

Collaboration Strategies for Digital Collections: The Australian Experience

International Conference on Libraries

Leading the Global Knowledge and Information Society

Seoul, Korea, 25-26 May 2009

Warwick Cathro
Assistant Director-General,
Resource Sharing and Innovation

Introduction

This paper will identify and discuss a range of opportunities for collaboration that are available to institutions which are developing, managing and providing access to significant digital collections. The modes of collaboration discussed in this paper reflect the recent experience of the National Library of Australia in cooperating, on both a national and an international level, with other institutions. Although I will be using the work of the National Library of Australia as an example, I will discuss collaboration opportunities which have widespread applicability.

But first, I wish to congratulate the National Library of Korea for its enterprise in developing the National Digital Library of Korea, and for organising this prestigious international conference.

On 5 March this year, the President of Korea and the Prime Minister of Australia met in Canberra. They agreed to declare the year 2011 as a “Korea-Australia Friendship Year”, because that year will mark the 50th anniversary of diplomatic ties between our two nations. On the same day, Korea and Australia entered into an agreement on cooperation in the field of cultural heritage. As part of that agreement, our countries will explore cooperation through joint research and seminars, and exchange of experts.

So, from Australia I bring you warm greetings and our best wishes for the successful development of your National Digital Library.

Setting the scene

At the National Library of Australia, we are building and managing a set of rapidly growing and complex digital collections¹. In April 2009, these collections comprised about 400 terabytes of data. They include archive copies of Australian web sites; digitised copies of historic Australian newspapers; digitised copies of oral history and other audio files;

photographs, e-mails and other items which were “born digital”; and digitised copies of analogue collection items such as pictures, music scores, maps, books and manuscripts.

Since we began to build our digital collections, we have found many opportunities to collaborate with institutions that are facing similar challenges, and with other stakeholders. These potential partners have included overseas national libraries, other Australian cultural institutions, university libraries, agencies that are developing relevant software, and standards bodies.

In a paper published last year Martha Anderson analysed the types of collaboration that may occur between stakeholders in digital collection communities¹. She called these the “components of a stewardship network”.

She identified four such components:

- Content custodians (such as research libraries, major archives, universities that maintain repositories of research outputs, and data archive centres) who are committed to long term preservation, including tackling the problem of obsolescence
- Communities of practice and information exchange (standards bodies, digital preservation experts, relevant professional associations)
- Providers of services (such as infrastructure providers, software developers, registry services, identifier resolution services)
- Capacity building organisations (funders of research, curricula developers, etc).

The collaborative opportunities discussed in this paper relate to the first three of these components.

In a recent book contribution, this author noted that collaborative activities can influence the economics of digital collections in several ways¹. For example, collaboration to use shared technology infrastructure can assist in reducing the costs of developing and maintaining digital collections and their related services. Collaborative activities can also increase the value of digital libraries from the user perspective.

That contribution cited a number of examples of relevant collaborative activities in Australia, including:

- newspaper digitisation;
- web archiving;
- collaborative projects resulting from investment in research infrastructure; and
- sharing of software components through the open source model.

Each of these examples will be discussed further below.

This paper will not attempt to list all of the major national and international collaborative activities that are taking place in the digital collecting field. An extensive survey of digital repository developments, preservation strategies and collaborative initiatives in 15 countries

was published a few years ago¹. However, some of the major developments, in cases where there is significant Australian involvement, will be referred to below.

Collaboration in collecting

During 2008, three Australian collecting institutions (the National Library, the National Archives and the National Film and Sound Archive) collaborated on a joint bid for additional resources. In the process they developed a joint business case for funding to enable them to deal with four challenges:

- *digital collecting*: the challenge of collecting and storing petabyte-level collections of “born digital” content
- *digital preservation*: the challenge of preserving digital content for long term access in the face of technical obsolescence
- *audiovisual obsolescence*: the challenge of migrating very large audiovisual collections to digital format to rescue them from obsolescence which will render them inaccessible
- *digital access*: the challenge of converting traditional content into digital form, and delivering digital content to make it easily accessible to the Australian people.

The agencies entitled their bid “Dealing with the Digital Deluge”.

As part of the business case, the three agencies identified a number of opportunities for closer collaboration, which are discussed further below.

In the sections below, opportunities for collaboration will be identified in the following areas associated with digital collecting:

- web archiving;
- managing digital resources on physical carriers;
- digitisation
- collection backup and disaster recovery; and
- national frameworks.

Web archiving

A major international conference on “Archiving Web Resources” was held at the National Library of Australia five years ago¹. At this conference, the theme of cooperation was central. Many aspects of cooperation were discussed, including:

- sharing information;
- pursuing co-operative projects; and
- developing and using common tools and standards.

A key conclusion of the Conference was

The task is too large for individual institutions to undertake in isolation and the resources required for successful and sustained archiving are too great to make duplication of effort a tenable position.

Australia was an early implementer of web archiving. Since 1996 the National Library of Australia has been developing and maintaining PANDORA, an archive of selected, significant Australian web sites and web-based online publications⁶. The purpose of PANDORA is to ensure that Australians of the future will be able to access a significant component of today's Australian web based information resources.

Because of the high cost of selective web archiving, it makes sense for one agency (such as a national library) to develop both the expertise and the infrastructure for web archiving, and for other agencies to leverage off this investment. Accordingly, PANDORA is a collaborative activity, as the archive is being built by the Australian state libraries and some other cultural institutions in addition to the National Library. This collaboration involves a shared software and database platform, and agreement on non-overlapping collection responsibilities. This collaborative activity provides an example of Martha Anderson's first mode of collaboration (between content custodians).

Since 2005, the National Library has also been undertaking an annual large scale harvest of the Australian web domain⁷. This activity was the result of a long-standing aspiration to complement the selective PANDORA approach with a whole domain approach. The Library contracted the Internet Archive to undertake the whole domain harvest. Effectively, the Internet Archive acts as both a supplier and partner to the National Library. This provides an example of Martha Anderson's third mode of collaboration (with providers of services).

There are prospects of widening the collaborative aspects of this whole domain approach. For example, as part of the "Digital Deluge" bid, the National Library of Australia and the National Archives of Australia identified a scenario under which the National Library would copy or transfer the ".gov.au" component of its whole domain web harvest to the National Archives so that it will form part of the archival resources of the Australian government.

The National Library of Australia, along with the National Library of Korea and 26 other institutions, is a member of the International Internet Preservation Consortium (IIPC)⁸. The IIPC is fostering the development and use of common tools, techniques and standards for the creation of web archives.

Another example of collaborative web archiving activity is the UK Web Archiving Consortium, which aims to build a corpus of websites selected by leading institutions in the United Kingdom for their historical, social and cultural significance⁹.

Managing digital resources on physical carriers

Digital collection content can also be created by ingesting born-digital files received on physical carriers such as USB drives, floppy disks, DVDs and CD ROMs.

The National Library of Australia has a small but important collection of such content. To deal with it, the Library developed a software application called Prometheus, which provides

a semi-automated, scalable process for transferring data from physical carriers to preservation-managed digital storage¹⁰. This software, together with the use of a customisable 'mini-jukebox', allows Library staff to copy the content from a wide range of carriers. Once the content is copied, integrity and virus checking is conducted and as much metadata as possible is harvested.

The development of such tools creates another opportunity for collaboration. It makes little sense for every agency to build and maintain its own equivalent of Prometheus. Instead, it is an attractive option for one agency to provide such a solution to a group of partners, particularly where these partners do not receive enough physical carriers to justify their own investment in the tool. A joint ingest facility could be used by all of the partners in the collaboration.

Digitisation

Another significant workflow for creating digital collection content is digitisation. Many libraries and cultural institutions have established their own in-house digitisation activities, or have outsourced all or some of the process.

However, some digitisation processes are highly specialised and require a major investment in equipment, workflow design, staff training, engagement of suppliers, and software development. A good example is the digitisation of newspapers. The National Library of Australia has made a major investment during the past three years in its Newspaper Digitisation Program¹¹. An investment on this scale is difficult for many libraries to justify. For this reason, the Australian state and territory libraries have partnered with the National Library to leverage off its investment. Where these libraries can generate funding to digitise titles (such as regional newspapers) that will not otherwise be digitised, these funds can be used to process the additional titles through the National Library's production process, and to mount the titles on the National Library's database.

Another potential area for collaboration in digitisation is the conversion of analogue audio content to digital form. Such collaboration formed part of the "Digital Deluge" funding bid, discussed above. In economic terms, it is an attractive option for one agency to generate the capability for undertaking audio conversion and to supply this capability to the other agencies.

Collection backup and disaster recovery

Digital collections, such as those created by the processes described above, must be protected against loss of data. To safeguard their collections, institutions need to maintain multiple copies of their digital collection items, with at least one copy held offsite.

One way of supporting this need is for institutions to collaborate on off-site disaster recovery facilities. For example, it is possible for institutions to act as offsite storage sites for each other. This opportunity was actively discussed as part of the "Digital Deluge" funding bid. However, such collaboration must deal with the challenge created by the combination of

required response times for data restoration, the terabyte-levels of data, and the cost of providing high bandwidth access to a backup facility located at some distance.

National frameworks

Some countries are attempting to develop a national approach to digital collection management. For example, the New Zealand Digital Strategy, launched in May 2005, includes a Digital Content Strategy which recognises the importance of managing digital content and digitising New Zealand's existing collections. The funding for this Strategy has included more than NZ\$30 million over four years to the National Library of New Zealand to build a national digital heritage archive and to coordinate national collection digitisation projects¹².

In Australia and New Zealand, the state, territory and national libraries are collaborating on a major program of work which they are calling "Reimagining Library Services". This activity will include a project to enhance the capabilities of all 10 partner libraries in creating, managing and curating their digital collections¹³.

Collaboration in digital preservation

It has been recognised for many years that digital collections are at risk of becoming inaccessible over the long term, due to continuous changes in hardware, software and standards, leading to obsolescence. Collecting institutions therefore face significant challenges in developing the strategies and tools to preserve long term access. To date, such institutions have been "learning by doing". In a sense, they are undertaking continuous research in digital preservation.

The issues involved in this challenge of long term sustainability have been described by Kevin Bradley of the National Library of Australia. This review of sustainability issues was compiled as part of the Australian Partnership for Sustainable Repositories (APSR) which was funded by the Australian Government between 2004 and 2007 as part of its program to improve higher education and research infrastructure¹⁴.

Many activities are in progress around the world in an effort to address these issues. We have already noted the work of the International Internet Preservation Consortium (IIPC). Other notable activities include:

- PLANETS, a project funded by the European Union to address digital preservation challenges (in part by building services and tools) and involving four national libraries, three national archives, five universities and two major IT corporations;
- the Digital Preservation Coalition and the Digital Curation Centre in the United Kingdom;
- the National Digital Information Infrastructure and Preservation Program (NDIIPP) in the United States; and
- Nestor, which aims to create a network for information and communication about digital preservation activities in Germany.

In Australia, some institutions have worked on the development of appropriate techniques and software tools. For example, the National Archives of Australia has developed software known as Xena (XML Electronic Normalisation of Archives)¹⁵, which converts newly acquired digital records into “open” formats, as well as tools to export to original formats and access converted data in the way it was originally presented.

In 2007, the National Library of Australia collaborated with the Australian National University, as part of the APSR Project, to develop a software tool called AONS (Automated Obsolescence Notification Service). This tool allows repository managers to automatically monitor the status of file formats in their repositories, to make risk assessments based on a standard set of questions, and to receive notifications when file format risks change¹⁶. The toolkit depends on there being useful data in external registries such as the Library of Congress Sustainability of Digital Formats Registry, and PRONOM, supported by the UK National Archives.

A number of Australian Government agencies have formed a working group known as MAGDIR (Managing Australian Government Digital Information Resources), which is exploring how to progress the following objective: *Australian Government information resources in digital form are controlled, managed and preserved so as to permit their ongoing and reliable use by government and the community for as long as needed*¹⁷.

These agencies include the National Library of Australia, the National Archives of Australia, the National Film and Sound Archive, the Australian Bureau of Statistics and GeoScience Australia. In the context of the “Digital Deluge” funding bid, the first three of these agencies proposed to establish a joint Digital Preservation Taskforce, which would develop and share innovative approaches, tools and processes for preserving and providing access to born digital and digitised analogue content.

A significant component of the digital collections of libraries is represented by the electronic journals and databases which are licensed from vendors. It is important that access to the backsets of these electronic resources is preserved. The initiative known as CLOCKSS (Controlled Lots of Copies Keep Stuff Safe) has been established to meet this challenge¹⁸. CLOCKSS is an international program under which scholarly publishers and research libraries are cooperating to build a sustainable, distributed archive to ensure the long-term survival of Web-based scholarly publications. The Australian National University is one of the governing libraries for the CLOCKSS program.

Collaboration in access

Individual institutions normally provide access to their collections through a catalogue or local portal. From the user’s perspective, it is highly desirable to be able to search across multiple collections so that users do not need to perform the same search many times in many catalogues.

The content of many collections will be indexed in Google, but with two significant drawbacks:

- even with well constructed site maps, digital collections are only partially indexed by Google; and
- Google's relevance ranking will not necessarily give preference to content from collecting institutions.

Over the past decade, the National Library of Australia has constructed eight separate "national portals" for discovery of Australian collections. An example is Picture Australia, which provides a single point of discovery for over 1.6 million pictures that have been digitised by Australian collecting institutions (libraries, museums, archives and other institutions).

The National Library is now in the process of integrating these services into a single national portal through its "Single Business Discovery Project"¹⁹.

A key part of this activity will be to build and extend partnerships with potential contributors of content and metadata for this national portal. The data that will be contributed or harvested will include:

- metadata found in Australian library catalogues
- the full text of books, journals and newspapers digitised by Australian collecting institutions;
- the full text of the Australian web archives;
- metadata from Australian universities and research data centres;
- metadata describing pictures digitised by libraries, museums, archives and galleries;
- metadata for journal articles that are collected or licensed by libraries, whether in print or electronic form; and
- finding aids and summaries for manuscript and archive collections, and for oral history interviews.

Collaboration in software

Of all of the factors involved in building digital collection services, software is usually the most important and the most challenging. Well designed and robust software is an essential foundation for these services. However, the software provided by vendors often fails to meet the rapidly developing needs of the collecting institutions.

The result is that the needs of collecting institutions are often met by a mixture of:

- vendor-supplied software;
- open source software; and
- in-house developed software.

In terms of the activities of the National Library of Australia, we have already seen one example of open source software development, in the AONS software developed under the auspices of the Australian Partnership for Sustainable Repositories¹⁶. This was an example of

the National Library creating open source software. There are other cases where the Library has used or is trialling open source software developed by others. Examples are the VuFind software for supporting library catalogues²⁰, and the Archivists Toolkit for managing manuscript and archive collections²¹.

Last year the Library was invited to join the Open Library Environment (OLE) Project, which has been funded by a grant from the Mellon Foundation. The OLE Project is led by Duke University, located in North Carolina. The goal of the Project is to define requirements for an open source library management system, based on a re-examined model of library operations. It aims to develop, by July 2009, a design for a next-generation library system using open source software, and a community of interest that could be tapped to help build this system²².

The OLE Project has adopted the framework of Service Oriented Architecture, under which a number of software modules inter-operate through the use of standard protocols and Application Programming Interfaces. The National Library of Australia is also working in this space, and has developed its own digital library service framework. The Library has announced its intention of working with other interested parties to develop further this framework as a collaborative activity²³.

It is also quite common for major software collaborations to occur between collecting institutions and software vendors. A notable example is the development by Ex Libris, in partnership with the National Library of New Zealand, of the Rosetta digital preservation module, which now forms one of the Ex Libris family of library system products.

Ex Libris has recently announced that it will pursue an "Open Platform Strategy"²⁴. This development should open up a wide range of possibilities for libraries to partner with that company in the use of Application Programming Interfaces and Service Oriented Architecture to integrate open source and in-house software modules with the vendor's product, creating the potential for a more flexible and innovative software regime for the collecting institution.

Collaboration and standards

Successful digital collection services depend on inter-operable data and inter-operable software. In both areas, achievement of inter-operability depends on the definition of technical standards, the achievement of consensus about these standards, and on a suitable long term maintenance arrangement for the standards.

In the case of data inter-operability, relevant standards include MARC (Machine-Readable Cataloguing), MODS (Metadata Object Description Schema), METS (Metadata Encoding and Transmission Standard) and PREMIS (Preservation Metadata – Implementation Strategies) all of which are maintained by the Library of Congress, and Dublin Core, which is maintained by the Dublin Core Metadata Initiative, now a company incorporated in Singapore.

The National Library of Australia participated in the original development of PREMIS and of Dublin Core. During 2006, as part of the APSR Project, the National Library defined an exchange format for repository content (including the PREMIS metadata associated with an

object) in a standard way using METS, and tested the profile by transferring digital objects between two university repositories that were using different platforms (DSpace and Fedora)²⁵.

In the case of software inter-operability, relevant standards include the NISO Circulation Interchange Protocol (NCIP), the Information Retrieval Protocols (Z39.50, SRU), and the Open Archives Initiative Protocol for Metadata Harvesting (OAI). The National Library of Australia has participated actively in the standards committees that monitor such protocols and that encourage their implementation. The National Library also participates in ICADS, the IFLA-CDNL Alliance for Digital Strategies, a partnership of six national libraries which is exchanging information about digital collection issues²⁶.

Concluding remarks

This paper has described a wide range of collaboration activities which are possible for the institutions that are building and managing digital collections. Collaboration requires effort, and a willingness to examine services from a perspective which does not place one's own institution at the centre.

Collaboration, of course, is not an end in itself. Such collaborative activities must to be motivated by user needs, and they should lead either to more content being available for users, to improved user access pathways, or to preservation of content for future users.

Collaboration is often encouraged by the funding of projects which have a limited life. Even where such projects are highly successful, a sustainability strategy is needed to ensure that the benefits of collaboration are not lost.

This raises the question of who should drive collaboration, and who should take on the responsibility for ongoing delivery of collaborative services. From the perspective of this author, national institutions – such as national libraries and national archives – are well placed to provide this leadership, because of their strong legislative mandates, and their inherent longevity.

References

1. National Library of Australia. Explore the National Library's digital collections. <http://www.nla.gov.au/digicoll/>
2. Anderson, Martha. Evolving a network of networks: the experience of partnerships in the National Digital Information Infrastructure and Preservation Program. *International Journal of Digital Curation*, Issue 1, Volume 3, 2008.
3. Cathro, Warwick. The Australian perspective. In *Digital library economics: an academic perspective*. Chandos, 2009. ISBN 1 84334 403 3.

4. Networking for digital preservation: current practice in 15 National libraries. (IFLA publications 119). Munchen : Saur, 2006. <http://www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf>
5. Archiving web resources: issues for cultural heritage institutions. Canberra, 9-11 November 2004. Conference report. <http://www.nla.gov.au/webarchiving/index.html>
6. Phillips, Margaret E. and Koerbin, Paul. PANDORA, Australia's Web Archive: how much metadata is enough? *Journal of Internet Cataloguing*, v. 7, issue 2 (2007).
7. Gatenby, Pam. Recent developments in digital archiving and preservation. Paper prepared for the CDNL meeting, Seoul, August 2006. http://www.nla.gov.au/nla/staffpaper/2006/documents/pgatenby_CDNL.pdf
8. International Internet Preservation Consortium. <http://netpreserve.org/about/index.php>
9. UK Web Archive. <http://www.webarchive.org.uk/ukwa/>
10. Elford, Douglas [et al]. Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia. Paper presented at the 74th IFLA Conference, Quebec, 2008. <http://www.ifla.org/IV/ifla74/papers/084-Webb-en.pdf>
11. Australian Newspapers Digitisation Program. <http://www.nla.gov.au/ndp/>
12. New Zealand Digital Strategy: smarter through digital. <http://www.digitalstrategy.govt.nz/>
13. National & State Libraries Australasia. Reimagining Library Services. <http://www.nsla.org.au/projects/rls>
14. Bradley, Kevin. APSR sustainability issues discussion paper. http://www.apsr.edu.au/documents/APSR_Sustainability_Issues_Paper.pdf
15. National Archives of Australia. Xena: digital preservation software. <http://xena.sourceforge.net/>
16. Pearson, David. AONS II: continuing the trend towards preservation software 'Nirvana'. Paper presented at iPres2007, Beijing, China, October 11-12, 2007. http://www.apsr.edu.au/aons2/pearson_ipres_2007_text.pdf
17. MAGDIR: managing Australian Government digital information resources. <http://www.nla.gov.au/MAGDIR/>
18. CLOCKSS: a trusted, community-owned archive. <http://www.clockss.org/clockss/Home>
19. National Library of Australia. Single Business Discovery Project. <https://wiki.nla.gov.au/display/LABS/2.+Single+Business+Discovery+Project>
20. Open source and the National Library of Australia catalogue. *Gateways*, No. 92 (April 2008). <http://www.nla.gov.au/pub/gateways/issues/92/story02.html>

21. Introduction to the Archivists' Toolkit. <http://www.archiviststoolkit.org/>
22. OLE Project in Australia. <https://wiki.nla.gov.au/display/GWP/Home>
23. National Library of Australia. Service framework.
<https://wiki.nla.gov.au/display/LABS/3.+Service+framework>
24. The Ex Libris Open Platform Strategy. The Ex Librian newsletter, July 2008.
<http://www.exlibrisgroup.com>
25. Pearce, Judith [et al]. The Australian METS Profile – a journey about metadata. D-Lib Magazine, March/April 2008. <http://www.dlib.org/dlib/march08/pearce/03pearce.html>
26. ICADS (IFLA-CDNL Alliance for Digital Strategies).
<http://www.nla.gov.au/padi/topics/712.html>