Digital Preservation: Some Preliminary Issues

A. S. Chandel

Abstract:

Digital preservation, not only preserving the digital contents, but alos increases their accessibility accross the globe. With this process the digital divide is getting bridged day by day. However there are certain issues before digital preservation. Inspite of all these, this activity needs to be carried out as users of todya are becoming more dependent on digital resources than printed ones.

Keywords: Digital Preservation; OCR; Hybrid Technology; Digitization.

1. Introduction

Libraries aim at collection, organization, storage, dissemination of library material, and give least attention of preserving them for posterity or for longer duration. Library materials have now extended meaning which includes any physical or non-physical carriers of information. Multiple formats and various media of information have given us various challenges to organize them for access. Maintenance and preservation of such new media are more complex than the traditional ones. Volume of digital publishing is alarming. According to *WorldWideWebsite.com*, the indexed web contains at least 27.85 billion pages as of June 2008 on an estimated 168,408,112 sites; 2.7 million sites were added in May 2008 alone. The figures might have gone much higher by now. In fact digitization was conceived for primarily for publishing and access not for preservation. As such digitization technology is mainly for publishing, though preservation is an important added feature. Preservation is important function irrespective of formats whether printed or digital. Imagine if challenge is thrown to this profession to preserve all these e-resources applying appropriate technology. We should first

think of archiving our rare material of cultural and heritage value which are likely to perish or unusable. No doubt that digitized material has wider use and accessibility and can be opened freely to whole of the world. Users are more concerned with digital resources and they expect that such resources should be preserved for long term if not for posterity. Realizing the importance of digital preservation LC had announced $15 million in awards to eight institutions and their partners to identify and preserve digital material. Professional aspiration is perpetual access and preservation of digital resources. Preservation includes to:

1. maintain in a safety from injury, peril, or harm; protect
2. keep in perfect or unaltered condition; maintain unchanged
3. keep or maintain intact
4. prepare for future use
5. prevent from decaying or spoiling

As per common observation, profession has not been giving due attention to care and preserve even printed material not to speak of digital resources which has now become a necessity and deserve priority. Book material need proper dusting, timely repairs and binding, exposure to sunlight to avoid humidity which most of us don't take it seriously. Neglect of such care has serious repercussions as valuable documents get spoiled beyond repairs if timely action is not taken. Library staff may not even know simple treatment of fungus, silver fish, and termite attacks. It is an accepted fact that professionals are not familiar even with the simple treatment of conservation and preservation of deteriorated materials, though every library needs at least one staff responsible for such jobs specializing in conservation and preservation. We may have a question "what to preserve" all or selective so far as digital resources are concerned. Do we

have appropriate and trusted technology to archive e-resources. But so far as printed materials are concerned, every document acquired in the library need to be preserved which is quite feasible and controllable subject to timely action.

Vasquez de Parga on UNESCO's world Panel on Communication and Information stated that the greatest challenge facing the archive community throughout the world is - the challenge of accepting its full responsibility for ensuring adequate archival processing of these new records and incorporating them fully into their countries' archival systems so that historical memory can be permanently preserved, administrative agencies can function properly and citizens' rights, based on evidential and legal value of such records can be protected. Increase in production of digital documents has further made it more complex and difficult. The process involves huge infrastructure and special professional ability and skills which are not easily available. Storage technology has progressed leaps and bounds and present storage devices have tremendous capacity to archive digital resources. Digitization is a buzz word in the present information society and more and more information materials are appearing in digital format. The society is going to face digital delusion in the near future. Digital divide is getting bridged day by day but their preservation and maintenance are still seeking solutions. Since 1990's onwards, more and more printed material are being converted into digital formats under various projects in the form of microfilm, or any other digital formats. Main advantage of digital resources is that they can be accessed anywhere, anytime. That is why increasing quantity of information is being created and stored digitally. The pertinent question being thought of is regarding the ultimate objective of digitization; *accessibility* or *preservation* or *both*. The birth of digitization technology of course was for wider accessibility

any time anywhere, not the preservation. Digitization was not primarily thought of for preservation, though it was implied that digitized material is to be accessed when it is stored. But long term preservation was not the objective. This accessibility could be temporary or permanent. Nature of digital information is unpredictable; which is accessible today may not be available in the next moment. Therefore, digital archival has to be created and maintained. Many digital resources are not purchased but subscribed for certain period of time. When license or subscription is over, access is denied and there is every likelihood that material could be withdrawn for ever which is as good as lost. Digital resources and its technology are vulnerable to be lost or fail. Whole heritage can be lost with no possibility of their recovery. It cannot be considered as a permanent record/data. No digital medium offers the life expectancy of ink or toner on permanent paper, and the usable life span of most of the digital media are likely to be much shorter than physical lives (Crawfield, 1999). There are various reasons for loss of digital data. Print media has more permanence in its physical form. Deterioration of printed material is quite slow and can be inspected from time to time. The visibility and inspection of physical form of digital resources are not possible as against printed resources. In comparison to printed resources, digital resources are more complex in their characteristics including preservation. Consequently, digital material require specialized knowledge and timely action to keep pace with changing technology. However, everyone thinks of digitization to happen on priority and users expect that such resources should be available and accessible in future also. There are benefits and risks involved in digitization. There is sharp contrast to paper records which, provided they are stored in appropriate conditions, are likely to remain readable for a long time (Sharpe, 2007, p.13). Author

quotes that 'The Domesday Book', William the Conqueror's survey of England in 1086, is still readable today in UK National Archives at Kew, London. A modern digital version of 'Domesday Book' was created by BBC Children's Programme Blue Peter in 1986 to celebrate the 900th anniversary of the original.

All libraries and information centers are aiming to build up their digital collection and has become important part of the collection development. This has multiplied the professional commitment and responsibilities too acquire, convert, process and maintain such resources.

In Indian libraries, collection of digital as well as printed material are to be developed and managed simultaneously. Therefore, competence of library staff to handle both types of resources are required. There are resources which may be called 'born digital' such as photographs, videos, audios. These are also to be processed, maintained and preserved with competence. In order to handle all types of information resources, like microfilm, films of various types, audio cassette tapes, video tapes, computer disks, CD ROM's and DVD's in a single environment is not a easy task to handle. Digital resources are in many formats with their individual features and physical description. It is interesting to note that within five to six years of embarking on chosen course, most of these libraries seemed to be at crossroads with regard to planning their future directions for digital archiving (Gatenby, 2002). Decision about preservation is also crucial which requires decision to select and de-select material for digitization and preservation. Some say that all online publishing should be collected and preserved for future use which may not be possible and practical. Libraries primarily digitize for access not for long term preservation though digitization process is inclusive of preservation. So far as digital archiving is

concerned it is relatively a new discipline though, microfilming and microfiche technology are quite old and still are in use with improved advantages of new technology.

2. Digital Preservation

2.1 Definition

There are quite a good number of definitions of digital preservation given by various authors but most of the people agree that it is a set of activities aimed at preserving digital resources for access in future. It covers:

'Digital preservation combines policies, strategies and actions that ensure access to information in digital formats over time ALCTS (2007)

'Digital preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary (Digital preservation, 2007) 'Digital preservation is series of actions and interventions required to ensure continued and reliable access to authentic digital objects as long as they are deemed to be of value' (JISS, 2007).

Literature review has revealed some of the following challenges of digital preservation:

- Lack of evaluation policy
- Lack of descriptive policy
- Physical vulnerability
- Logical vulnerability
- High technological obsolescence
- High technological dependence
- Lack of trained manpower
- Difficulty in preserving the original look of the document
- Preserving functionality

2.2 Why Digital Preservation

Libraries particularly national libraries and other depository centres have been using various methods of preservation of manuscripts and other such like material were being preserved since centuries. But by the advent of digital technology, professional started rethinking to use new technology for preservation also which has both advantages and disadvantages. Even today, debate continues whether digitization is appropriate technology for preservation. If so, what type of materials deserve priority. It is ideal to preserve all digital resources for future but it is not possible. Selective material which have social, cultural and research value for future should be given priority for preservation. Technology being expensive and time consuming compels for wise selection and choices so that adoption of technology is cost effective. The assumption is that any attempt to define perfect architecture for the system that solves the entire problem once for all is naïve and creates too much risk for the organization that depends on it. (Smith, MacKenzie, 2005). Obsolescence of technology continues to be a problem.

2.3 Objectives

The objectives of digital preservation have been summarized as under:

- Each item in the archive is quality assessed and functional to the fullest extent by current technical capability.
- A gathering schedule can be individually tailored for each selected title taking into account its publication schedule or the frequency with which the website changes, thus enabling the content gathered to be as complete as possible.

- Each item in the archive can be fully catalogued and therefore become part of national bibliography.
- Each item in the archive can be made accessible via the web to readers immediately because permission to do so can be negotiated with the publisher.
- The 'significant properties' of individual resources and classes of resources can be analyzed and determined.
- Sites that are in accessible to harvest robots can be identified and achieved using other methods, by arrangement of other methods with the author (Phillips)

A digital resource which would be rarely used don't deserve to be digitized mainly for access. It is the call of the modern society to have digital environment for easy access in a friendly way, in spite of its serious pitfalls. This technology is not dependable at all. Your whole resource dataset can vanish in seconds accidentally or deliberately. When George Bush took over the presidency on January 20, 2001, he ordered to wipeout the website of Whitehouse. AltaVista reported that 170,000 links were suddenly broken (Wiggins, 2001). Therefore digitized resources need proper care and timely migration from one format to another when the existing formats are becoming obsolete. This also would involve additional finances as well as manpower. If this technology is seriously evaluated, we may find that this technology is primarily more useful to publishing industry. While using digital text, users have to revert back to the print media for proper utility. It provides instant access but has limitation of being used in digital version. Only fact finding information can be made use of in a better way, otherwise long text has to be brought back to print version for careful reading. The question of what to be digitized has to be rationally answered, though all national libraries and institutions of national levels

have the responsibility of preserving some material of cultural, social and economic importance. In this paper, emphasis is being given to preservation of digital resources born digitally or created. It is estimated that more than 90% of new information is being produced digitally. Therefore, it becomes increasingly important to seriously think as to how such resources are stored for future use.

2.4 Role of Digital Preservation

Digital preservation is series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term. The key roles of digital preservation include:

Usability of intellectual contents: There are many factors involved in providing accessibility and usability to archived material. The archived material has to be described in detail so that right information is identified. This calls for creation of metadata following international standards. How these resources are to be catalogued/indexed for effective use requires professional competence and users' approach to such material.

Authenticity of intellectual contents (true replica of original): Before any record is opened for use, it needs to be edited verified ensuring that the original document has been faithfully migrated to archival file. Authenticity and authority should be decided and included in the policy document

Discoverability/identity through authentic metadata: Whatever objects or text is digitized, the same need to be made available online or offline as per requirement. This involves the permission or the copyright to archive for specific period of time or perpetual. Digital environment by its nature is sensitive to changes. Sometime even management may like

to retain the history even being not authentic. We should understand that invisible objects or text are to be made visible. This has to be given due consideration.

Fixity: It also has to be ensured that digitized objects remain protected and well secured from virus or any other threat.

Viability and readability: If hardware and software go out-of-date and are not available, the archived data would not be read. File formats may go obsolete. Hardware and storage media are inherently unstable. There has been always a look for robust archiving solution and scholarly contents in digital form must be ensured to be preserved for future use avoiding all possibilities of their disappearance.

Ingesting the data: While considering digital archiving, most important decision to be taken is choice of material to be digitized. Choices and priorities have to be worked out. Digitization technology is still expensive which all organizations cannot afford. Therefore, selection of material is an important issue. A digital resource which would be rarely used don't deserve to be digitized. It is the call of the modern society to have digital environment due to its convenience and being more user friendly, in spite of its serious pitfalls. We should take a right decision what to archive and with what priority. Since it is still expensive technology therefore, we have to be very selective. If large data is archived, its maintenance and organization may become more complex. Moreover, many material have restrictions of intellectual property right and many other considerations. There are two possibilities of ingesting data. One is already digitally born data, and another is to create. Both would go under different stages and processes. All possible material has to be reviewed thoroughly before final decision is taken. There could be such material or objects which are to be archived for long term,

such as newspapers and old manuscript etc. whose physical characters have deteriorated and are likely to be destroyed. In those cases accessibility, availability may not be top priorities. If newspapers and old manuscript are to be archived, microfilm technology is still the best which has added many new features during the last few years. It was the time when microfilm technology was being considered as out of date. Microfilm technology is the oldest technology used for preservation. The first newspaper microfilmed was *London Evening Times News*, filmed in 1853. That time the life expectancy of the film was about one generation, which has subsequently increased perhaps more than 500 years now. Its technology does not change provided microfilm rolls are properly maintained. and does not need new hardware/software to read even after a century There has been a big revolution in storage technology including microfilm. Progress is still going on which can be well estimated from the capacity of computer memory and other storage devices. The work of Library of Congress has been very significant in preservation and conservation of newspapers under different projects. It had launched this programme during 1980's. The objective of its programme is worth reproducing here which states U S Newspaper *Program is to produce high-quality microfilm surrogates of newspapers that will satisfy the needs and requirements of researchers, archivists, librarians and general public for all reasonably anticipated purpose, including both text and image material for general and scholarly research, as well as the use of newspaper images for display or exhibition purposes. Every effort shall be made to record in the new format all information contained in and used with the original newspaper, and to maintain the integrity and authenticity of representation of the original document on microfilm.* In such cases ingesting or capturing data for archiving is not a big deal. Such decision

can be taken quite easily. Therefore, decision has to be taken what is going to be digitized primarily for long term preservation and immediate access. Some material is digitized with a purpose that the stored material would not be frequently used. The use would be for a very selected and dedicated researchers.

Structuring of records: Data structuring is an intellectual process which needs conceptual clarity to organize data which is being archived. There could be different objectives to store records. For example DSpace using Dublin Core for record structuring and retrieval is commonly used. Preparing a document for access is most complex function which only very competent professionals would be able to handle. Another crucial decision to be taken is as to how text, visuals or any other form of digital records is to be indexed for access and in which format. Ultimately archived information is to be retrieved from the storage devices.

Need for Integrated System: Preservation and access functions should go together hand in hand compatible to each other.

The steps in migration of print material to digital format includes:

- Image preparation of documents to be scanned
- Scanning of pages
- Editing of pages
- Using OCR to edit text reads the text from the image of a document and converts it into ASCII text
- Creating a searchable database
- Liking text to image
- Mounting the product on web

It has to be ensured that images are of high quality that are readable and printable

3. New Hope from Hybrid Technology

By the advent of CD-ROM, it was being thought of that microfilm technology would be replaced by such other technology like CD-ROM which has a tremendous capacity of storage. But microfilm media stayed and confirmed its permanent place for storage with a little of bit inconvenience of retrieving text as is possible from the digital sources. Now microfilm technology enables to convert microfilm into digital from for accessing data as from other digital formats. It uses microfilm for storage and computer for accessing data. This is the ideal solution for preserving such resources which are likely to be accessed rarely.

4. Conclusion

As the technology is new, most of us are not much familiar with the infrastructure required including hardware and software and lack practical experience. We may need technical know-how and consultancy to begin with. Most important initiative would be to prepare a strategic policy framework and model for creating, maintaining and preserving digital collection for access and preservation both. We should always be prepared to fight the modes of digital deaths of preserved material and should go on refreshing and migrating stored information to a new format taking into consideration of media/ format obsolescence. We cannot wait and delay, have to take initiative to take off. We have to be prepared to face the challenges of present digital world we are living in. In the present knowledge society, initiatives are to provide access to world repository. This access should be provided today, tomorrow, and ever after which is only possible provided information is preserved. Librarians were also called custodian of knowledge which we may not like this title bu this function is attached to the profession and we have t

perform our duty as a custodian by preserving everything we collect and preserve for posterity providing access to the world's knowledge for those who are in search of the treasure. Care must be taken that file format and software don't become obsolete and readability is always maintained. Information users now depend more on digital resources than printed ones, therefore need for preservation of such sources is increasingly more.

Reference

1. ALCTS. Preservation and reformatting selection Committee (PARS) Digital Preservation Group blog, http://blog.ala.org/digipres.php
2. Crawfield, Walt (1999) Bits is bits pitfalls in digital reformatting. *American Libraries*, May, p.47.
3. Digital preservation Coalition: **website: http://www.dpconline.org/graphic/intro/definitions/html**
4. Gatenby, P. (2002). Report on the senior executive fellowship to research digital archiving in national libraries. In what should be preserved.../Margaret E Phillips
5. JISC(2006)Digital preservation briefing paper; www.jisc.ac.uk/publications/publications/pub_digipreservationbp.aspx
6. Phillips, Margaret, E. (2005) What should we preserve? The question for heritage libraries in a digital world. *Library Trends*, 54(1), 61-62
7. Sharpe, Robert (2007) Solving archive challenges. *Research Information*, June/July, 13
8. Smith, MacKenzie (2005) Exploring variety in digital collections and the implication for digital preservation. *Library Trends*, 54(1), 6

9. Thomaz, Katia (2006) Critical factors for digital records preservation. *Journal of Information Technology, and organizations*, **1**