

## Recent Developments in Digital Archiving and Preservation

Prepared for the CDNL Meeting, Seoul, 23 August 2006  
Pam Gatenby, Assistant Director General, Collection Management,  
National Library of Australia

This presentation provides an overview of some recent developments relating to digital archiving and preservation that would be of particular interest to national libraries. The developments covered are ones that the National Library of Australia is involved in has particular interest in.

The presentation draws on *What's new in digital preservation*, the quarterly digest available on PADI (<http://www.nla.gov.au/padi>), the digital preservation subject gateway compiled and managed by the National Library of Australia. The quarterly digest is a collaborative effort compiled by the UK Digital Preservation Coalition (DPC) (<http://www.dpconline.org>) and the National Library of Australia. It is highly recommended as an easy way of keeping up to date with international developments relating to digital archiving and preservation. The quarterly digests include news from organisations, information about new projects and recent and forthcoming events, and developments relating to particular topics. Topics highlighted include web archiving, repositories, repository certification, preservation standards and approaches, and tool development.

## Background

The PADI quarterly digests as well as recent surveys and studies, suggest that there is now a high level of awareness by collecting organisations of the issues associated with managing digital resources for long-term access. The level of commitment to taking on this challenge is also growing - however, commitment is not yet translating into action to the same extent.

This is demonstrated by a survey of RLG members undertaken in January 2006 which found that 60% of respondents said web archiving was part of their mission, but that the same percentage did not yet have a "plan of attack." [http://www.ohio.rlg/en/page.php?Page\\_ID=399](http://www.ohio.rlg/en/page.php?Page_ID=399). In late 2005, the Digital Preservation Coalition (DPC) surveyed a wide range of UK organisations in different sectors to gauge their response to accepting responsibility for digital preservation. From the 104 responses received, the survey found that management was committed to digital preservation in 52% of cases but that clear responsibilities for the task were assigned in only 33% of cases, and a strategy for action in only 18 % of cases.<sup>1</sup>

It is likely that the findings from these surveys would reflect the situation in most other countries.

Some of the issues the DPC report identifies that inhibit organisations from making the leap from commitment to action, include:

- Funding to support sustainable programs
- Access to shareable tools and services
- Uncertainty about standards
- Skills development
- Legal framework
- Confusion about where to start

However, it is important to acknowledge that there has been real progress in recent years at the international level with addressing these and other barriers to action. The body of knowledge has grown significantly, recommendations for best practice are evolving, a range of digital preservation programs are in place, and tools designed to aid practical action are emerging. So, while problems remain, the future is optimistic.

A very good overview of developments and the current state of affairs which I recommend to you is the IFLA 2006 publication [Networking for Digital Preservation: current practice in 15 national Libraries](#) by Ingeborg Verheul at the National Library of the Netherlands (the KB). The publication is based on a survey conducted by the KB as part of its contribution to the work of ICABS (the IFLA CDNL Alliance for Bibliographic Standards.) (A PDF versions of the publication is available at <http://www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf> )

In my presentation, I will focus in particular on some recent developments and trends relating to collaborative alliances; web archiving, and digital preservation.

---

<sup>1</sup> *Mind the gap : assessing digital preservation needs in the UK*. Digital Preservation Coalition, 2006.

### **Collaborative alliances**

Collaborative alliances and projects that have been set up in many countries have had a significant impact on the growing level of understanding of the issues associated with digital archiving and preservation and on progress with addressing the issues. Many have been successful in attracting major funding support. Collaboration is of course essential as the problems, costs and expertise required are beyond the resources of any single institution.

Information about many collaborative alliances, including all the major national and international ones, can be found through the *Organisations and websites* listing in PADI (<http://www.nla.gov.au/padi/format/org.html>).

A scan of the work programs of some of these initiatives indicates that topics of current interest include for instance, policy development; cost benefit modelling; research use of web archives; curating specific data types such as emails, raw scientific and geospatial data; preservation metadata; digital repository software and standards; repository certification; and tools for specific preservation tasks.

#### International Internet preservation Consortium (IIPC)

A collaborative alliance that is now well known to national libraries which was set up with an action-based agenda is the IIPC – the International Internet Preservation Consortium (<http://www.netpreserve.org>).

The IIPC was established in July 2003 for an initial three year period which will be extended until the end of 2006. The main goals of the Consortium are to:

- encourage and support national libraries to address Internet archiving and preservation; and to
- foster the development and use of common tools, techniques and standards that enable the creation of web archives.

From the outset it had a practical focus with members expected to contribute to working groups and projects. In this respect it was different from most other consortia interested in digital resources. Membership comprises 12 institutions (national libraries and the Internet Archive (<http://www.archive.org>)), all of whom had some experience with archiving web sites and publications. The National Library of Australia is a member and we have benefited considerably from active participation.

In the lead up to the completion of phase 1 of the IIPC this year, its Steering Committee reviewed the achievements and future directions of the consortium. There is support for the IIPC continuing but with a new membership model so more institutions can participate. The model is still being considered by the Steering Committee which will meet again in September this year to finalise governance and membership details, as well as the work plan for 2007-2010.

During its first 3 years, the IIPC has delivered tangible outputs that will assist national libraries to collect and manage digital resources. These include the following:

#### *Standards*

- the WARC (Web ARchive) file format, a standard for storing data that has been harvested during crawls of the web, in large compressed files. (The WARC file format has been submitted for certification as an ISO standard);

## Tools

- Heritrix, web crawler software designed to scale to whole national domains and to gather data in an archival format;
- DeepArc, a desktop software tool for exporting the content from publisher-supplied databases into XML for archiving;
- Xinq, a tool for searching and browsing archived databases;
- NutchWax, software which indexes web archives using the Lucene indexing engine and provides a web based user interface; and
- WERA, a search application which allows users to navigate the contents of a web archive.

Another important tool for national libraries that is being developed under the auspices of the IIPC (but managed by the National Library of New Zealand with funding assistance from the British Library) is the Web Curator Tool. This is intended to help collecting institutions with limited technical support available, to manage the activities associated with collecting web content including selection, description, permissions, harvesting and quality review. The first alpha release of the Curator Tool is due later this year.

The IIPC work plan for 2007-210 includes projects to improve the effectiveness of the harvesting and access tools already developed by making them “smarter”, to package the existing tools with simple installation procedures and good documentation, and to develop standards to support searching and navigation across multiple archives.

## Web archiving

Web archiving by collecting institutions has really taken off in the last couple of years. Many institutions are now doing routine archiving or else experimenting with different approaches. The PADI section on digital archiving provides an overview of activity.

There are 3 main approaches to web archiving – thematic (on a particular subject or topic); selective (representative of a number of subject areas and types of resources); and whole domain. The trend is towards thematic and whole domain archiving, with some institutions taking both approaches. The selective approach is not as common, probably because it is quite resource intensive and obviously very limited in scope. However, its advantage for collecting institutions is that it can provide a quality assessed collection of resources on a range of topics that are deemed to have documentary heritage significance and to which enhanced access is provided by way of descriptive metadata.

The UK Web Archiving Consortium (UKWAC) (<http://www.webarchive.org.uk>) and the National Library of Australia's PANDORA web archive (<http://www.pandora.nla.gov.au>) are examples of the selective approach.

The thematic approach is followed by the Library of Congress who in May this year launched their Web Capture website (<http://www.loc.gov/webcapture/>). This explains their approach to capturing historically important web sites and provides information about the thematic web collections they have built so far. Topics covered include Unites States National Elections, the Iraq war, the events of September 11 and the Winter Olympics 2002.

While Australia has been archiving web resources selectively since 1996, we are now moving to a dual approach – annual whole domain harvests supplemented by selective archiving. We commissioned the Internet Archive to carry out our first harvest of the Australian domain in July last year and they will do a second harvest for us in August this year. We will continue to

select important resources for archiving in the PANDORA Archive but our policy on the scope of this selective activity will change once we are able to provide public access to our Whole Domain harvests. (Legal deposit restrictions currently prevent us from doing this.) We intend to carry out an annual harvest and will work out how to integrate access with PANDORA. The Internet Archive offers collecting institutions a very efficient and (for us at least) cost-effective approach to collecting web resources. Under our arrangement with them, as well as conducting the harvest they indexed the resources gathered and delivered and installed the archive at our Library.

Several other national collecting institutions are now routinely carrying out whole domain harvests or experimenting in this area. Some recent developments follow.

- In April this year, the University of Lisbon announced the availability of their new web archive called Tomba which contains the resources crawled over the last 4 years. A prototype is available at <http://www.tomba.tomba.pt>.
- In July last year, following revisions to legal deposit laws that gave them authority to collect and preserve the Danish Internet, the national libraries in Denmark initiated a three-pronged strategy to collect web resources. This involves whole domain harvesting four times a year; selective harvesting of approximately 80 domains with greater frequency; and event harvesting of 2 or 3 events annually. A very interesting paper outlining the approach was issued early this year. ([http://www.netarkivet.dk/publikationer/DFreyv\\_english.pdf](http://www.netarkivet.dk/publikationer/DFreyv_english.pdf)).
- A new web archiving initiative in Croatia called the DAMP Project which involves the university and national libraries was announced late last year. DAMP is developing system architecture needed to support harvesting and managing web resources. A report on the project is available at [http://widwisawn.cdlr.strath.ac.uk/Issues/Vol3/issue3\\_3\\_1.html](http://widwisawn.cdlr.strath.ac.uk/Issues/Vol3/issue3_3_1.html).

As well as carrying out whole domain harvesting for institutions, the Internet Archive announced a new service earlier this year that makes it possible for organisations to carry out small scale, subject focused archiving without having to install their own management system. This new subscription service is called Archive-It (<http://www.archive-it.org>) Archive-It, allows institutions to build, catalogue and search their own web archive through a user friendly web application accessible via the Internet Archive site. The sites selected are stored by the Internet Archive.

RLG Web Archiving Program ([http://www.ohio.rlg.org/en/page.php?Page\\_ID=399](http://www.ohio.rlg.org/en/page.php?Page_ID=399))

Archive-It is being offered by RLG to its members as part of its Web Archiving Program which was also announced earlier this year. This Program aims to “demystify and simplify the process of web archiving” through a number of initiatives involving working groups which will consider:

- procedures for collaborative collection development aimed at avoiding duplication of effort;
- descriptive metadata requirements for easy searching and browsing;
- user interaction with web archives and their needs and expectations; and
- IP concerns.

RLG also plans to put together information on web archiving options covering software, and the organisational and technical resources required to undertake web archiving.

### Legal deposit

An issue for national libraries that is closely linked to web archiving is legal deposit. There has been pleasing progress in recent years with the recognition by governments of the importance of extending legal deposit provisions to electronic resources. Several national libraries now have legislation that applies to electronic resources in some form and others (according to the IFLA publication Networking for Digital Preservation mentioned earlier); expect legislation to come into effect this year or next. The Deutsche Bibliothek is the latest national library to announce amendment to its legislation, which took effect in June this year.

For those wanting an up-to-date overview of the issues related to the building, management and usage of web archives, the *New Review of Hypermedia and Multimedia* journal will publish a special issue on web archiving next year. It will include case studies and examples of new research and approaches being pursued.

### **Digital preservation**

The significant growth over the last couple of years in knowledge and experience of collecting and managing digital resources has been accompanied by an ever bigger escalation in research into digital preservation issues. (By digital preservation I mean the actions required to ensure that the essential characteristics of data are maintained and that data can be used into the future.)

Research in this area is now on a much larger scale and better coordinated than a few years ago and general agreement is starting to emerge on the nature of the problems that need to be addressed and on approaches to dealing with them.

### National Digital Information Infrastructure and Preservation Program (NDIIPP)

One of the biggest and best known digital preservation projects underway is the National Digital Information Infrastructure and Preservation Program (NDIIPP) (<http://www.digitalpreservation.gov/about/index.html>) which is being led by the Library of Congress. NDIIIP is a collaborative strategy involving government and non-government entities that received just under \$100m funding from the US Congress in late 2000. Its broad goal is to provide a national focus on important policy, standards and technical components necessary to preserve digital content. The NDIIPP plan will be implemented over several years and result in recommendations to the U.S. Congress about the most viable and sustainable options for long-term preservation.

In the last two years several partnerships have been entered into to pursue the program goals. For instance, in May 2005 NDIIPP, in partnership with the National Science Foundation, launched a digital preservation research grants program by awarding 10 universities a total of \$3 million to undertake pioneering research to support long-term management of digital information.

### DigitalPreservationEurope (DPE)

A new European digital preservation initiative that was established in April this year is DigitalPreservationEurope (DPE). DPE "fosters collaboration and synergies between many existing national initiatives across the European Research Area." It "addresses the need to improve coordination, cooperation and consistency in current activities to secure effective preservation of digital materials." (<http://www.digitalpreservationeurope.eu/>)

### Digital Institutional Repositories

The development of digital repositories emerged as a new strategy to manage changes in scholarly communication at the start of 2000. After some tentative experiments with implementation in a couple of universities they have now become very popular. With the rapid increase in the amount of digital content being produced and the availability of affordable open source software, the implementation of digital repository systems is now a priority for most collecting and educational institutions.

There is also increasing recognition that there is a difference between digital repositories and sustainable digital repositories – i.e., repositories that are capable of managing and providing access to meaningful data over time. This has triggered focused research in the last 2 years in particular into the architecture, capabilities and standards required to support sustainable repositories.

### *Australian Partnership in Sustainable Repositories (APSR) Project*

An example of research in this area is the collaborative Australian Partnership in Sustainable Repositories (APSR) Project (<http://www.apsr.edu.au>) which is funded by the Federal government as part of a national information infrastructure strategy for higher education.

APSR aims to increase the awareness and understanding of the risk environment in which repositories operate, to identify or develop tools and systems to deal with main risks, and to encourage or mandate widespread practices that will increase the likelihood that both the content and the responsibility for its management can be sustained.

Its work is based on repository facilities already within partner institutions that are under development. These are used as test bed projects within APSR.

The National Library is a member of APSR with particular responsibility for identifying sustainability issues and leading discussion on how they may be dealt with. For example, we have been working on assessment of the risks these repositories have managed within the test bed projects; specification of preservation metadata standards; and the development of an automated risk identification system.

The APSR project recognises that modular solutions rather than stand alone ones are required. It is making links with related projects such as the Global Digital Format Registry (<http://hul.harvard.edu/gdfr/>) so that appropriate tools can be plugged into its architecture, and replaced if necessary without destroying the entire system.

### Certification of digital repositories

An issue related to digital repositories which has emerged as a core digital preservation topic of current interest, is the audit and certification of digital repositories. As the number of repositories grows there is a need to be able to apply agreed measures of trustworthiness to distinguish those that are sustainable or that have the capability of preserving digital content overtime.

Work in this field has been undertaken over several years by the RLG-National Archives and Records Administration (NARA) Digital Repository and Certification Taskforce ([http://www.rlg.org/en/page.php?Page\\_ID=367](http://www.rlg.org/en/page.php?Page_ID=367)) which was set up to develop audit criteria for digital repositories and archives.

At the end of August 2005, the task force released its audit checklist as a draft for public comment. Also in 2005, an 18 month project to develop the processes and activities required to actually audit and certify repositories was commenced. This project is being undertaken by the Centre for Research Libraries. It will test the RLG-NARA metrics through actual audits of subject digital archives and will refine and deliver specifications for the auditing processes, develop a plan for certification, and will outline a business model for certifying agencies.

As a result of the work of the Digital Repository and Certification Taskforce, repositories will be able to self evaluate and to benchmark their performance using an agreed methodology and audit criteria.

## PREMIS (Preservation Metadata: Implementation Strategies)

To manage digital content for long-term access, it is essential to use metadata to document the provenance, the characteristics and preservation history of the data. A major contribution to this core area of digital preservation has been achieved through the PREMIS Working Group (<http://www.oclc.org/research/projects/pmwg/>).

The PREMIS Working Group was established in 2003 and concluded its activities in May 2005. It was jointly sponsored by OCLC and RLG.

The group was composed of international experts from a variety of domains interested in digital preservation. Its objectives were to:

- develop a core preservation metadata set with broad applicability across the digital preservation community; and
- Identify and evaluate alternative strategies for encoding, storing, and managing preservation metadata in digital preservation systems.

The final report of the PREMIS Working Group is available from the project's website (see above.) Its products, which include a data dictionary for preservation metadata and examples of use cases, and XML schema for the elements in the data dictionary, are available for downloading. A maintenance process for the data dictionary and XML schema, hosted by the Library of Congress, has been organized.

As a result of the work of the PREMIS Working Group we now have an international, open-sourced standard for preservation metadata. In recognition of the important contribution PREMIS has made to the development of effective digital preservation solutions, the Working Group was awarded the prestigious Digital Preservation Award for 2005 which is funded by the UK Digital Preservation Coalition.

## **Conclusion**

There are many developments in digital archiving and preservation underway that will assist national libraries in carrying out their documentary heritage responsibilities for information in digital form. While some fundamental issues remain for most collecting institutions, especially relating to funding, skills development and legal frameworks, shareable technical systems and solutions are emerging that will make it possible for more institutions to move forward from commitment to action.

Contact: Pam Gatenby  
Assistant Director General: Collection Management Division  
National Library of Australia