Is Digital Archiving Simple ?

J P S Kumaravel

Abstract

Explains what is digital preservation – and the benefits of digital preservation - discusses the barriers in digital archival – elucidates the steps in digital archiving.

Keywords : Digital Archiving, Digital Library, Digital Preservation

0. Introduction

The history of Digital libraries spans more than a decade and are the most important and influential institutions to emerge in recent times, their mission is to dissolve barriers to 'information equity'; Digital libraries seek to bring the highways of knowledge to everyone. The explosion of interest in research and practice of digital libraries has spanned the boundaries of computing, networking, and information systems and sciences.

Digital libraries are complex information systems. Their design, development, management and use require application of scientific, technological, methodological, economic, legal and other innovations. Digital library technologies are rapidly developing and are still evolving. The grand challenges of Digital libraries research and development centers on interoperability-the ability to access

1. Digital preservation

There is enormous growth in the creation and dissemination of digital objects by authors, publishers, corporations, governments, and even librarians, archivists and museum curators. Much has been emphasized in the speed and ease of short-term dissemination. But the regard for the long-term preservation of digital information is very low. However, digital information is fragile in ways that differ from traditional technologies, such as paper or microfilm. It is more easily corrupted or altered without recognition.

Digital storage media have shorter life spans, and digital information requires access technologies that are changing at an ever-increasing pace. Some types of information, such as multimedia, are so closely linked to the software and hardware technologies that they cannot be used outside these proprietary environments[1]. Because of the speed of technological advances, the time frame in which we must consider archiving becomes much shorter. The time between manufacture and preservation is shrinking. While there are traditions of stewardship and best practices that have become institutionalized in the print environment, many of these traditions are inadequate, inappropriate or not well known among the stakeholders in the digital environment. Originators are able to bypass the traditional publishing, dissemination and announcement processes that are part of the traditional path from creation to archiving and preservation.

Librarians and archivists who traditionally managed the life cycle of print information from creation to long-term preservation and archiving, must now look to information managers from the computer science tradition to support the development of a system of stewardship in the new digital environment. There is a need to identify new best practices that satisfy the requirements and are practical for the various stakeholder groups involved

2. Benefits of Digital Archiving

Some of the benefits in the migration to digital data are:

- Shelf space conservation, a valued commodity in constant demand in the library
- Improved access to reference data
- Easy to cross-reference or cross-link data
- Easy-to-use reference resource that includes on-line training tutorials
- Easy-to-upgrade resource
- Standard format for spectral data
- Greater searching capability via a quick “searcher friendly” system
- Intuitive user interface
- Ability to incorporate laboratory data generated by students and faculty
- Simplification of teaching and research

3. Barriers in the Digital Age

Some of the barriers which still exist, even for digital data are:

- Diverse needs from user types (academic, forensic, polymer); “one size doesn’t fit all”
- Content available but at moderate to high cost; generating new spectral content is costly
- Scaling product to variable funding sources
- Migration issues for old or “heritage” data on the part of users
- Wide variations in skill with computing, software learning curve, and consumer’s time available to learn software
- Constant changes in instrument operating systems

4. Digital Archiving – steps

Digital archiving is not just storing the digital information for future use. It involves various steps such as:

- creation,
- acquisition,
- cataloging/identification,
- storage,
- preservation and
- access.

4.1 Creation

Creation of digital information is the act of producing the information product. The producer may be a human author or originator, or a piece of equipment such as a sensing device, satellite or laboratory instrument. Creation is viewed here in the broadest sense, as increasingly science is based on a variety of data types, products and originators. Creation though it is the initial stage the other phases of work like long-term archiving and preservation must start. Even in rigorously controlled situations, the digital information may be lost without the initial awareness on the part of the originator of the importance of archiving. Practices used when a digital object is created ultimately impact the ease with which the object can be digitally archived and preserved.

The creator has to assess the long-term value of the information. In lieu of other assessment factors, the creator’s estimate of the long-term value of the information may be a good indication of the value that will

be placed on it by people within the same discipline or area of research in the future. For example the U.S. Department of Agriculture's Digital Publications Preservation Steering Committee has suggested that the creator provide a *preservation indicator* in the document.

The preservation and archiving process is made more efficient when attention is paid to issues of consistency, format, standardization and metadata description in the very beginning of the creation. The best practice would be to create the metadata at the object creation stage, or to create the metadata in stages, with the metadata provided at creation augmented by additional elements during the cataloging/ identification stage.

In the case of data objects, the metadata is routinely collected at the point of creation. Many of the datasets are created by measurement or monitoring instruments, and the metadata is supplied along with the data stream. This may include location, instrument type, and other quality indicators concerning the context of the measurement. In some cases, this instrument-generated metadata is supplemented by information provided by the original researcher.

For smaller datasets and other objects such as documents and images, much of the metadata continues to be created "by hand" and after-the-fact. Metadata creation is not sufficiently incorporated into the tools for the creation of these objects to rely solely on the creation process. As standards XML (eXtensible Mark-up Language) and RDF (Resource Description Framework) architectures during the creation of metadata as part of the origination of the object will be easier.

4.2 Acquisition

Acquisition and collection development is the stage in which the created object is "incorporated" physically or virtually into the archive. The object must be known to the archive administration. There are two main aspects to the acquisition of digital objects namely

- collection policies
- gathering procedures.

4.2.1 Collection Policies

The major difference in collection policies between formal print and electronic publications is the question of whether digital materials are included under current deposit legislation. Guidelines help to establish the boundaries in such an unregulated situation. Much of the digital material could be archived from the Internet, so guidelines are needed to tailor the general collection practices of the organization.

4.2.1.1 Selecting What to Archive

Both the National Library of Canada (NLC) and the National Library of Australia (NLA) acknowledge the importance of selection guidelines. The NLC's Guidelines state, "The main difficulty in extending legal deposit to network publishing is that legal deposit is a relatively indiscriminate acquisition mechanism that aims at comprehensiveness. In the network environment, any individual with access to the Internet can be a publisher, and the network publishing process does not always provide the initial screening and selection at the manuscript stage on which libraries have traditionally relied in the print environment… Selection policies are, therefore, needed to ensure the collection of publications of lasting cultural and research value." [NLC 1998]

The guidelines for the selection on online publications intended for preservarion should include scholarly publications of national significance and those of current and long term research value. Other items

should be archived on a selective basis “to provide a broad cultural snapshot of how people of the country are using the Internet to disseminate information, express opinions, lobby, and publish their creative work.

4.2.1.2 Determining Limitations

Directly connected to the question of selection is the issue of extent. What is the extent or the boundary of a particular digital work? This is particularly an issue when selecting complex Web sites. The extensive use of hypertext links to other digital objects in electronic publications raises the question of whether these links and their content should be archived along with the source item.

4.2.1.3 Refreshing the Archived Contents

In cases where the archiving is taking place while changes or updates may still be occurring to the digital object, as in the case of on-going Web sites, there is a need to consider refreshing the archived contents. A balance must be struck between the completeness and currency of the archive and the burden on the system resources. Obviously, the burden of refreshing the content increases as the number of sources stored in the archive increases.

4.2.1.4 Gathering Approaches

There are two general approaches to the gathering of relevant Internet-based information — hand-selected and automatic. The sites have to be reviewed and hand-selected and then monitored for their persistence before being included in the archive. The other way is to develop a program which automatically captures sites from the known Web servers.

4.2.1.5 Intellectual Property Concerns

Intellectual property remains a key issue in the acquisition process. The approaches to intellectual property vary based on the type of organization doing the archiving. In the case of data centers or corporate archives where there is a close tie between the center and the owner or funding source, there is little question about the intellectual property rights related to acquisition. However, in the case of national libraries, the approaches to intellectual property rights differ from country to country. The differences are based on variant national information policies or legal deposit laws. In many countries, the law has not yet caught up with the digital environment, and the libraries must make their own decisions.

4.3 Identification and Cataloging

Once the archive has acquired the digital object, it is necessary to identify and catalog it. Both identification and cataloging allow the archiving organization to manage the digital objects over time. Identification provides a unique key for finding the object and linking that object to other related objects. Cataloging in the form of metadata supports organization, access and curation. Cataloging and identification practices are often related to what is being archived and the resources available for managing the archive.

4.3.1 Metadata

All archives use some form of metadata for description, reuse, administration, and preservation of the archived object. There are issues related to how the metadata is created, the metadata standards and content rules that are used, the level at which metadata is applied and where the metadata is stored. There is increasing interest in automatic generation of metadata, since the manual creation of metadata

is considered to be a major impediment to digital archiving. The metadata formats depends on the data type, discipline, resources available, and cataloging approaches used.

4.3.2 Persistent Identification

For those archives that do not copy the digital material immediately into the archive, the movement of material from server to server or from directory to directory on the network, resulting in a change in the URL, is problematic. The use of the server as the location identifier can result in a lack of persistence over time both for the source object and any linked objects. Despite possible problems, most archives continue to use the URL when referencing the location for the digital object.

4.4 Storage

Storage is often treated as a secondary nature in digital archiving, but storage media and formats have changed with legacy information perhaps lost forever. Block sizes, tape sizes, tape drive mechanisms and operating systems have changed over time. The most common solution to this problem of changing storage media is migration to new storage systems. This is expensive, and there is always concern about the loss of data or problems with the quality when a transfer is made. Check algorithms are extremely important when this approach is used.

4.5 Preservation

Preservation is the aspect of archival management that preserves the content as well as the look and feel of the digital object. Depending on the particular technologies and subject disciplines involved the hardware/software migration must be decided.

4.5.1 Hardware and Software Migration

New releases of databases, spreadsheets, and word processors can be expected at least every two to three years, with patches and minor updates released more often. While software vendors generally provide migration strategies or upward compatibility for some generations of their products, this may not be true beyond one or two generations. Migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features. There is generally no backward compatibility, and if it is possible, there is certainly loss of integrity in the result. Emulation, which encapsulates the behavior of the hardware/software with the object, is being considered as an alternative to migration.

4.5.2 Preservation of the Look and Feel

At the specific format level, there are several approaches used to save the “look and feel” of material. For journal articles, the majority of the projects reviewed use image files (TIFF), PDF, or HTML. TIFF is the most prevalent for those organizations that are involved in any way with the conversion of paper backfiles. HTML/SGML (Standard Generalized Mark-up Language) is used by many large publishers after years of converting publication systems from proprietary formats to SGML. PDF versions can also be provided by conversion routines.

For purely electronic documents, PDF is the most prevalent format. This provides a replica of the Postscript format of the document, but relies upon proprietary encoding technologies. PDF is used both for formal publications and grey literature.

4.5.3 Transformation vs. Native Formats

A key preservation issue is the format in which the archival version should be stored. Transformation is the process of converting the native format to a standard format. On the whole, the projects reviewed favored storage in native formats. However, there are several examples of data transformation.

4.5.4 Standards and Interoperability

One of the paradoxes of the networked environment is that in an environment that is so dynamic and open to change, there is a greater and greater emphasis on standards. Those projects that have been archiving for a long period of time indicated that while they started out with a large number of incoming formats — primarily textual — the number of formats have decreased.

4.6 Access

All the steps previously discussed are performed for the purpose of ensuring continuous access to the material in the archive. Successful practices must consider changes to access mechanisms, as well as rights management and security requirements over the long term.

4.6.1 Access Mechanisms

Today it is the Web, but there is no way of knowing what it might be tomorrow. It may be possible in the future to enhance the quality of presentation of items from the digital archive based on advances in digitization and browser technologies.

4.6.2 Digital Rights Management and Security Requirements

One of the most difficult access issues for digital archiving involves rights management. What rights does the archive have? What rights do various user groups have? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? Rights management includes providing or restricting access as appropriate, and changing the access rights as the material's copyright and security level changes. Security and version control also impact digital archiving. In cases where conservation issues are at stake, it is important to have metadata to manage encryption, watermarks, digital signatures, etc. that can survive despite changes in the format and media on which the digital item is stored.

5. Conclusions

Within the sciences, there are a variety of digital archiving projects that are at the operational or pilot stage. Standards for creating digital objects and metadata description, which specifically address archiving issues, are being developed at the organization and discipline levels. Regardless of whether acquisition is done by human selection or automated gathering software, there is a growing body of guidelines to support questions of what to select, the extent of the digital work, the archiving of related links and refreshing the contents of sites. Standards for cataloging and persistent, unique identification are important in order to make the material known to the archive administration. A variety of metadata formats like Dublin core and RDF, content rules and identification schemes are currently in use, with an emphasis on crosswalks to support interoperability, while standardizing as much as possible.

Current practice is to migrate from one storage medium, hardware configuration and software format to the next. This is an arduous and expensive process that may be eliminated if emulation strategies are developed among standards groups and hardware and software manufacturers. Access mechanisms,

being hardware and software based, have their own migration issues. In addition, there are concerns about rights management, security and version control at the access and re-use stage of the life cycle.

While there are still many issues to be resolved and technology continues to develop a-pace, there are hopeful signs that the early adopters in the area of digital archiving are providing lessons-learned that can be adopted by others in the stakeholder communities. Through the collaborative efforts of the various stakeholder groups — creators, librarians, archivists, funding sources, and publishers — and the involvement of information managers, a new tradition of stewardship will be developed to ensure the preservation and continued access to our scientific and technological heritage.

6. References

1. Neil Beagrie and Daniel Greenstein. "A Strategic Policy Framework for Creating and Preserving Digital Collections." July 14, 1998. <http://www.ahds.ac.uk/manage/framework.htm>
2. Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS): Recommendation Concerning Space Data Systems Standards." White Book CCSDS 650.0-W-4.0, September 17, 1998. <http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html>
3. John Garrett and Donald Waters. "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information." Commissioned by the Commission on Preservation and Access and the Research Libraries Group, Inc. 1996. <www.rlg.org/ArchTF/tfadi.index.htm>
4. Margaret Hedstrom and Sheon Montgomery. "Digital Preservation Needs and Requirements in RLG Member Institutions." A study commissioned by the Research Libraries Group. December 1998. <http://www.rlg.org/preserv/digpres.html>
5. Alan R. Heminger and Steven B. Robertson. "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents" November 21, 1998. <http://tuvok.au.af.mil/au/database/research/ay1996/ afit_la/rober_sb.htm>
6. Gail Hodge. "Digital Electronic Archiving: The State of the Art, The State of the Practice." April 26, 1999. <http://www.icsti.org/>
7. Brewster Kahle. "Preserving the Internet." *Scientific American*, March 1997.
8. Terry Kuny. "The Digital Dark Ages? Challenges in the Preservation of Electronic Information." International Preservation News, No. 17, May 1998.
9. Jeffrey Rothenberg. "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.". Report to CLIR, January, 1999. <http://www.clir.org/pubs/reports/rothenberg/contents.html>

About Author

Shri J P S Kumaravel is Information Scientist from Dr. T.P.M. Library, Madurai Kamaraj University. He holds M.L.I.Sc., M.C.A., M.Phil., (Lib. Sci.) and is currently engaged for Ph.D. He has 27 years of experience in the profession and has been actively involved in teaching and training. He has a number of publications to his credit.