGORIES

5.1. Technical Challenges

- Storage media may have short life, less space, costly, un-available and obsolescent.
- File formats depends on software and specific schemes;
- Integrity of the files, including safeguarding the content, context, references, and provenance.
- Storage capacity and processing devices depend on technological environment which can change both capabilities.
- Distributed retrieval and processing tools, such as embedded Java scripts and applets, may not work with different browser versions.

5.2. Organizational and Administrative Challenges

- Long-term preservation: terms and conditions
- Lack of preservation policies and procedures
- Lack of technical staff and financial resources
- Problem for funding: financial resources
- Collections depend on users and surrounding environment and maintenance
- Lack of good management
- Copyright and fair use for digital collection

6. DEFINING PRESERVATION

At a basic level, digital preservation is preserving the digital medium that holds the digital information by storing it in the correct environment and following agreed storage and handling procedures; copying the digital information into newer, fresher media before the old media deteriorates.

Digital resources can be converted in three ways. These are:

6.1. Preserving Bit Streams through Copying/Refreshing

Digital resources can be stored on any medium that can represent their binary digits. Rothenberg defines a "bit stream" as "an intended meaningful sequence of bits with no intervening spaces, punctuation or formatting". A bit (short for *binary digit*) is the smallest unit of data in a computer. A bit has a single binary value, either 0 or 1. A bit stream is a contiguous sequence of bits, representing a stream of data, that is transmitted continuously over a communications path, serially (one at a time).

However, simply preserving the digital information on several copies of a document or digital medium is not sufficient. It is needed to be sure that the digital information can be retrieved and processed in future.

6.2. Data Interpretation as a Machine Language

The bit stream is read physically– the next step is to be able to interpret it. Interpreting a bit stream depends on understanding its implicit structure which cannot be explicitly represented in the stream. A bit stream that represents a sequence of alphabetic characters may consist of fixed

length bytes each representing a code for a single character. Most files contain information that is only meaningful to the software that created them. For example, Word Processing files embed format instructions describing typography, layout and structure. Spreadsheet files embed formulas relating to their cells etc. This embedded information and all aspects of the representation of a bit stream - including the byte length, character code and structure - comprise the encoding of a file. The files contain both instructions and data that can only be interpreted by the appropriate software.

6.3. Ensuring the Decode Data in Future

A word processing file does not represent a document in its own right. It merely describes a document that comes into existence when the file is interpreted by the program that produced it.
To preserve a digital resource it is needed to ensure that it can be decoded in future. There are two main approaches taken to solving this complex requirement:

a) The conservative approach that is able to fully decode the bit streams held in a file in future is to preserve the program used to create it. It leads to one of two preservation strategies – "technology reservation" or "technology emulation".
b) The optimistic approach that would be able to fully decode the bit streams held in a file in future is to ensure that they are encoded in a format that is independent of the particular hardware and software used to create them. It leads to one preservation strategy – "digital information migration". This approach is discussed below as the part of preservation strategy.

7. DIGITAL PRESERVATION STRATEGIES

There are three potential strategies for ensuring long-term access to digital information.
The three strategies are:

- Technology preservation
- Technology emulation
- Digital information migration

A librarian can select the most appropriate long-term preservation strategy for any library or archive for preserving the digital resources. The three strategies are described in more detail below.

7.1. Technology Preservation

This strategy involves the following criteria:

- Converting the information through machine language as a stable medium,
- Digital medium should be preserved by technology;
- Information can be refreshed and copied as a new media according to the requirements,
- Application programs need to create or access the digital documents
- Preserving the integrity of the digital information during the copying process
- Hardware or system software should support the application software.
- Preserving the computer hardware platform that the operating system software was designed to run on.

This strategy could be adopted where a valuable digital resource is accessed by application software that should have been run on operating system software.

7.2. Technology Emulation

This strategy also has some common criteria with the technology preservation strategy described above.

It involves the following criteria:

- A stable digital medium information should be stored in system
- Digital medium will be preserved when document converted as a machine language
- Data to be represent as a new media format through reconverting or reformatting
- Integrity of digital information will increase through copying process
- Original application propramme should preserve and use to create or access the digital resource.

Migrating or digitizing of digital material depends on both environments such as hardware and application software. If software become obsolete and hence less commercially valuable, copyright restrictions may expire and hence may be made available to future users.
Emulation strategy is used as a short to medium term strategy to maintain the original digital resource. Technology emulation should be used where digital resources cannot be converted into software independent formats and migrated forward. This would usually be due to the complexity of the digital resource and the fact that it was created on a proprietary and obsolete application program.

7.3. Digital Information Migration

The third digital preservation strategy is digital information migration. For this strategy software availability always should be there to decode the current format.
Digital information migration facilitates.

- "Backward compatibility" for application software.
- Application programs interoperability with new product
- Standard formats for converting digital resource independent of both hardware and software

8. DIGITAL INFORMATION STRATEGY

Based on the above discussion it is essential to derive a digital preservation policy for Digital Libraries. Such a policy would help formulate a strategy to tackle digital preservation. To summarise the basic criteria involved:

8.1. Changing Media

This strategy involves printing digital information onto paper or recording it on microfilm. Paper and microfilm are more stable than most digital media and no special hardware or software is needed to retrieve information from them. This may be appropriate in those cases where collection managers are faced with having to preserve hardware and/or software and when digital resource is hardware/software dependent for long time with relatively low budgets. So, changing media should be regarded as a back-up strategy or as a last resort so that digital resource can be accessed as a require formats.

8.2. Backward Compatibility

A second migration strategy relies on making sure popular application software being "backward compatible". The latest versions of most popular word processing packages will be capable of decoding files created on earlier versions of the same package – particularly the previous two or three versions. If the leading application packages are "backward compatible" then migration simply involves testing the process and then loading files created on previous versions. While this strategy may work over the short term for simple digital resources created on some of the leading application packages it cannot be relied upon over the medium to long term or for more complex digital resources.

8.3. Interoperability

The third migration strategy relies on "interoperability" between application programs. Digital resources created on one application program can be exported in a common interchange format and then imported into another application program.

When digital information migrates some problems would involved in interchange process and some valuable data can be lost in this process. Compared to all the data that is lost when digital resources are printed out to paper or microfilm, the data lost during such an "interchange" may be minor. On the other hand – when interchanging the data held in complex databases such as GIS databases and groupware databases – it could involve the loss of thousands of links that have taken years of effort to create and that represent the bulk of the value of the database. The interchange formats themselves may cease to be supported or may be replaced by newer, richer formats.

8.4. Conversion to Standard Formats

The converted digital information should be having standard format that can encode the complexity of structure and form of the original. Digitized information can be accepted as textual documents in several commonly available commercial word processing formats or require that documents conform to standards like SGML (ISO 8879). Databases should follow the standard format so that a file can be stored as a standard file format for example, for bibliographic data MARC 21 may be used as a standard.

9. DEFINING THE DIGITAL RESOURCES

Digital resources which depend on computer and application software. To produce an authoritative classification of all the data types; digital resource categories; application software; data structures and data management systems which are in current use or which have been used at some time in the past would be a lifetime's work.

There are many application programs that still involve the processing of one basic data type and hence the creation of digital resources that comprise just one basic data type. Examples include the huge range of numeric and alphanumeric data processing applications; image processing applications and simple text processing applications.

However, in the increasingly sophisticated world of computing, many of our most popular application programs now involve the processing and creation of multiple data types and the creation of digital resources that comprise multiple data types. Comparison study list is given below. (3)

Applications Used to Create the Digital Resource Categories

	Application Program	Data Type/s Created	Category Of Digital Resource Created	Notes
1	Data Processing; Environmental; scientific; finance; administration	Alphanumeric data	Data sets	Survey data; results of experiments; Transaction data; event data; administrative data; attribute/bibliographic

				data
2	Word Processing	Alphanumeric data; mark-up codes; graphic data; tables	Office documents; structured texts	Simple text documents; reports; literary texts; text for input to publishing systems; text for mark-up;
3	Desktop/ Corporate publishing	Alphanumeric data; mark-up codes; compound documents; tagged graphics; indexes;	Structured texts;	Reports; directories; catalogues; corporate publications; commercial publications
4	HTML Editors	Alphanumeric data; mark-up tags; Web pages	Structured texts	Simple documents; Web pages
5	SGML Editors	Alphanumeric data; mark up codes;	Structured Texts;	Reports; Corporate Publications; Commercial Publications
6	Web page design	Alphanumeric data; mark-up tags; raster /vector graphics	Web pages with graphics	HTML; JPEG & GIF
7	Spread-sheet packages	Alphanumeric data	Data sets/ Office documents	Can be stored in native form for access via spreadsheet or ASCII data can be extracted and held as data set; Data Interchange File (DIF)
8	Business graphics packages	Vector graphics; alphanumeric data;	Flow charts; line drawings;	Can produce graphics for import into compound documents
9	Creative graphics/ clip art	Vector graphics; raster graphics; alphanumeric data	Art work; advertising copy; clip art;	Can produce graphics for import into compound documents

10	Presentation graphics	Vector/raster graphics; alphanumeric data; moving graphics;	Presentations; Courseware; Slides	Adobe Persuasion; MS PowerPoint; Slide manager etc
11	Document Image Processing	Raster graphics;	Office documents; maps; designs; image collections	Used to capture digital images of paper documents; drawings etc. Form of image processing software; With recognition software can capture text from paper documents
12	Image editing/ processing	Raster graphics	Visual Images	Used to capture bitonal, greyscale or colour images of fine art, photographs etc and to manipulate and enhance the images
13	Computer Aided Design (CAD)	Vector graphics; alphanumeric data	Design Data; Mapping Data	Used to create 2 and 3 dimensional designs, models, plans etc
14	Simulation, Modelling & Testing	Vector graphics; raster graphics; animation; alphanumeric data	Ground modelling; flight simulation; 3D visualisation	Used to create 3D models and images and in simulation roles
15	GIS	Vector graphics; raster images; alphanumeric data; links	Mapping data; land cover; population trends	Used to create and manage (see 3.5) links between maps and overlaid data types e g Arc/Info; Arc/View MapInfo etc
16	Speech Processing	Speech coding; Speech synthesis; Speech recognition	Store and playback speech; Computer communication to humans; Create office documents;	Speech recognition used to capture dictated text and load it into word processing packages for editing and to control computer systems

			control computers	
17	Music/ Audio Processing/ Recording	Audio data	Music recordings	Digital recording of live music or existing analogue recorded music; digital composition
18	Digital Video Recording; Proc essing Editing; Generation	Video data; audio data; interleaved audio and video	Video recordings	Digital recording of live video broadcast or existing analogue recorded video; computer generation
19	Animation Processing; Generation	Moving graphics; animation	Presentations; games; entertainment	Manual creation of moving graphics with authoring tools; computer generation of moving graphics

These would need to be preserved if a "technology preservation" or "technology emulation" strategy was adopted. They may also have an impact on the choice of migration strategy if a "digital information migration" strategy is adopted.

10. STRUCTURE OF DIGITAL RESOURCES

The structures that may have been used to store and interchange data for different types of material are enlisted (3).

Structuring Digital Resources

	Category Of Digital Resource	Data Type/s	Proprietary Processable Forms	Standard Processable Forms	Standard Formatted Forms	Notes
1	Data Sets	Alphanumer ic data		ASCII; CSV; Delimited	PDF Postscript	
2	Structured Texts	Alphanumer ic data; mark-up data; tags to graphics;	WP Formats; DTP Formats;	SGML; HTML;	PostScript PDF TeX DSSSL	
3	Office Documents	Alphanumer ic data; raster & vector	WP; Images Spreadsheets; Presentation Graphics;	ASCII; RTF; HTML; SGML; TIFF; CGM;	PostScript PDF; DSSSL	

		graphics; Moving graphics				
4	Design Data	Vector/ raster graphics; alpha-numeric data	CAD formats;WP formats;	DXF/DWG; IGES; CGM; TIFF; ASCII/RTF	HP GL PostScript EPS	
5	Presentation Graphics	Vector/ raster graphics; alpha - numeric data; Moving graphics	Graphics formats; PowerPoint etc;		PostScript PDF	
6	Visual Images	Raster graphics	BMP; PCX;	TIFF; GIF; JPEG;	PostScript PDF	
7	Speech & Sound Recordings	Audio Data	Sun AU (UNIX) MS Wave	MPEG-1 Audio Layers 1/2/3 MIDI		
8	Video Recordings	Video Data	MS AVI; Apple Quick-time	MPEG-1 MPEG-2 MPEG – 4		
9	Geographic/ Mapping Data	Vector graphics; raster graphics; Alpha-numeric data	Arc/Info Arc/View MapInfo AutoCAD Map	TIFF; ASCII CGM	PostScript EPS HPGL	
10	Interactive Multimedia Publications	Audio/ video data moving graphics Raster/ vector graphics; alpha-numeric data	Macro-media; Apple Quick-time;	MPEG-1 MPEG-2		

11. METADATA FOR DIGITAL PRESERVATION

As the archives community are seriously considering using metadata to ensure the integrity and longevity of records, it might be useful to investigate whether a similar approach would be useful

for digital preservation in a digital library context - and in particular for networked documents. Resource discovery metadata like Dublin Core Element Set has the specific aim of supporting in a network environment. It can be used to give basic details about the technical or legal context of a document, but this would need to be extended so that future systems would know exactly how to accurately interpret the document itself, or to migrate the data to a non-obsolete format.

12. CONCLUSION

The long-term digital preservation problem calls for a long-term solution that does not require continual effort or repeated invention of new approaches every time formats, software or hardware paradigms, document types, or record keeping practices change. This approach must be extensible. It must handle current and future documents of unknown type in a uniform way, while being capable of evolving as necessary.

Most approaches that have been suggested as solutions to this problem—including reliance on standards and the migration of digital material into new forms as required—suffer from serious inadequacies. The pros and cons of different method in digital preservation have been discussed. However the emphasis here is to give ample consideration to the issue of digital preservation in planning of Digital Libraries

13. REFERENCES

1. *Handbook For Digital Projects: A Management Tool for Preservation and Access*. from http://www.nedcc.org/digital/II.htm
2. *Preserving Digital Information: Objects in the Digital Landscap.* from http://www.rlg.org/ArchTF/info.html
3. *Moving Theory into Practice: Digital Imaging Tutorial*. from http://www.library.cornell.edu/preservation/tutorial/preservation/preservation-05.html
4. *Digital Preservation and Deep Infrastructure*. from http://www.dlib.org/dlib/february02/granger/02granger.html
5. *Digital preservation: a time bomb for Digital Libraries, Margaret Hedstrom.* from http://www.uky.edu/~kiernan/DL/hedstrom.html
6. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation by Jeff Rothenberg*. from http://www.clir.org/pubs/reports/rothenberg/references.html
7. *Digital Preservation, Ralph Kimball.* from http://www.intelligententerprise.com/000301/webhouse.shtml/database
8. *Comparison of Methods & Costs of Digital Preservation.* from www.ukoln.ac.uk/services/elib/papers/ other/jisc-npo-dig/app1.pdf
9. Day, M. *Metadata for digital preservation: an update.* from http://www.ariadne.ac.uk/issue22/metadata/
10. *Digital Preservation of Moving Image Material?* from http://www.gseis.ucla.edu/~howard/Papers/amia -longevity.html
11. Andrew, K. P. (2000, February). *Digital Preservation: Everything New Is Old Again*, from http://www.infotoday.com/cilmag/feb00/pace.htm
12. *Preserving Digital Objects: Recurrent Needs and Challenges*. from www.lesk.com/mlesk/auspres/aus.html
13. DeHoff, N. M. *Problems with Preservation.* from http://www.geocities.com/Area51/Corridor/5447/termpaper.html