Digital Preservation and Persistence of Digital Collection

Agrapu Dharini

Abstract

The growth and development in Information and Communication and Technology is multifaceted and turbulent. Rapid revolutionary advances have taken place in the field of computers and their application in libraries. Internet, high band width networks, explosive growth of web based information resources, sophisticated search engines to access information, high resolution capture devices and growing electronic publications have created the development of digital libraries. Digital Libraries build information resources in the Digital form .Digital materials cannot be said to be preserved if access is lost. The purpose of preservation is to maintain the ability to present the essential elements of authentic digital materials. Digital preservation must ensure that they can hold the collection for future generations .. Digital preservation will only happen if organizations and individuals accept responsibility for it. The starting point for action is a decision about responsibility. Everyone does not have to do everything; everything does not have to be done all at once. The present paper defines the concept of digital preservation , challenges, strategies and an Open Archival Information System ,a reference model for digital preservation.

Keywords: Digital Preservation, Electronic Resources, Digital Management, Open Archival Information System.

1. Introduction

The term 'Library' means a collection of documents, their preservation and dissemination of human knowledge. The rapid developments that have occurred since man started recording his information ranges from clay tablets to modern paper and finally to the present electronic digital format.

The rock and stone inscriptions of ancient times can be called the first open library as they depict the cultural heritage and historic value of that society during that era to the present and new generations yet to come. They throw light on that society which no longer exists. In ancient times libraries were established in royal court, temple, mosque, monastery and church. These divine manuscripts is a gift of these libraries. Here the work of conservation and preservation has been carried out by the custodians. Thus libraries served as an agent /bridge for connecting societies of the past, present and future by preserving the knowledge of different phases of time. The history of knowledge preservation began since ancient times through different forms of libraries i.e from outside open library to the present modern library.

Preservation of information resources for the future generations is one of the important objective of a library. The custodian of a library plays a major role in increasing the life span of the documents. Taking due care and attention to the entire document collection is a major challenge for the librarians.

The latest developments in Information Communication Technology transformed the static classical libraries to dynamic multimedia libraries

all over the world. The present century libraries look forward for new innovative methods to transform from libraries within the four walls to digital libraries and virtual libraries. Having learned from developed countries, developing countries are making serious efforts for the development of digital libraries to keep pace with the changing trends in technology and information needs of the user community so as to face global competition in the 21 century.

Libraries realized that users require access to a wide range of information resources distributed throughout the world of information. Libraries must provide access to a broad range of distributed information resources in all types of formats in a timely and effective manner. The need of the hour in the libraries of the 21 centaury is to revise their services, redesign their activities, integrate relevant modern

technologies , computerize functions, build digital information resources, provide access to web resources, repackage their products to add value to the services in order to satisfy the changing complex information needs of their users. All this is one part of the challenge faced by the librarians to provide efficient, effective and timely information to the information needs of the users while on the other hand the preservation of information resources for the future generations is another tedious task faced by the libraries.

The present paper focusses on digital preservation and digital persistence of digitization and digitally born resources and the challenges faced by the libraries that are tuning to engage the readers to read digital resources and digitally born resources.

2. Definition

Chris ,Rusbridge defined digital preservation as" a series of holding positions. Make your dispositions on the basis of the timescale you can foresee and for which you have funding. Preserve your objects to the best of your ability, and hand them on to your successor in good order at the end of your lap of the relay. In good order here means that the digital objects are intact, and that you have sufficient metadata and documentation to be able to demonstrate authenticity, provenance, and to give future users a good chance to access or use those digital objects".

Digital preservation is used to describe the processes involved in maintaining information and other kinds of heritage that exist in a digital form. In these Guidelines, it does not refer to the use of digital imaging or capture techniques to make copies of non-digital items. That is done for preservation purposes. Of course, digital copying (also known as digitization, or digitalization), may well produce digital heritage materials needing to be preserved.

Digital materials is generally used as a preferred term covering items of digital heritage at a general level. This term has been used interchangeably and generically: they do not imply a particular kind of item unless that is clearly stated.

Preservation programme is used to refer to any set of coherent arrangements aimed at preserving digital materials. More commonly used digital materials are digital archive and digital Repository

Presentation, re-presentation describes the processes of providing access to digital materials. The second term means that digital preservation seeks to re-present what was previously stored.

3. Need

Preservation of digital information is widely considered to require more constant and ongoing

attention than preservation of other media. This constant input of effort, time, and money to handle rapid technological and organisational advance is considered as the main stumbling block for preserving digital information. Digital preservation can, therefore, be seen as the set of processes and activities that ensure continued access to information to all kinds of records, scientific and cultural heritage existing in digital formats. This includes the preservation of materials resulting from digital reformatting particularly information that is born-digital and has no analog counterpart. In the language of digital imaging and electronic resources, preservation is no longer just the product of a program but an ongoing process. In this regard the way digital information is stored is important in ensuring their longevity. The long-term storage of digital information is assisted by the inclusion of preservation metadata.

Digital preservation is in its infancy state worldwide. It faces some difficult technological issues. Since the creation of digital media, over 200 different storage mediums have been invented ranging from magnetic tape to CD-Rom. Each of these mediums present a variety of their own preservation issues and also require a diverse range of technology which in many cases is no longer manufactured. In addition to this, there are thousands of different formats in which data can be stored on each medium; and each type of storage format may also require a specific piece of software to interpret the data's meaning.

Digital preservation is not a new concern: it began when the first computers were introduced. A number of national archives, data archives, and other cultural institutions in many countries established digital preservation programs as early as the late 1960s. Those programs reflected the prevailing technology and digital content of that time. Each generation of technology brings changes in potential capabilities to both create and preserve digital content—and will affect a suitable institutional response.

Preservation strategies in academic and research libraries are not new concepts. However, with an increasing amount of digital content, organizations are having to cope with a new set of preservation issues.

4. Challenges In Digital Preservation

4.1 The Viewing Problem

All digital formats require computer technology to view them. By nature technology (software/ hardware/formats) move at such a rapid pace that, odds are, they won't be around when you want to view your data.

4.2 The Scrambling Problem

Data is often compressed or "scrambled" to assist in its storage and or protect it's intellectual content. These compression and encryption algorythms are often developed by private organisations who will one day cease to support them. If this happens the data is stuck between a rock and a hard place. If you don't want to get into legal trouble you are no longer able to read your data; and if you go ahead and "do the unwrapping yourself" it's quite evident that the copy right law is violated.

4.3 The Inter-Relation Problem

Digital information is often linked to other items. This is much more evident in the digital world than the physical. If these links aren't maintained the information is either incomplete, incorrect, or just plain doesn't make any sense. Due to the diversity of digital linkages they're often overlooked. During

the migration process it is possible to lose data. It is also a costly process in terms of work hours and expertise.

4.4 The Custodial Problem

Who is the custodian of a digital document? Is it a librarian's job? What if someone changes the content without telling the custodian, after all digital content is dynamic and easily changed.

4.5 Type of Storage Medium

Magnetic tape, on which most of the world's computer backups are stored, can degrade within a decade.

4.6 Translation Problem

If we need software to interpret data (due to formats etc), and software changes from one version to another advanced version then it will be translated differently in subsequent versions. Even if the software claims it will sometimes it might lose formatting, a font. This is particularly dangerous where the changes are subltle or so small that no one notices them.

4.7 Preservation Management

Librarians know very little about digital preservation. we can lose digital content through poor selection of storage mediums. The progression of technology often motivates institutions to move from one storage medium to another through forced obsolescence. The new and "improved" mediums do not always deliver the same level of functionality.

ronically, many libraries who initially began migrating data to digital repositories under the banner of "economy" have found that it is too costly to keep up with technology and are using microfilm and acid-free paper to back up digital content. When you compare the potential preservation value of a CD-Rom with an undetermined shelf life beginning at only two years, all of a sudden keeping paper records doesn't seem too bad.

4.8 Massive Storage Failures

Basically no matter how much money the library management spends on the system housing digital data there are still many ways in which it can fall over and create opportunities for data to be lost. This may be from hardware/software failure. The longer you try to store data the more likely this will occur.

4.9 Erasing the Data by Mistake

Sometimes people accidentally delete things and if it's the only copy, then it's gone. On the other hand sometimes people think that they no longer need a piece of data and delete it on purpose only to find that it was in fact useful.

4.10 Bitrot

No affordable digital storage is completely reliable over a long period of time. For example some CD's have recently been shown to have a life span of only 2 years which could cause significant problems for anyone relying on them. Other media such as magnetic tape also suffers various types of bit rot. The worse thing about this threat is that it is often undetected until it's too late to recover the material. Bit rot is inevitable with any storage medium over a period of time.

4.11 Outdated Media

Over time all kinds of digital media become outdated. Technology is driven by innovation which unfortunately leads to very short periods of relevancy before redundancy. Data stored on redundant media becomes effectively useless if the appropriate hardware is not available to read it. This

is a particularly difficult issue to manage where data is stored over long periods of time. Ideally, long term data storage should be technology independent, however this is not practical.

4.12 Outdated Formats, Applications and systems

As hardware becomes redundant, so do file formats and the software which interprets them. A good example of this is Word Perfect; try to find a computer today which can read a Word Perfect document properly. Fortunately, system and format redundancy does not usually happen at quite as rapid a pace as hardware.

4.13 Loss of Context

Some data can be related, and their relationship can be vital to data interpretation.. The longer the data is kept without this relationship, the less likely it is to ever be resolved.

4.14 Intenational Attacks

Unfortunately in the world we live in there are some people who intentionally destroy or damage digital assets for a variety of reasons. As much of the information is currently located in open access repositories accessible via the internet it is also vulnerable to attack. This is a threat to both long and short term storage.

4.15 Lack of Resources

Many institutions simply do not have the resources, finance for digital preservation. These strategies are often overlooked as low priority and are likely to remain so until a major data loss scares people into action.

4.16 Organisation Failure

This is a massive threat to long term digital storage of any kind. Technology is so dynamic not only in innovations but also the movement of vendors to compete with each other. It is reasonable not to rely too heavily on any one vendor/system/ sponsoring organisation because they change quickly. Digital assets which need to be preserved long term must be protected from the failure of any one organisation. Unfortunately this is easily said but hard to plan for in such a dynamic environment.

5. Preservation Strategies

Digital technologies are enabling information to be created, manipulated, disseminated, located and stored with increasing ease, preserving access to this information poses a significant challenge. Unless preservation strategies are actively employed, this information will rapidly become inaccessible. Choice of strategy will depend upon the nature of the material and what aspects are to be retained.

5.1 Refreshing, that is, copying information without changing it, offers a short-term solution for preserving access to digital material by ensuring that information is stored on newer media before the old media deteriorates beyond the point at which the information can be retrieved.

5.2 The migration of digital information from one hardware/software configuration to another or from one generation of computer technology to a later one, offers one method of dealing with technological obsolescence.

While adherence to standards will assist in preserving access to digital information, it must be recognised that technological standards themselves are evolving rapidly.

5.3 Technology emulation potentially offers substantial benefits in preserving the functionality and integrity of digital objects. However, its

practical benefits for this application have not yet been well demonstrated.

5.4 Encapsulation**,** a technique of grouping together a digital object and anything else necessary to provide access to that object, has been proposed by a number of researchers as a useful strategy in conjunction with other digital preservation methods.

The importance of documentation as a tool to assist in preserving digital material is universally agreed. In addition to the metadata necessary for resource discovery, other sorts of metadata, including preservation metadata, describing the software, hardware and management requirements of the digital material, will provide essential information for preservation. The requirement to keep every version of all software and hardware, operating systems and manuals, as well as relevant skills, generally makes the preservation of obsolete technologies not a feasible strategy.

5.5 Strategies for Preserving Digital Materials

Digital preservation involves choosing and implementing an evolving range of strategies to achieve the kind of accessibility discussed above, addressing the preservation needs of the different layers of digital objects. The strategies include:

- Working with producers (creators and distributors) to apply standards that will prolong the
- effective life of the available means of access and reduce the range of unknown problems that must be managed.
- Recognising that it is not practical to try to preserve everything, selecting what material should be preserved.
- Placing the material in a safe place.
- Controlling material, using structured metadata and other documentation to facilitate access and to support all preservation process.
- Protecting the integrity and identity of data.
- Choosing appropriate means of providing access in the face of technological change.
- Managing preservation programmes to achieve their goals in cost-effective, timely, holistic, proactive and accountable ways.
- To preserve a copy of the appropriate software and make it available wherever that data is stored. This becomes increasingly unmanageable as the types of systems required increases.
- To migrate data to an acceptable format, for example all text files might be migrated to pdf thus only requiring copies of Adobe Acrobat to be preseved.

6. Reference Models

Various frameworks designed to assist in managing the preservation of digital material have been developed. These include tools designed for assisting in the development of digital preservation strategies. Often these will entail the identification of the various stages at which the provision of long-term access should be considered. Understanding the dimensions and requirements of user communities is increasingly recognized as an essential part of repository design, both from a systems and a human perspective.

6.1 OAIS Reference Model

6.1.1 About OAIS

The Reference Model for an Open Archival Information System (OAIS) has proved an extremely useful model in relation to 'archival systems'. OAIS, the Reference Model for an Open

Archival Information System, was developed by the Consultative Committee for Space Data Systems (CCSDS) to provide a framework for the standardization of long-term preservation within the space science community. OAIS was created with a view to bring widely applicable long-term preservation of digital material. The model exists at an abstract level, providing a conceptual framework. Its usefulness lies in providing a common terminology and act as a communication tool.

OAIS is very clear in its focus on Long-Term preservation. However in a well-managed repository there should be some consideration for preservation and the OAIS model is useful in ensuring that preservation is not forgotten. OAIS draws attention to the important role of preservation for repositories.What it does not do is demand a specific level of preservation, allowing repositories the scope to first assess the needs of their community and information.

6.1.2 Need

The OAIS must

- 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 authentic.

6.1.3. Features

The 'Designated Community' acts as the point where the external environment and the OAIS model interact, it enables repositories to identify who they are providing for, their stakeholders and users, to create policies about what they will offer and to frame their service-provision. This is extremely useful for repositories, yet, on the downside although OAIS makes a small concession towards the existence of multiple communities, it does imply a single knowable Designated Community.

The OAIS Environment can contain multiple communities made up of users, depositors and other stakeholders. In OAIS terms, the environment contains the Producer, Consumer and Management, where management "is the role played by those who set overall OAIS policy as one component in a broader policy domain" (CCSDS 2002, p. 2-2). This simple model provides a good foundation for mapping the external interactions with the repository and can include external systems that might act as Consumer or Producer. It does not incorporate attempts to map the relationships between Consumers and Producers, between the different communities and the stakeholders that do not fit within these three categories.

The Information Model is another key aspect of OAIS, also required for conformance to OAIS. It provides a loose framework identifying the different blocks of data and metadata that make up the Information Package. In OAIS terminology the

information package is made up of: content information, which comprises the data object (either digital or physical), together with its representation information (structural and semantic information) to enable interpretation of the data object, used in conjunction with the external Knowledge Base of the Designated Community.

OAIS defines three information packages handled by a repository

1. The Submission Information Package (SIP)
2. The Archival Information Package (AIP)
3. The Dissemination Information Package (DIP)

The SIP represents the data and metadata that comes from the Producer; the AIP is the data and metadata preserved by the repository and the DIP is the data and metadata that are sent to the Consumer on request.

OAIS defines six functional entities, each with a set of sub-functions, further identifying how information flows between these entities.

Ingest is the function that accepts an information package from the Producer, checks and updates it, generates the version for storage and creates, or augments, the descriptive information. It is only during the Ingest stage that the repository and its Producer interact and it is at this stage that the information received is enhanced to ensure its usability by the Consumer. The pre-ingest function includes contacts and negotiations between the Producer and repository (the preliminary phase), SIP design and submission agreement (formal definition phase), transfer of the SIP (transfer phase) and validation processing and Producer follow-up (validation phase). At this crucial point, where the information first flows into a repository, OAIS appears to provide little guidance, despite the fact that it is this information flowing in that drives the remaining functions.

Archival Storage offers the basic storage and backup of data, rather than the metadata, receiving it from the Ingest function and providing it to the Access function. Error checking, media replacement and disaster recovery are part of Archival Storage. For repositories this might be a file store with a structure and backup.

Data Management is where the descriptive and system information are stored, most likely in a database. This function is also responsible for maintaining the database, performing queries sent by the Access function and generating reports. For repositories, this might be an open-source database, such as MySQL, and a series of scripts and configuration files. It might incorporate a web-accessible report generator, and other automated processes, as well as the effort of a database administrator to run specific queries and undertake any necessary development work.

Administration connects to every other function and also interfaces with the Producer, Consumer and Management. For any organization or repository, this function will undoubtedly be the most difficult to understand and is likely to involve different staff across different departments depending on the size of repository and its role within the organisation or institution as a whole. For example, an Institutional Repository might be administered by a Librarian, liaising with various technical support staff with responsibilities for different systems tasks. It might have an advisory board, or management group, with representatives from across the institution and there may also be interactions with Academic Schools regarding

material submission and with administrative departments for policy support.

OAIS provides a useful set of sub-functions for Administration, it provides too much detail about a set of discrete functions without providing an overview or indication of the full set of administrative functions a repository might need to fulfill.

Logically part of the Administration function is Preservation Planning. It has drawn out into its own high-level functional entity because of the preservation focus of OAIS. Concerned with monitoring and setting policy, the Preservation Planning function does not carry out actual preservation activities, rather it is responsible for carrying out a technology watch function, monitoring changes in community requirements, recommending changes and updates, designing the information packages and developing preservation policies.

The final function, Access, is where the repository interfaces with its Consumers, receiving queries and requests, delivering responses and connecting with the Data Management and Archival Storage functions to generate the DIP. It might be useful here to consider the two levels of communities identified by CD-LOR (Community Dimensions of Learning Object Repositories).The user who simply wants materials, and the wider stakeholder community who might want a different set of information These users can be a local or remote and might include interactions with external systems, such as OAI-harvesters or federated search services that rely on existing standards (e.g. OAI-PMH or SRU) and pre-defined metadata schemas (e.g. Dublin core or IEEE LOM). For all repositories, access is a necessary and existing function. The OAIS Access function provides a simple, abstract, model for the way a repository interacts with its Consumers. Points of contention do exist, though, and further analysis may be necessary of existing repository practice.

Hence the OAIS ensures good practice. Requirements for compliance to OAIS are low-level. To fulfill the mandatory responsibilities, a repository must define its long-term preservation commitment. One should dedicate some time and effort to understand and document its processes, practices, functions, information, workflow and Designated Communities. Policies, guidelines and agreements should exist and these should demonstrate the sustainability and viability of the repositories' business model. Arguably, if repository developers and administrators are guided by reference model, they are more likely to consider the right issues. By using OAIS as that reference model, awareness of long-term preservation is heightened and this could help embed preservation into the workflow, as well as demonstrating a commitment to long-term viability and sustainability and engendering trust.

One of the key strengths of OAIS is its abstract nature, allowing the model to be adapted for specific needs, such as repositories for different communities, functions or material types. Detailed implementation models could be layered beneath the high-level OAIS to provide additional context, guidance and examples, and to identify technical standards and specifications. In fact, there is nothing in the OAIS model that presents insurmountable difficulties for repositories. The conceptual and flexible nature of OAIS allows repositories to adapt and extend their own functional and informational models to take local practices into account.

7. Conclusion

Libraries have been used to preserve materials for future generations. As information is being produced processed and stored in digital format and distributed in the electronic environment such as the Internet and CD- ROM.. Little progress has been made in archiving digital information to produce the nation's cultural heritage and record intellectual discourse. Digitization of library collection is important to provide wider access, space for more collection and to preserve for future generations. Libraries should prepare criteria for identification and selection of the materials that need to be digitized. The most important criteria to be considered is the information communication technology infrastructure, availability of funds, trained staff, copyright clearance and cost effectiveness.

OAIS was created to serve a specific community (space science) in carrying out a specific business requirement (long-term preservation). A heightened awareness of preservation has both direct and indirect benefits for repositories and compliance to the OAIS framework is relatively easy to achieve.

References

1. **Marcum, D B,** Ed. Development of Digital Libraries. Westport:Greenwood Press.2001.
2. **OAIS Reference Model** available at http://www.ccsds.org/documents/650x0b1.pdf (Accessed on 04/01/2009).
3. **DPC Technology Watch** Report on OAIS model by Brian Lavoie (OCLC Research):available at http://www.dpconline.org/ (Accessed on 04/01/2009).
4. **RLG/NARA Task Force on Digital Repository** Certification: available at http://www.rlg.org/ (Accessed on 06/01/2009)

About Author

Mr. Agrapu Dharini, Research scholar, DLIS, Andhra University, Visakhapatnam.530004
Prasad_bode@yahoo.com