Digital Preservation: A Software Approach

R K Joteen Singh

Abstract

In today's ever developing and rapid growing world, the national heritage site- the pride and identity of a country, seems to be in deteriorating state though various preservative measures are being taken up. What is important here is to understand which preservative method and technique will serve us best in this technological and computerized era. Since there is a continuous technological growth and advancement, the technological component does also undergo transformation on both software and hardware levels. One of the suitable preservation techniques available is the digitization using Xena software. This paper attempts to show how the particular software will help in preserving the national heritage in long run.

Keywords: Digital Preservation, Preservation Strategies, Preservation Software, Xena Software, Open Archival Information System

1. Introduction

The national heritage is the pride and identity of the nation and its people. Heritage that gives an identity to the people needs to be preserved in order to pass down to generations to come. Various elements of national significance like the architecture, landscapes, documents and other artifacts should be preserved using advanced scientific technology and methods to save them from decay and disappearance. India among other countries is known for its rich heritage and culture in the world. It has been one of the main contributors in the field of medicine, mathematics, science, technology, philosophy, theology, literature, linguistics, graphic arts, music, dance and many other disciplines. The means of documentation available to our pioneers was to record their ideas and thoughts by writing on the palm-leaf. One of the main forms of Indian traditional preservation method is maintaining a palm-leaf book. The amount of books on all kinds of disciplines that India has contributed using this method is remarkable. However, these fascinating heritage including paper, palm-leaves and birch bark are organic materials that have their own limited life span. German scientists have determined that most of the Indian palm-leaf books will naturally decay within the next 50 to 100 years. It implies that may be after 100 to 200 years all these kinds of books will disappear. And this will be a lost to human civilization. With the development it has also undergone a great transformation through ages now with the advancement in our science and technology. In this computerized era it is apt to utilize the technology in preserving such heritage of a country and as many scientists have suggested, the best, effective and advantageous way is digitizing the materials, and the national heritage sites to save it from being lost. Using sophisticated equipments and computers, it has produced ultra sharp digital

colour pictures of the Indian national heritage like palm-leaf books, paper manuscripts, books, newspapers, letters, birch-bark texts, drawings, paintings, sculptures, inscriptions and many other heritage objects though the originality and the essence of the heritage item may not be maintained up to the point. Now the concern here is how far these digitized materials will serve our purpose? Is it free from limitation? The answer, perhaps, is no, we are not safe from its loss. The technology is changing very fast and the present day information format may not be accessible after a few decades. Keeping this in mind, the main objective of this paper is to attempt and show the feasibility of the application of Xena software in long-term preservation of digitized objects.

2. Digital Preservation

Digital preservation is the process of maintaining accessibility to information and all kinds of records including scientific and cultural heritage existing in digital formats. Many of the national heritage including rare books have digitized which promotes accessibility to the organizations or even individuals irrespective of the location of the national heritage.

3. Necessity Of Digital Preservation

Most of the media which digital information is recorded have less life span than some analog media such as paper. While acid paper is prone to deterioration however, the rate of deterioration is quite slow. It is also possible to retrieve the information without loss once deterioration is noticed. In case of digital environment, the digital data recording media deteriorate more rapidly and once the deterioration is detected, it is the matter of chance to retrieve even a piece of information from the particular media. In addition to the above, the digital technology is developing in its full speed and retrieval and coding technologies can become obsolete within very short period of time. When more performance and less expensive storage devices are developed, older versions may abruptly replace. Even a software decoding technology is abandoned or a particular hardware may no longer be in production, records created with such technologies are at a great risk, because they are no longer accessible which is known as digital obsolescence. In this scenario, preservation of digital information is intense required to have focus at right time than preservation of other media [1].

4. Common Preservation Strategies

Regarding the long-term preservation of digital objects the Online Computer Library Centre (OCLC) has developed a set of strategy, which consists of:

- ❖ Assessing the risks for loss of content posed by technology variables such as commonly used proprietary file formats and software applications
- ❖ Evaluating the digital content objects to determine what type and degree of format conversion or other preservation actions should be applied
- ❖ Determining the appropriate metadata needed for each object type and how it is associated with the objects
- ❖ Providing access to the content [2].

Some other strategies which are commonly use by individuals and organizations may be highlighted as refreshing, migration, replication, emulation, etc. All these long-term preservation strategies are mutually important however, digital obsolescence

is the main issue, simply because lack of established standards, protocols and methods for preserving digital information. Therefore, standardization of digital file format is again a basic requirement for long-preservation of digital objects.

5. Digital Preservation Standards

To standardize digital preservation practice and produce a set of recommendations for preservation programme implementation, the Reference Model for an Open Archival Information System (OAIS) was developed. The reference model includes the following responsibilities that an OAIS archive must abide by:

❖ Negotiate for and accept appropriate information from information producers

❖ Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation

❖ Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and therefore, should be able to understand the information provided

❖ Ensure that the information to be preserved is independently understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information

❖ Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies and which enable the information to be disseminated as authenticated copies of the original or as traceable to the original.

❖ Make the preserved information available to the Designated Community [3]

6. Digital Preservation Software

❖ The Digital Preservation Software Platform (DPSP) is free and open source software developed by the National Archives of Australia. The DPSP is a set of software applications which promote the process of digital preservation. There are four components of DPSP such as:

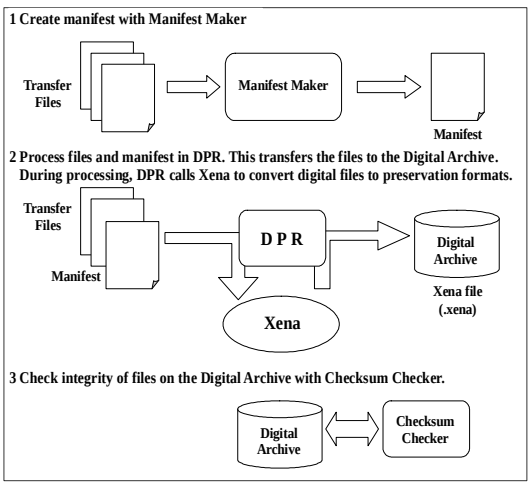
❖ **Xena:** Xena stands for XML Electronic Normalising for Archives. Xena converts digital files to standards based, open formats.

❖ **Digital Preservation Recorder (DPR)**: DPR handles bulk preservation of digital files via an automated workflow.

❖ **Checksum Checker:** Checksum Checker is a piece of software that is used to monitor the contents of a digital archive for data loss or corruption.

❖ **Manifest Maker:** Manifest Maker produces a tab-separated list of digital files in a specified location. The manifest includes the checksum, path and filename of each digital file.

The digital preservation process is described in the following diagram:



The new features of the latest version of Xena includes:

- ability to normalise harvested websites;
- integration with Tesseract OCR and the ability to create raw text versions of file formats (such as Word, TIFF and PDF);
- support for audio files in OGG container format using Vorbis, FLAC or Speex codecs;
- improved MP3 guesser;
- support for more image formats (such as CUR, PCX and XPM);
- new character set detection library;
- automatic configuration of Xena output and log directories;
- ability to preserve directory structures;
- ability to handle files normalised with previous versions of Xena;
- major refactoring of the source code for the external libraries used by Xena and an update of license to GPL version 3;
- creation of an automated installer for Microsoft Windows and MAC OS X versions of Xena

6.1 Supported Formats

During the process of normalisation, Xena will convert the following file types to the specified open format [4].

Archives and Compressed Files

GZIP Files are extracted from the archive and normalised into separate Xena files.

JAR	Files are extracted from the archive and normalised into separate Xena files. A Xena index file is created, which when opened in a Xena Viewer, will display the files in a table.
MAC BINARY	Files are extracted from the archive and normalised into separate Xena files.
TAR	Files are extracted from the archive and normalised into separate Xena files. A Xena index file is created, which when opened in a Xena Viewer, will display the files in a table.
TAR.GZ	Works as a combination of 'GZIP' and 'TAR'. All files are extracted from the archive and normalised into separate Xena files.
WAR	Files are extracted from the archive and normalised into separate Xena files. A Xena index file is created, which when opened in a Xena viewer, will display the files in a table.
ZIP	Files are extracted from the archive and normalised into separate Xena files. A Xena index file is created, which when opened in a Xena viewer, will display the files in a table.

Audio

AIFF	Audio Interchange File Format files are converted to FLAC.
FLAC	Free Lossless Audio Codec files are preserved and wrapped in XML.
MP3	MPEG-1 Audio Layer 3 files are converted to FLAC.

OGG	OGG container format files are converted to FLAC.
WAV	Waveform Audio Files are converted to FLAC.

Databases

SQL	Structured Query Language files are preserved and wrapped in XML.

Documents

CSV/TSV	Comma and Tab Separated Values-based files are stored as a special case of plain text.
DOC/PPS /PPT/XLS	Microsoft Office documents are converted to the Open Document Format.
DOCX/PPTX/XLSX	Microsoft Office Open XML documents are converted to the Open Document Format.
HTML	Hypertext Markup Language files are converted to XHTML.
MPP	Microsoft Project documents are converted to XML.
ODS/ODP ODT	Open Document files are preserved / and wrapped in XML.
RTF	Rich Text Format is converted to Open Document Format.
SYLK	This spreadsheet format is converted to Open Document Format.
SXC/SXI/ SXW	StarOffice formats are open, but are converted to the newer Open Document Format.
TXT	Text files are preserved and wrapped in XML.
WPD	Word Perfect files are converted to Open Document Format.
XHTML	Extensible Hypertext Markup Language files are preserved and wrapped in XML.
XML	Extensible Markup Language files are preserved and wrapped in XML.

Email

MBX/ MBOX	Mailboxes are converted to individual XML files and a Xena index file is created which will display the files in a table when opened with Xena Viewer.
PST	Mailboxes from Microsoft Outlook are converted to individual XML files and a Xena index file is created which will display the files in a table when opened with Xena Viewer.
TRIM	Messages from TRIM are converted to XML and a Xena index file is created Mailboxes are converted to individual which will display the files in a table when opened with Xena Viewer.

Graphics

BMP	Bitmap image files are converted to PNG.
CUR	Windows cursor files are converted to PNG.
GIF	Graphics Interchange Format image files are converted to PNG.
JPEG	JPEG image files are preserved and wrapped in XML.
ODG	Open Office Document Drawings are preserved and wrapped in Xena XML.

PCX	Personal Computer eXchange image files are converted to PNG.
PDF	Portable Document Format files are preserved and wrapped in XML.
PNG	Portable Network Graphics are preserved and wrapped in XML.
PNM	Portable Anymap graphic bitmap files are converted to PNG.
PSD	Photoshop image files are converted to PNG.
RAS	Sun raster graphics are converted to PNG.
SVG	Scalable Vector Graphics are preserved and wrapped in XML.
TIFF	Tagged Image File Format image files are converted to PNG. Embedded metadata is preserved in Xena XML.
XBM	X11 Bitmap Graphics are converted to PNG.
XPM	Unix Icon files are converted to PNG.

7. Conclusion

Digital obsolescence is exacerbated by the lack of established standards, protocols and methods for preserving digital information. Such problems can be minimized if open formats based on open standards is used for preserving digital information. Xena digital preservation software can converts files into an openly specified format or else it performs ASCII Base 64 encoding on binary files and wraps the output with XML metadata headers and footers. The resulting .xena file is plain text, although the content of the data itself is not directly human-readable. The exact original file can be retrieved by stripping the metadata and reversing the Base 64 encoding, using an internal viewer which includes an export function. This way Xena software can be used for long term digital preservation up to some extent.

Acknowledgements

The author thanks Ms. Ningthoujam Somola Devi for her generous help and support.

References

1. McLeod, R. Wheatley, P. and Ayris, P. Lifecycle information for e-literature: full report from the LIFE project. http://eprints.ucl.ac.uk/ (accessed on 23/03/2010)
2. Online Computer Library Center, OCLC Digital Archive Preservation Policy and Supporting Documentation. http://www.oclc.org/ (accessed on 08/06/08/2011)
3. Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS). http://public.ccsds.org/ (accessed on 14/04/08/2011)
4. National Archives of Australia, http://xena.sourceforge.net/ (accessed on 01/05/2011).

About Author

Dr. R K Joteen Singh, Information Scientist, Manipur University Library.
E-mail: joteenrk@yahoo.com