



IFLA
2005
OSLO

World Library and Information Congress: 71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:

<http://www.ifla.org/IV/ifla71/Programme.htm>

May 28, 2005

Code Number:

032-E

Meeting:

150 SI - ICABS (IFLA/CDNL Alliance for Bibliographic Standards)

Getting started: What needs to be in place to maintain access to digital collections¹

Pam Gatenby

Assistant Director General, Collection Management
National Library of Australia

Abstract

The challenges that face national collecting institutions in carrying out their role in the digital age are immense. Faced with a plethora of information and insufficient resources, the way forward can seem daunting- however, many examples around the world demonstrate that it is possible to make an impact by starting small and by doing something rather than waiting for all issues to be solved. To support the management and long-term preservation of digital resource, appropriate institutional policies, processes, work practices, skills and technical infrastructure are required. However, it is not necessary, and may not even be desirable, to set up well developed digital archiving and preservation programs from scratch before action is taken. Minimal programs can still safeguard significant heritage resources that might otherwise be lost. The paper discusses the processes that are core to programs aiming to provide on-going access to digital resources. These are:

¹ This presentation is based on the documents listed in the bibliography at the end of the paper. Additional information relevant to the theme of the presentation is contained in those documents. I would like to acknowledge Colin Webb, Director of Preservation at the National Library of Australia, the author of the UNESCO guidelines document, which the presentation draws on in particular.

- **Selecting and gathering** resources considered to have lasting value into a safe place (generally referred to as “digital archiving”);
- **Organising** the gathered resources to facilitate their management and access; and
- Implementing a **digital preservation** (or data management) program to overcome threats to the resources posed by changes to the software and technology on which they depend.

Introduction

Digital resources, the World Wide Web and the Internet are now a central and essential part of the world’s information environment. Most national libraries and other collecting institutions accept that they have some role to play in keeping at least a small part of these global resources available for use by future generations. This role is considered an extension of the responsibility we have for safeguarding other records of human endeavour and achievement.

We now all understand that the challenges that face us in carrying out our role in the digital age are immense. Faced with a plethora of information and insufficient resources, the way forward can seem daunting- however, many examples around the world demonstrate that it is possible to make an impact by starting small and by doing something rather than waiting for all issues to be solved.

In my talk I will address the topics covered by the title of my presentation, in reverse order – I will first of all provide an overview of the key processes and systems that need to be in place to support ongoing, effective management of digital resources for long term access. I will then provide some comments on how institutions that have not yet committed to this might go about getting started. My focus is on web resources and on the responsibilities of national libraries although many of the points made apply more widely to other forms of digital resources and other types of institutions.

The presentation provides a broad overview of the topics and does not attempt to provide detailed guidance. In this way it serves as a framework or introduction to the other presentations on the program today which will address in more detail some of the processes or actions I mention in my overview.

Points to keep in mind

Before looking at the more tangible aspects of what needs to be in place to support a digital resources management program, I would like to suggest some principles or points to keep in mind when setting up a program². They are as follows:

- While most programs commence as projects, at some stage they need to be established as sustainable business operations.

² These are taken from Guidelines for the preservation of digital heritage (<http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>) prepared by the National Library of Australia under contract to UNESCO. The full list is available in Chapter 5.

- Institutions need to clarify their legal right to collect, copy, preserve and provide access to resources.
- Not all digital resources need to be kept, only those judged to have ongoing value – however, while a decision to preserve can be subject to later review, a decision not to preserve is most likely final.
- For those resources that warrant keeping, continuity of access is the objective - digital resources cannot be said to be preserved if access is lost.
- Ongoing access will only be achieved if organisations and individuals accept responsibility for it.
- Having accepted responsibility, ongoing access requires sustained, direct action – it will only happen if it is made to happen; benign neglect is not enough.
- Waiting for comprehensive and reliable solutions to appear before taking action will probably mean material is lost. It can be better for non-comprehensive and non-reliable action to be taken than for no action at all. Small steps are usually better than no steps.
- Managing digital resources involves the assessment and management of risks and decisions about acceptable levels of loss of content and means of access.
- Collaboration is often a cost effective way to build digital preservation programs as it can provide wide coverage, mutual support and shared expertise – but collaboration also involves costs and choices.
- Standards based systems and processes are highly desirable to reduce the range of options needing to be managed and to foster collaboration, support interoperability over time and distance, and reduce guesswork in applying preservation processes.

What needs to be in place to support management of digital resources?

In a nutshell, what needs to be in place to support the management and long-term preservation of digital resources are appropriate institutional policies, processes, work practices, skills and technical infrastructure. However, a pre-condition to embarking on a digital preservation program is an institution's acceptance of responsibility for maintaining access to a collection of digital resources.

Programs for managing digital resources are more likely to be sustainable if they fit within the context of an institution's broader policies and responsibilities regarding collecting, access services, preservation and public accountability. This context should inform decisions and policy concerning what should be collected, what additional infrastructure support might be needed, what standards should be employed and the appropriate legal framework for carrying out the work. Success can also depend on where the program is located within the organisational structure and the extent of corporate support for what is likely to be a resource intensive and complex undertaking.

Key Actions

While there are different, equally valid business models for managing digital resources, to achieve the objective of on-going access to the resources certain core processes must be carried out. These can be grouped around a set of key actions:

- **Selecting and gathering** resources considered to have lasting value into a safe place (generally referred to as “digital archiving”);
- **Organising** the gathered resources to facilitate their management and access; and
- Implementing a **digital preservation** (or data management) program to overcome threats to the resources posed by changes to the software and technology on which they depend.

Selecting and gathering resources

Before resources are gathered, policy and procedures need to be established to guide decisions on what should be collected, how comprehensive collecting should be and how collecting should be carried out.

The *selection* of digital heritage is conceptually the same as selection of non-digital materials. Many of the same approaches are required – for instance, selection based on collection development policies, consideration of the lasting value of the information, and knowledge of the context and provenance of the resource. However, digital resources present some new challenges including the large amount of it available, inconsistent quality due to the widely available means of producing it, its transitory nature, and the difficulty of defining its boundaries and of establishing ownership. While selection must allow for judgment and be able to deal with uncertainty, it should be based on policy that reflects the objectives of the organisation and that, for accountability purposes, is available to the public. It should state clearly what factors influence the policy and why resources are included or excluded.

To protect digital resources that have been selected, it is generally agreed that it is necessary to remove them from an operating environment into safe storage (or a digital archive) where they can be managed for ongoing accessibility. Some institutions will create their own archives while others operating under a distributed and collaborative model, may decide to use the services provided by another organisation. The archival storage used must be secure and reliable and support processes that protect the data stream from unintended damage, change or loss. Such a capability can be achieved with modest equipment so long as the equipment and the system are well managed. However, the more diverse and complex the resources are, the more sophisticated the storage system needs to be.

Key management issues associated with gathering digital resources into a safe place or archive include:

The legal basis for gathering

Rights under Copyright law and Legal Deposit provisions need to be considered before resources are gathered. In particular, permission to provide access to archived resources and to copy them for preservation purposes is required.

The approach to gathering

Most transfer strategies are variations on two basic approaches – producers depositing resources or the institution gathering the resources. (For web resources, gathering means obtaining a copy of the original resource.) Whichever approach is used,

appropriate workflows and tools need to be in place that support decisions on matters such as the frequency of gathering non-static resources, and how more complex, restricted access resources will be dealt with. The International Internet Preservation Consortium (IIPC)³ is currently developing a web harvesting tool (called Heritrix) that is more suitable for the archiving needs of cultural heritage institutions than other public domain software. It has also developed tools and processes for supporting the collecting and archiving of data-base driven web resources and is currently developing a curators' tool kit to support small institutions without much IT support to gather and manage web resources.

Organising resources

Once digital resources have been transferred to an archive (or safe storage) they must be controlled and organised so they can be easily located, accessed, used, managed and preserved. This entails the following activities:

Assigning identifiers

Each digital file within a digital repository must be given a unique file name so that it cannot be confused with any other file and a resource must be assigned a persistent identifier that can be resolved so links to the resource do not break if the resource changes its location.

Describing the digital objects using standard metadata schemes

Metadata (structured information about resources) enables resources to be found through resource discovery services and, from a preservation point of view, to be managed and represented. A standardised approach to use of metadata is desirable as software tools can automatically recognise standard metadata elements, the metadata can be shared with other services, it provides consistency of searching and encourages standardisation of the preservation processes it describes. There are several important metadata initiatives underway including the PREMIS project, convened by OCLC and RLG, which aims to recommend metadata elements and implementation practices for managing the sustained survival and use of digital resources. The Library of Congress is also very active in metadata standards development and their work in this area as it relates to support for ongoing access to digital resources, will be reported later in today's program.

To support resource discovery and use a basic level of metadata (such as title, creator name, publisher, and subject terms) should be recorded, and the metadata included in a resource discovery service such as a library catalogue or public database. To support digital resource management and preservation, metadata about the technical nature of a resource (such as resource type, file format, data structure, processes used in its creation, and software required to operate it) is needed. Information about the changes and processes that have been applied to a resource throughout its life and about the means of providing access to it is also required.

Providing a means of access

³ The International Internet Preservation Consortium (IIPC) is a consortium of 12 national libraries working together to develop standards and tools to support web archiving. (<http://www.netpreserve.org>)

For most collecting institutions, the objective of safeguarding digital resources is current and future access. To enable archived resources to be used it is necessary to provide a means of access to them. This can involve a searchable interface to the repository in which they are managed, a mechanism for delivering the resources for viewing, and management of restrictions and any other conditions associated with access. The extent of access provided will depend on rights under appropriate Copyright law and/or permissions negotiated with creators and publishers. Effective access is dependent on reliable descriptive metadata that can be employed in whatever form of resource discovery interface provided. For instance, the National Library of Australia's PANDORA archive (<http://www.pandora.nla.gov.au>) can be accessed via bibliographic records in the Library's OPAC, or via title listings or an index search available from the archive's homepage. Other archives, such as the Internet Archive, provide time line searching that allows different versions of web resources captured over time to be compared.

The digital archiving management system used by the National Library of Australia is called PANDAS (PANDORA Digital Archiving System). It was developed by the Library to meet its own requirements and is available free of charge to other institutions. An Evaluation System is available to assist institutions assess the suitability of PANDAS for their needs. (For more information go to <http://www.pandora.nla.gov.au>)

Implementing a digital preservation program

Digital preservation consists of processes aimed at ensuring the continued accessibility of digital resources. This requires the maintenance of intact data streams and the means to read and interpret them and to re-present to users what was originally presented. Good understanding of the characteristics of the resources and of the threats to ongoing accessibility is central to achieving this. Digital storage media are vulnerable to damage or deterioration, and hardware and software become obsolete, so the time frame for making preservation decisions is short compared to paper-based materials. Action needs to be taken early and sustained for as long as the resources are needed.

Key approaches to digital preservation include:

Clarifying objectives

Decisions need to be made about what needs to be preserved and for how long and the scope of the institution's responsibilities and decisions outlined in a policy statement. The level of functionality and authenticity of resources that will be required by future users needs to be considered as well as the information that needs to be recorded to understand the resources into the future.

Knowing your collection

It is important to identify what is in your digital collections and to record the technical characteristics of resources using a standard metadata schema. Knowledge of the collection can be used to determine priorities including pressing risks such as aging or vulnerable media and software and file formats that are becoming obsolete.

Protecting the data streams

This is essential to maintaining access to digital resources and entails managing ongoing data protection in accordance with good IT practices. Specifically, it can involve appropriate storage for physical carriers, copying data to a reliable digital storage system with regular back-ups and disaster recovery measures in place, and regular copying to fresh media.

Managing risks

Digital preservation programs must seek to understand and respond to threats that would jeopardise ongoing accessibility to resources. A risk management approach provides an appropriate basis for deciding what risks warrant attention and for planning action that will lower the level of risk. It also assists with planning ahead and setting priorities for action.

Maintaining accessibility

The key challenge for digital preservation programs is future use of resources when technological changes make original formats and operating environments obsolete. Long term solutions are still being developed and may include migrating data to new formats, creating software tools to present original data and emulating original operating environments to run original software.

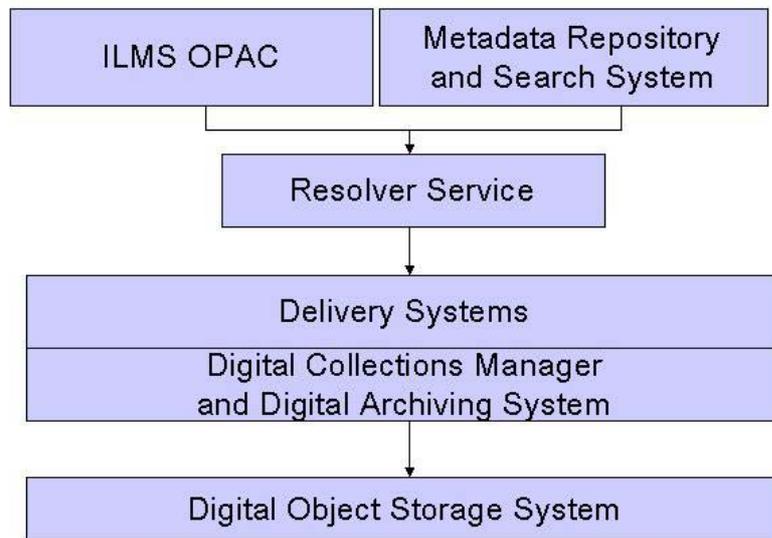
Steps that can be taken in the meantime include:

- Retaining for as long as possible the hardware and software required to access the resources in the collection;
- Limiting the range of formats to be supported by transferring resources to standard formats;
- Retaining and managing original data in the event that future solutions may be able to restore access;
- Monitoring the technological environment for signs that equipment, formats and standards are becoming obsolete; and
- Working with other institutions in developing solutions.

Technical infrastructure

To support the key actions and approaches outlined in this paper, a robust technical infrastructure needs to be in place. The architecture currently used by the National Library of Australia provides an example of an infrastructure that supports digital library activities including digitisation of existing collection materials, collecting born digital resources, and managing digital resources for long-term access. It is represented diagrammatically below.

NLA Digital Services Architecture



Getting started

It is not necessary, and may not even be desirable, to set up well developed digital archiving and preservation programs from scratch before action is taken. Indeed, this would be a daunting prospect for most collecting institutions. Instead, much can be gained from starting small with a modest amount of resource, possibly limited to straightforward data, with the aim of providing the best level of management you can within existing constraints and learning as you go. The following steps may help in setting up a program to manage digital resources for long-term access:

- Decide what resources you are responsible for and liaise with others who have similar responsibilities to see if a cooperative approach is possible.
- Identify other programs that have experience in managing the kind of resources you are interested in and seek their advice and guidance.
- Identify the most pressing threats that require immediate attention to prevent valuable information from being lost.
- Identify any immediate steps that can be taken, especially simple steps that can be taken quickly, to deal with the threats. These could include protecting the data by storing and managing it under appropriate conditions, and getting a better understanding of the characteristics of the material and how users might expect to be able to use it.
- Work out the rights or permissions you need to take action and approach creators to negotiate these if necessary.
- Develop at least early policies that will guide the commitments that you make.
- Evaluate every step of the action you take and review what else is required to develop a comprehensive and reliable program over time – learn from experience and build on it.

- Estimate the resources that will be needed to develop a more robust, reliable program over time.

While obviously limited in impact, minimal programs can still safeguard significant heritage resources that might otherwise be lost. As mentioned at the beginning of the presentation under Points to keep in mind, waiting for comprehensive and reliable solutions to appear before taking action will probably mean material is lost. It can be better for non-comprehensive and non-reliable action to be taken than for no action at all.

Bibliography

1. *Guidelines for the preservation of digital heritage* / prepared by the National Library of Australia. Information Society Division, United Nations Educational, Scientific and Cultural Organization, 2003. (<http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>)
2. *Managing web resources for persistent access*. National Library of Australia, 2001. (www.nla.gov.au/guidelines/2000/persistence.html)
3. *Recommended practices for digital preservation*. National Library of Australia, 2004. (www.nla.gov.au/preserve/digipres/digiprespractices.html)
4. *Safeguarding Australia's web resources: guidelines for creators and publishers*. National Library of Australia, 2002. (www.nla.gov.au/guidelines/webresources.html)