DIGITAL ARCHIVE

Introduction:

The vast amounts of information produced in the world are now for a large part digital. The task of managing the ever-increasing digital objects throughout their life cycle and to preserve them in perpetuity becomes more and more complex because of they are fragile, volatile and ephemeral in nature. Their viability depends on technologies that are rapidly and continuously changing. Furthermore, as newer digital technologies rapidly appear and older ones are discontinued, information that relies on obsolete technologies soon becomes inaccessible.

Definition:

Digital preservation is defined as the managed activities necessary:

a) For the long term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and

b) For the continued accessibility of the document contents through time and changing technology.

The vast amounts of information produced in the world are now for a large part digital and include a wide variety of materials: text, databases, audio, film, images. They range from medical records to movie DVDs, from satellite surveillance data to websites presenting multimedia art, from data on consumer behaviour collected by supermarket to a scientific database

documenting the human genome, from news group archives to museum catalogues. The problem on preserving digital records for long term access requires careful consideration about process and technology.

Until today there has been no storage platform that can be trusted to store critical electronic record for long time. Preserving digital information is more difficult than preserving record on materials such as paper or film. The sheer volume and the volatility introduced by digital, demand new software architecture capable of scaling and of preventing accidental changes to the records.

Procedures need to be put in place to identity, classify, move, evolve, access and occasionally dispose of digital records. Library and Information science and traditional archival practice provide an extensive body of knowledge that can be leveraged with technology create a true modern archive.

Architecture for Digital archiving and preservation:

Of late, the framework used for archiving and preservation is based on the Open Archival Information System Reference Model (OAISRM). The OAIS RM is used by most major preservation projects including those in Australia, the United Kingdom, the Netherlands, and the United States.

Ingest: Acquisition and collection development

The first function to be performed by the archive itself is acquisition, or ingest. This is the stage at which the created object is “incorporated” physically or virtually into the archive. The

acquisition of electronic information for archiving involves the development of collection policies and gathering procedures.

Production and creation of electronic information

Information that is born digital may be lost if the producer is unaware of the importance of preservation, and practices used when electronic information is produced will impact the ease with
which the information can be digitally archived and preserved. The archiving and preservation process is more efficient when attention is paid to issues of consistency, format, standardization and metadata description before the material is considered for archiving.

Metadata for preservation

Archiving and preservation require special metadata elements to track the lineage of a digital object (where it came from and how it has changed over time), to detail its physical characteristics, and to document its behavior in order to reproduce it on future technologies.

Formats for preservation

Without a through understanding of the internal details of the formats in which digital objects are encoded, the log term preservation of the objects is not feasible. Specific instances of formatted objects must also be interpretable so that the significant properties of those objects can be retrieved. Most electronic journals, reference books, or reports use TIFF image files, PDF, or HTML.

Preservation planning: Migration and emulation

Two strategies for preservation are migration and emulation. *Migration* means copying the object to
be archived and moving it to newer hardware and software as the technology changes. It is the process of transferring data from a platform that is in danger of becoming obsolete to a current platform Migration is, of course, a more viable option if the organization is dealing with well-established commercial software such as Oracle or Microsoft Word.
Emulation requires software to be developed that can simulate the original experience using the original file format but with current technologies. The essential idea behind emulation is to be able
to access or run original data/software on a new/current platform by running software on the new/current platform that emulates the original platform.

Access: Current and Future

The way in which access is viewed depends on the purpose of the archive, the audiences it will serve and the anticipated needs of those audiences over the long term. For example, national and institutional archives must be concerned with the ability to provide long-term access to the electronic information in a way that virtually replicates the look and behavior of the object today. This is a requirement because of the legal functions served by these archives of record.

Systems Development: Present Scenario

Various initiatives have been taken and various systems have been developed all over the world of digital preservation and archiving, such as:

DIAS

In 2003, the National Library of the Netherlands started a joint project with IBM to develop the preservation subsystem of DIAS, called Preservation Manager. The subsystem will consist of a preservation manager, a preservation processor and tool(s) for permanent access. The Preservation Manager will manage and control the long-term durability of the digital objects using technical metadata.

DIAS' research was extended through a new project called KOPAL, which began in October 2004 with the German national library.

CAMILEON

The CAMILEON Project (Creative Archiving at Michigan & Leeds: Emulating the Old on the New) is developing and evaluating a range of technical strategies for the long term preservation of digital materials.

The project is a joint undertaking between the
Universities of Michigan (USA) and Leeds (UK).

PADI

The National Library of Australia's Preserving Access to Digital Information (PADI) initiative aims to provide mechanisms that will help to ensure that information in digital form is managed with appropriate consideration for preservation and future access. Its objectives are: to facilitate the development of strategies and guidelines for the preservation of access to digital information; to develop and maintain a web site for information and promotion purposes and to provide a forum for cross-sectoral cooperation on activities promoting the preservation of access to digital information.

OCLC Digital Archive

OCLC's Digital Archive offers real-world solutions for the challenges of archiving and preservation in the virtual world. This flexible system allows archiving assets in two ways. (a) Web archiving and (b) Batch archiving. No matter how a document is being submitted to the archive, that can be made available to users in multiple ways - through FirstSearch, Connexion, through OPAC or a Web portal.

PANDAS

The PANDORA Digital Archiving System, known as PANDAS, was developed in-house following an unsuccessful attempt to find an off-the-shelf system (or systems) to provide an integrated, web-based web archiving management system. The need for such a system was evident as the scale of the Library's archiving activity increased and if the best possible efficiencies were to be achieved in building a collaborative, selective and quality assessed web archive.

NDIIPP

The National Digital Information Infrastructure and Preservation Program of the Library of Congress, seeks non-binding expressions of interest for collaborative projects that model innovative solutions for the preservation of commercial digital content. The Library is interested

in digital content intended for distribution through commercial channels, specifically moving images (film, television), digital photography and other forms of pictorial art, multimedia and literary arts, recorded sound, and video and computer games.

FEDORA

Flexible Extensible Digital Object Repository Architecture was developed jointly by the University of Virginia Library and Cornell University's digital Library Group in May, 2003. Fedora open source software gives organizations a flexible service-oriented architecture for managing and delivering
their digital content. At its core is a powerful digital object model that supports multiple views of each digital object and the relationships among digital objects.

LOCKSS

Lots of Copies Keep Stuff Safe is an automated, decentralized preservation system developed by
Stanford University to protect libraries against loss of accesses to digital materials. The evolution of the Web has disrupted this critical library role. Libraries have not had an easy way to build digital collections, nor had any assurance that a digital collection — once obtained — would remain accessible to future generations. Publishers are being asked to assure persistent access to content. The LOCKSS Program addresses these issues. It is a open source, peer-to-peer software that functions as a persistent access preservation system.

PMC

Portable PubMed Central is a digital archive of life sciences journal literature at the U.S.National Library of Medicine (NLM). Participation by publishers in PMC is voluntary, although participating journals must meet certain editorial and technical standards. PMC, itself, is not a publisher. Access to PMC is free and unrestricted. PubMed Central was developed and is operated by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the U.S. National Institutes of Health (NIH). PMC is an electronic archive of full-text journal articles,
offering free access to its contents.

GREENSTONE

Greenstone is a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato, and developed and distributed in cooperation with UNESCO and the Human Info NGO. It is open-source, multilingual software. Greenstone runs on all versions of Windows, and Unix, and Mac OS-X. It is very easy to install.

The EPrints

This software is a free, open source product that creates online archives and is produced by the University of Southampton. EPrints is configured by default to create online archives of research papers, but can be configured to archive various types of documents. The EPrints archive is currently in use in many libraries and centers throughout the world.

DSpace

DSpace is a groundbreaking digital repository system that captures, stores, indexes, preserves, and redistributes an organization's research data. Jointly developed by MIT Libraries and Hewlett-Packard Labs, the DSpace software platform serves a variety of digital archiving needs. Research institutions worldwide use DSpace to meet a variety of digital archiving needs, such as:

- Institutional Repositories (IRs)
- Learning Object Repositories (LORs)
- Theses
- Electronic Records Management (ERM)
- Digital Preservation
- Publishing and more.

DSpace accepts all forms of digital materials including text, images, video, and audio files. Possible content includes articles and preprints, technical reports, working papers, audio files, video files etc.

Indian Scenario

Ministry of Human Resources Development, Govt. of India has advised all the consortium members of INDEST to set up e-print archives using appropriate OAI(Open Archives Initiative) compliant e-print software. MHRD also recommended that a central server may be deployed to harvest metadata from all such epirnt archives. Again INFLIBNET, the inter University Centre of UGC under Ministry of HRD has initiated Institutional Repository and archive its publications, proceedings etc using DSpace. INFLIBNET's Institutional repository and dArchive-INDIA is an online electronic repository
especially created for Indian Academia by INFLIBNET Centre (UGC).
Some others major initiatives in this context are: Indian Institute of Science (NCSI); Search Digital
Library (SDL) at DRTC Bangalore; Nalanda Digital Library, NIT, Calicut; IIT Kharagpur, IIM Kozhikode, Million Book Universal Digital Library Project, Indira Gandhi Centre for the Arts etc.

Conclusion

Archiving and related issues of digital preservation are becoming ever more significant within the scientific and scholarly communication chain. Recently, several approaches for digital preservation have been identified and presented. Conventional methods are mainly technology emulation, information migration, and encapsulation. However, there is a lack of proven preservation methods to ensure long term safe preservation for digital objects.There is a burning question: what is appropriate material for preservation and what can be 'edited out'. The issue of the copyright of intellectual and intangible properties is also a problem towards digital preservation.