DIGITAL ARCHIVING AND PRESERVATION: PRESENT SCENARIO

Biswajit Saha

Abstract

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 contiouusly changing. Furthermore, as newer digital technologies rapidly appear and older ones are discontinued, information that relies on obsolete technologies soon becomes inaccessible. This paper addresses the present scenario, framework, formats, planning and the systems and projects developed all over world for preservation, the metadata used to manage them. It also discusses the practical considerations necessary to create a digital archive. This article would help Library & Information Professionals (LIP) to navigate this complex new field.

Keywords: Digital object; Metadata; Digital preservation; OAIS

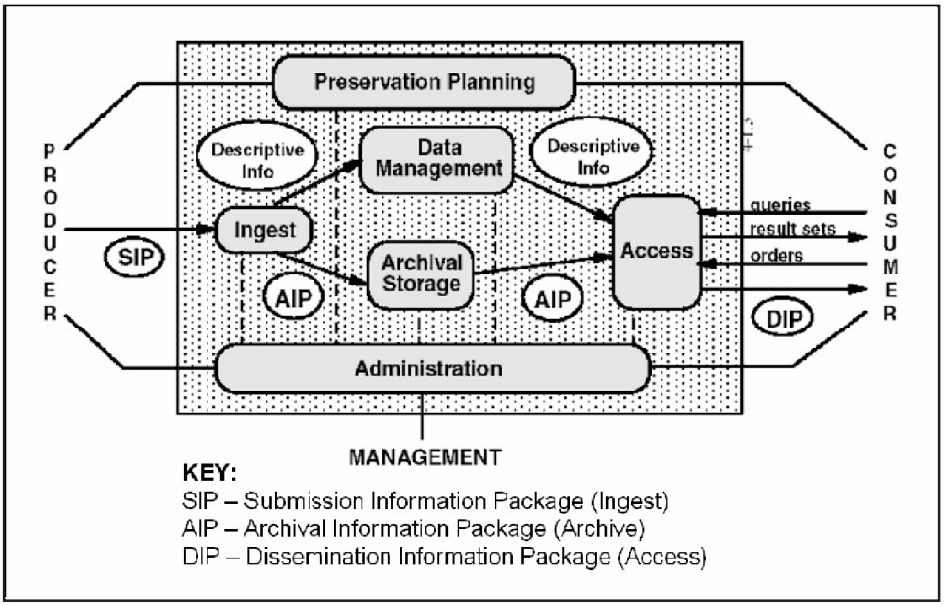
1. Introduction

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 tills 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.

2. 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. The OAIS RM is used by most major preservation projects including those in Australia, the United Kingdom, the Netherlands, and the United States.

Fig.1



The OAIS Reference Model for digital preservation

2.1 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, and these policies and procedures should be considered in tandem with the development of archiving system requirements.

2.2 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. Limiting the variability of the incoming

material is easier for a small institution or a single company to enforce than for a national archive or library, where a variety of formats must be ingested, managed and preserved. In the case of more formally published materials, such as electronic journals, efforts are underway to determine standards that will facilitate archiving, long-term preservation and permanent access. Such Standardization is considered key to efficient archiving and preservation of electronic journals by third-party archives. Creators can also create metadata at the producer stage, where an expert can support the description of the technical content. With recent incorporation of XML and other architectures into software applications, such as MS Word, and PDF, the creation of metadata by creators should become easier and more automatic.

2.3 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. In 2001–2002, the Preservation Metadata Working Group developed a draft set of over 20 elements and numerous sub-elements for metadata preservation in the framework of the OAIS Reference Model. In order to gain consensus on this set and to provide operational and implementation guidance, a follow on group, PREMIS, the Preservation Metadata: Implementation Strategies working group was formed. The draft element set for preservation metadata and the results of the implementation survey were published in 2004–2005. The plan is to provide the preservation metadata set for testing and prototype implementations before moving the results into a standards process.

2.4 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. TIFF is the most prevalent for those organizations that are involved with conversion of paper issues of journals. For purely electronic documents, Adobe's PDF (Portable Document Format) is the most prevalent format. PDF provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. While PDF is increasingly accepted, concerns remain for long-term preservation and it may not be accepted as a legal depository format, because it is a proprietary format. Therefore, Adobe, the Association for Information and Image Management (AIIM) and several other organizations have developed a draft standard for archival PDF, called PDF-A. This provides a file specification for a minimal set of PDF features and functions that will continue to be migrated from one version of PDF to another. The draft is currently in the ISO process.

2.5 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. However, even in these cases migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features.

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.

2.6 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.

3. 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:

3.1 DIAS

In 2003 the Koninklijke Bibliotheek (KB), the National Library of the Netherlands started a joint project with IBM to develop the preservation subsystem of DIAS, called Preservation Manager. The work began with a series of studies around key preservation issues such as authenticity, media migration management, archiving of web publications, and a proof of concept of the Universal Virtual Computer. This 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. This is considered to be an essential part of the DIAS solution, since technical metadata will allow a future hardware environment to take the software bit stream and the content bit stream and provide access to the content.DIAS' research is being extended through a new project called KOPAL, which began in October 2004 with the German national library, Die Deutsche Bibliothek

3.2 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. User evaluation studies and a preservation cost analysis are providing answers as to when and where these strategies will be used. The project is a joint undertaking between the Universities of Michigan (USA) and Leeds (UK) and is funded by JISC and NSF.

3.3 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. The PADI web site is a subject gateway to digital preservation resources. It has an associated discussion list padi-forum-l for the exchange of news and ideas about digital preservation issues.

3.4 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. FirstSearch is an online service that gives library professionals and end users access to a rich collection of reference databases. With FirstSearch, materials in a library's collection are highlighted in results from searches in dozens of leading databases. Connexion is OCLC's flagship cataloging service, a powerful, flexible suite of tools with built-in access to WorldCat, the world's largest bibliographic database.

3.4.1 Web archiving: Item-by-item

Introduced in September 2002, the Web Harvester for OCLC's Digital Archive allows for item-by-item archiving of Web pages and Web documents, notably HTML, PDF and associated files that are often "born digital." Simply create a Digital Archive record in Connexion, and then send the Web Harvester out to extract the components one want to place into the archive. After the content is ingested, one can manage it through the Administration Module, and users can access it through Connexion, FirstSearch, your OPAC or any Web links you create.

3.4.2 Batch archiving: For collections

We can submit our collected assets in TIFF and other formats for inclusion in the Digital Archive. We don't even have to use Connexion. Just we have to run a simple program supplied by OCLC and send digital collection(s) and basic metadata on CD-ROM or tape. OCLC ingests our collections, automatically generates the metadata records and notifies us when our collections have been archived for administration and access. The OCLC Digital Archive provides a reliable, standards-based solution for the life cycle and long-term management of digital collections.

From TIFFs and GIFs to PDFs and HTML pages, OCLC's Digital Archive can manage multiple file formats, and features an intuitive interface that makes it easy to harvest, organize, and archive digital assets. Based on the OAIS (Reference Model for an Open Archival Information System) ISO standards and utilizing the METS (Metadata Encoding and Transmission Standard), the Digital Archive brings context, accessibility and longevity to the fast-shifting and often ephemeral world of digital assets.

3.5 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. PANDAS was first implemented in June 2001 and a second much enhanced version was released in August 2002. The second version was more modular in design to facilitate future enhancement by allowing development to component parts of the system. Consequently the current development program includes a number of incremental upgrades - as at the end of June 2004 version 2.1.5 is in production and version 2.2 is under development - as well as concurrent development of what for working purposes is characterized as PANDAS version 3 and which will include more systemic features.

3.6 NDIIPP

The National Digital Information Infrastructure and Preservation Program of the Library of Congress: The Library of Congress, through NDIIPP, is seeking expressions of interest in a project to preserve the digital content produced by the private sector. Expressions of interest are due Sept. 22, 2006. 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. Project participants may include content creators, service providers, distributors, technology firms, associations that represent such entities and cultural institutions that are entrusted with preserving content.

3.7 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. Digital objects can encapsulate locally-managed content or make reference to remote content. Dynamic views are possible by associating web services with objects. Digital objects exist within a repository architecture that supports a variety of management functions. All functions of Fedora, both at the object and repository level, are exposed as web services.

Preservation worthy: Fedora repositories incorporate a number of features that facilitate the complex tasks associated with digital preservation. Internally all Fedora digital objects are represented in the file system as files in an open XML format. These XML files include data and metadata for the objects plus relationships to services and other objects. The entire structure of a Fedora repository can be rebuilt from the information in these files. In addition, Fedora repositories are compliant with the Reference Model for an Open Archival Information System (OAIS) due to their ability to ingest and disseminate Submission Information Packages (SIPS) and Dissemination Information Packages (DIPS) in standard container formats such as METS and MPEG-DIDL.

This unique combination of features makes Fedora an attractive solution in a variety of domains. Some examples of applications that are built upon Fedora include library collections management, multimedia authoring systems, archival repositories, institutional repositories, and digital libraries for education.

3.8 LOCKSS

Lots of Copies Keep Stuff Safe is an automated, decentralized presrvation system develpoed 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 — a function well outside of their core mission. The LOCKSS Program addresses these issues. It is a open source, peer-to-peer software that functions as a persistent access preservation system.

OCLC joins the LOCKSS Alliance in support of its collaborative effort to explore new uses of the LOCKSS technology to benefit the community and to build new capabilities for digital preservation. OCLC will work collaboratively with LOCKSS to explore the expansion of the LOCKSS technology to operate with different types of digital content.

The Library of Congress NDIIPP has entered into a three-year cooperative agreement with Stanford University in support of the CLOCKSS digital archive pilot and related technical projects. CLOCKSS (Controlled LOCKSS—Lots of Copies Keep Stuff Safe), a not-for-profit community approach to securing access to electronic scholarly content for the long term. More than 53,000 libraries in 96 countries and territories around the world use OCLC services to locate, acquire, catalog, lend and preserve library materials.

3.9 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. PMC contains over half a million articles, most of which have a corresponding entry in PubMed.

3.10 GREENSTONE

Greenstone (http://www.greenstone.org) 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.

3.10.1 Features

- **Interoperability:** Greenstone is highly interoperable using contemporary standards, It incorporates a server that can serve any collection over the Open Archives Protocol for Metadata Harvesting (OAI-PMH), and Greenstone can harvest documents over OAI-PMH and include them in a collection.
- **Metadata formats:** Users define metadata interactively within the Librarian interface. These metadata sets are predefined: Dublin Core (qualified and unqualified),RFC 1807, NZGLS (New Zealand Government Locator Service), AGLS (Australian Government Locator Service).New metadata sets can be defined using Greenstone's Metadata Set Editor. "Plug-ins" are used to ingest externally-prepared metadata in different forms, and plug-ins exist for XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace, METS.

Document formats, Plug-ins are also used to ingest documents. For textual documents, there are plug-ins for PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, Email (various formats), source code For multimedia documents, there are plug-ins for Images (any format, including GIF, JIF, JPEG, TIFF), MP3 audio, Ogg Vorbis audio, and a generic plug-in that can be configured for audio formats, MPEG, MIDI, etc.

3.11 The EPrints

This software is a free (http://software.eprints.org), 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. Using the EPrints software makes the documents OAI compliant by inputting the appropriate OAI metadata tags, the developing standard for online digital repositories and thus searchable and interoperable. The latest recommended version is eprints-2.3.11. In India NCSI has been conducting workshops and training in this context. Ease of setup and installation. An installation script automates most of the installation process, The archive can use any metadata schema, Can store documents in any format The EPrints archive is currently in use in many libraries and centers throughout the world.

3.12 DSpace

DSpace (http://www.dspace.org) 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)
- eTheses
- 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.

Metadata: DSpace uses a qualified Dublin Core metadata standard for describing items intellectually (specifically, the Libraries Working Group Application Profile). Only three fields are required: title, language, and submission date, all other fields are optional. There are additional fields for document abstracts, keywords, technical metadata and rights metadata, among others. This metadata is displayed in the item record in DSpace, and is indexed for browsing and searching the system (within a collection, across collections, or across Communities). For the Dissemination Information Packages (DIPs) of the OAIS framework, the system currently exports metadata and digital material in a custom XML schema while we work with the METS community to develop the necessary extension schemas for the technical and rights metadata about arbitrary digital formats.

4. 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 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 UGE under Ministry of HRD has initiated Institutional Repository and archive its publications, proceedings etc using DSpace (DSpace @ INFLIBNET). 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.

5. 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. The issue of the copyright of intellectual and intangible properties is also a problem towards digital preservation. There is also a burning question: what is appropriate material for preservation and what can be 'edited out'.

References

1. NLA, PANDAS Manual (2006). Available at: http://pandora.nla.gov.au/about.html (August 12, 2006).
2. OCLC, OCLC Digital Archive. Available at: http://pandora.nla.gov.au/about.html (July 28, 2006)
3. Digital Library Federation. http://www.diglib.org/preserve.htm.
4. MacKenzie Smith et.al.(2003). DSpace: An Open Source Dynamic Digital Repositaty.D-LibManagime,9(1),Available-at: ttp://www.dlib.org/dlib/january03/smith/01smith.html

5. http://www.greenstone.org/cgi-bin/library
6. http://www.eprints.org/
7. Fedora 2.0: A Powerful Open-source Solution for Digital Repositories. Summary Report in D-Lib Magazine, 11(3) (March 2005). Available at: http://www.dlib.org .
8. Rodriguez, Andres (2005). Preserving the Last Copy. Computer Technology Review, 25 (3); ABI/INFORM Global. p.17.
9. Han, Yan (2004). Digital Content Management: the Search for Content Management System. Library Hi Tech, 22(4); ABI/INFORM Global. P.355.
10. http://www.lockss.org/lockss/About_LOCKSS.
11. Hodge, Gail (2005). Preservation of and Permanent Access to Electronic Information Resources: A System Perspective. Information Services & Use, 25, p.47-57, IOS Press.
12. http://dspace.inflibnet.ac.in/handle/1944/283
13. Lavoie, Brian F (2003). The Incentives to Preserve Digital Materials: Roles, Scenarios and economic decision making. Available at http://www.oclc.org/research/projects/digipres/incentives-dp.pdf