COLLABORATION IN BUILDING A SUSTAINABLE REPOSITORY ENVIRONMENT: A NATIONAL LIBRARY'S ROLE

OPEN REPOSITORIES 2008, SOUTHAMPTON, 1-4 APRIL 2008

Warwick Cathro
Assistant Director-General, Innovation
National Library of Australia

Background

This paper is based on the experience of the National Library of Australia during the past four years in collaborating with the higher education sector to advance the development of sustainable cyberinfrastructure.

During that period the evolution of cyberinfrastructure in Australia has been notable for:

- the strong involvement of policy officers from the government department responsible for higher education and research, in shaping that evolution;
- the establishment of repositories in more than 70% of Australia's universities;
- after a slow start, increasingly rapid growth in the content of those repositories;
- the establishment of a national discovery service, now containing more than 90,000 records, which provides a single access point to those repositories;
- the development or commissioning of three Fedora-based software platforms (VITAL, Fez, and Muradora) which, in addition to other software platforms, are being used to support university repositories;
- the attention given to sustainability issues, led by one of the repository projects (the Australian Partnership for Sustainable Repositories); and
- the decision to establish the Australian National Data Service, which is due to commence operations in July 2008.

The National Library of Australia was invited to participate in two of the key repository projects because of its experience in discovery services, digital preservation and standards activities. This provided a valuable opportunity for the National Library to collaborate with university partners and to work with them to produce tangible outcomes (Cathro, 2006).

Since the change of government in Australia in late November 2007, policy responsibility for research infrastructure has been vested in the new Department of Innovation, Industry, Science and Research. The key funding initiative, the National Collaborative Research Infrastructure Strategy (NCRIS) will be maintained by the new government. NCRIS will support a range of disciplinary and systemic infrastructure investments and ongoing services. Among these is the planned Australian National Data Service (ANDS).

The Australian National Data Service

The scope of ANDS has been defined in a recent policy paper (Towards the Australian Data Commons, 2007). The projected funding for ANDS from the NCRIS initiative is a relatively modest A$7M (£3M) for each of the three financial years ending in June 2011.

ANDS will facilitate the curation, discovery and use of data created by Australian research through its four proposed programs:

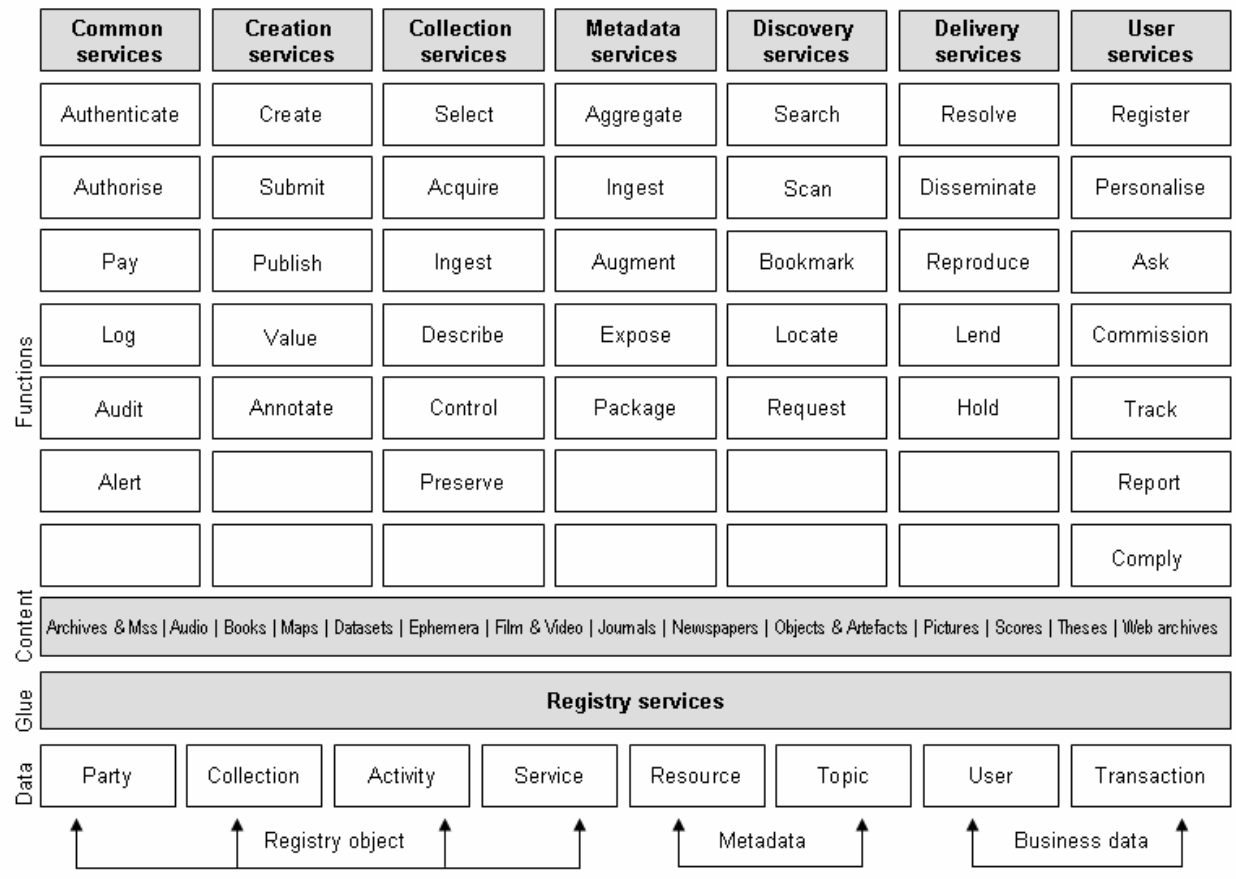
- A **framework program** will develop the policies for the network of research repositories
- A **utilities program** will provide services that support access and use of data, including a persistent identifier service, a discovery service and a registry service
- A **repositories program** will help repository managers in their task of ingesting and sustainably managing Australian research data
- A **research practice program** will help Australian researchers and institutions to develop the necessary skills to create, manage, and share the research data.

ANDS will subsume, extend and place on a recurrent footing the work of several projects that were funded in the period 2004-2007, including:

- ARROW (Australian Research Repositories Online to the World), led by Monash University;
- APSR (Australian Partnership for Sustainable Repositories), led by the Australian National University; and
- PILIN (Persistent Identifier Linking Infrastructure), which reported to the University of Southern Queensland and the ARROW Project.

Focus of this paper

There are many issues involved in managing and accessing digital repositories, and this paper necessarily deals with only a subset of them. In order to establish a planning context for digital library services, the National Library of Australia has developed a Service Framework which it has mapped to other SOA (Service Oriented Architecture) frameworks such as the E-Framework for Education and Research. This work has been described in a publicly accessible wiki (National Library of Australia. Library Labs wiki). For each service the NLA framework provides a definition, use cases, and related protocols and schemas. The framework describes 39 services, organised into seven service groups, as follows:



The services dealt with in this paper are:

- The Ingest and Preserve Functions in the Collection Services column
- The Aggregate and Expose services in the Metadata Services column
- The Discovery Services column
- The Resolve and Disseminate services in the Delivery Services Column
- The Registry Services layer.

"Discovery Services" and "Metadata Services"

There is an obvious need for services which allow the outputs of research, whether they are data sets, electronic publications or print publications, to be discovered by other researchers and the general public. There will always be many discovery services, some discipline based, some provided in a library context, and of course public search engines such as Google.

As well as supporting Search (through standard protocols such as OpenSearch and SRU) and the other services in our Discovery Services column, these services often function by aggregating metadata, through standard protocols such as OAI-PMH and SRU Update. In turn they often expose metadata through the same protocols.

These services may function merely as mechanisms to assist Google by pre-aggregating metadata from individual repositories or library catalogues, but often they have their own search interfaces which may be designed to give greater prominence to particular data or document collections in the relevance ranking than would be given by Google.

Like a number of other countries, Australia has a national discovery service for research outputs. Because this service was sponsored by the ARROW Project it is currently branded as the ARROW Discovery Service. It was developed by (and is currently delivered by) the National Library, one of the ARROW Project partners. The ARROW Project asked the National Library to undertake this role because of the Library's track record in operating other national federated discovery services such as Libraries Australia, Picture Australia and Music Australia.

The ARROW Discovery Service was recently redeveloped using a Solr back-end and a Lucene search platform, in order to improve its performance and to support relevance ranking and faceted search refinement. It now provides a single point of access to about 100,000 research outputs harvested from 28 university and related repositories.

The ARROW Discovery Service is not intended purely as a native search service. It is set up to *syndicate* the aggregated data using OAI-PMH, and also supports the trio of search protocols (OpenSearch, SRU and Z39.50) so that the metadata can be reused by third parties. And, of course, it facilitates harvesting of the nationally aggregated metadata by Google and Google Scholar.

Looking to the future, we can expect the ARROW Discovery Service to evolve into a wider service supported financially by the ANDS Utilities Program. The future service will need to embrace all research outputs, including data sets, and will need to support authenticated access to some resources using protocols such as Shibboleth and OpenID.

The ANDS Establishment Project will select the provider of the ANDS Discovery Service through a competitive selection process. The National Library may or may not be selected for this role, but at the least it will aim to facilitate access to that service as an "external target" from its own integrated discovery service, which will expose content such as Australian library catalogues, the national digital newspaper collection and the national web archive. More details of this planned discovery service are given in the Library Labs wiki.

The "Registry Services" Layer

Federated discovery services need to know which data collections exist, what search and harvesting protocols they support and how to access them (Pearce, 2005). One way of tackling this problem is to register collections in a central repository of some kind. There is also a benefit in storing access policies in a central registry. Potentially, such registries can support some automated intelligent shaping of discovery and delivery services. For example a discovery service could select particular delivery options for the user after querying the access policies in the registry.

Given this background, the APSR Project in 2007 built a prototype registry which it called *ORCA: Online Research Collections Australia* (Australian Partnership for Sustainable Repositories, 2007). The ORCA Registry contains structured, machine readable descriptions of collections, services, parties and activities. It is intended to serve a similar function as that planned for the JISC Information Environment Service Registry (IESR).

The ORCA Registry was developed by the Australian National University. The National Library's contribution to this project was to lead the work on the international standard (ISO 2146) on which the registry schema is based.

ISO 2146 describes a data element directory, based on an object-oriented data model which has a registry object as its primary object class. A registry object may be a collection, party, activity or service. (A party is a person or organisation which may own a collection. In turn, parties are involved in activities, and provide services, including services supporting access to collections).

ISO 2146 provides a framework for machine readable descriptions of about 150 data elements covering service types, functions, products, access policies, protocol information and other required registry data.

At present, the ORCA Registry is a demonstrator only. On present plans, however, it will develop into a functioning national registry supported financially by the ANDS Utilities Program.

It has not escaped our attention that a registry of this kind has value beyond the higher education and research sector. For example, such a registry could support the effective sharing of government data. It would seem to be sensible not to duplicate or re-invent this kind of infrastructure across different sectors.

The "Resolve Service" and Persistent Identifiers

For the proper functioning of delivery services, we need to ensure that data sets and other research resources are reliably cited, so they can be accessed efficiently through actionable links in discovery services. As we all know, URLs which identify a resource in terms of a current location do not provide a persistent reliable link to the resource.

The use of a persistent identifier, if coupled with good management practices, will ensure that when a resource is moved, or its ownership changes, the links to it will remain actionable. A commonly used persistent identifier system is the Handle System, which is supported by the Corporation for National Research Initiatives (CNRI).

In Australia, most university repositories have implemented the Handle System to identify repository content and to provide the required resolver service.

In 2006 the Australian Government funded the PILIN Project (Persistent Identifier Linking Infrastructure Project). PILIN developed a pilot set of tools and it defined the requirements for a future national operational service (PILIN Project Home Page, 2008). Although PILIN

issued its final report in December 2007, it is remaining in existence until June 2008 to support a transition to the proposed National Persistent Identifier Service (NPIS) which will form part of the ANDS Utility Program.

It is envisaged that the NPIS will operate a Global Handles Mirror and an Australian Handles Proxy Service, and will administer the Australian Handles namespace. It will develop tools to support changes in data custody, and tools to ensure resolution to the "appropriate copy" when there is more than one location for a resource.

It is envisaged that the NPIS will serve the higher education sector initially, with the longer term aim of expanding into a broader national service, taking in sectors such as government, non-profit organisations and cultural institutions.

The ANDS Establishment Project will select the provider of the NPIS through a competitive selection process.

The National Library has had a close relationship with the PILIN Project and has provided advice on the scope of the future NPIS. The Library also has a longstanding interest in persistent identifiers. As long ago as 2000, the Library issued guidelines on persistence (National Library of Australia: Managing Web Resources for Persistent Access). In 2001 the Library developed an identifier naming scheme and a resolver service to support reliable access to its own digital collections (Boston, 2002).

The "Preserve Service" and Obsolescence notification

In the NLA Service Framework, "Preserve", once closely analysed, reduces to a continuum of processes or events. Each event has an input copy (submission information package) and an output copy (dissemination information package). There are uses cases such as "Commission preservation action", "Track preservation event" and "Approve preservation outcomes". The PREMIS event ontology provides a list of the event types.

One of these events is "migration" - a transformation of an object which creates a version in a more contemporary format. An important aspect of repository sustainability is therefore concerned with the need to alert repository managers to the obsolescence of file formats in those repositories.

During 2006 and 2007 the APSR Project undertook a sub-project known as AONS: Automated Obsolescence Notification Service. AONS built on the pioneering work of Jane Hunter and her colleagues in the PANIC Project (Hunter, 2006). The National Library, as a partner in APSR, took a leading role in the sub-project.

AONS was concerned with processes for monitoring and assessing the risks of file format obsolescence. At its conclusion in October 2007, the sub-project delivered a pilot open source toolkit which 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 change (Pearson, 2007).

The toolkit builds a profile of the formats in a repository. The profile can be derived from a repository crawl using purpose-built adaptors for a given repository type (DSpace, Fedora, etc). The crawl results may be obtained from existing repository metadata or from automated format recognition tools (such as DROID and JHOVE), or both.

The toolkit can compare the profile with information derived from external registries, including the Library of Congress Sustainability of Digital Formats Registry, and PRONOM, supported by the UK National Archives. The AONS toolkit is designed so that adaptors for new registries, such as Global Digital Format Registry, can be created with minimal effort.

The usefulness of the AONS tools will improve as the international format registries develop. Currently, these registries do not provide adequate format obsolescence risk metrics, and do not provide format data in a sufficiently structured manner to be used by tools such as AONS without human intervention. The National Library would like to see such registries take account of automated obsolescence notification as an important use case.

Australian METS Profile and PREMIS metadata

In addition to migration, the PREMIS ontology includes events such as ingest, normalization and dissemination. In order to support such events we need an open, standard and extensible way of exchanging objects between repositories, and hence of supporting both submission and dissemination workflows.

Submission workflows could include deposit workflows initiated by creators of content, and capture workflows initiated by a repository. Delivery workflows could include dissemination initiated by the repository's own delivery system, requests for representations of an object for use by another system, and the transfer of content from one repository to another.

During 2006 and 2007 the APSR Project undertook two related sub-projects, concerned with describing an exchange format for repository content (including the preservation metadata associated with an object) in a standard way using METS. Again, the National Library took a leading role in the sub-project.

In addition to the Preserve Service, this work is clearly relevant to the Ingest Service (in the Collection Services column) and the Disseminate Service (in the Delivery Services Column).

The first sub-project concerned the encoding in METS of preservation metadata. While METS does not natively support the encoding of preservation metadata, it is extensible by plugging in other schemas, including the PREMIS schema for preservation metadata. By combining METS and PREMIS data elements in a single METS document, an object can be packaged for submission or dissemination in a way that is completely self-describing.

The first sub-project also developed a profile supporting the transfer of an object from one repository to another. Based on this work, the Australian National University and the University of Queensland were able to implement demonstrators for exchanging content between a DSpace repository and a Fez-Fedora repository. This demonstration proved that METS extended by PREMIS can be used to support the transfer scenario. However, it

exposed the need to test the profile against a range of data content models and submission and dissemination scenarios.

This led to the second sub-project, involving the development of an Australian METS Profile. At the conclusion of the sub-project in December 2007, this Profile was registered with the Library of Congress.

The Australian METS Profile describes a three layer model. At the top level, a generic profile specifies the mandatory elements and attributes, extension schemas and controlled vocabularies applicable to all content models. At the second level, sub-profiles inherit the generic profile and clarify requirements for specific content models. The first of these sub-profiles – the Australian Journals METS Profile – has now been registered. Work is now progressing on content models for digital publications stored on physical media, and for audio content. At the third level is the local registration of implementation profiles as part of a national registry of repositories.

It has occurred to us that the Collection/Service Registry, referred to earlier in this paper, could provide the infrastructure for registering implementation profiles.

Further details of this sub-project are given in an article published this month in D-Lib Magazine (Pearce, 2008).

Conclusions

During the past four years, the National Library of Australia has actively collaborated with the higher education sector to advance the development of sustainable cyberinfrastructure in Australia. The Library was able to play a constructive role due to its experience in discovery services, digital preservation and standards activities. Its key contributions have been:

- its development of a national discovery service which provides a single access point to Australia's university repositories and exposes metadata for harvesting by third parties;
- its development of a rich schema (ISO 2146) to support registries of collections, services, parties and activities;
- its strong support for the goals of the Persistent Identifier Linking Infrastructure (PILIN) Project and its advice on the scope of the future NPIS;
- its leadership of an activity to develop tools to alert repository managers to the obsolescence of file formats; and
- its leadership of the development of an Australian METS Profile.

Acknowledgments

The author's contribution to the above activities has been at the level of broad strategic direction setting. The real work was undertaken by the following people whose ideas and efforts were fundamental to the outcomes:

- ARROW Discovery Service: Debbie Campbell, Alison Dellit, Tor Lattimore, Vinita Tuteja, Joanna Meakins, Lynda Hurley, Aaron Defazio
- ORCA Prototype Registry: Scott Yeadon, James Blanden, Adrian Burton, Chris Blackall (ANU: APSR Project)
- PILIN Project: Dennis Macnamara, Nigel Ward, Kerry Blinco
- AONS: David Pearson, Matthew Walker, David Levy
- Australian METS Profile: David Pearson, Bronwyn Lee, Scott Yeadon, Megan Williams, Gerard Clifton
- Standards development (NLA Service Framework, ISO 2146, and Australian METS Profile): Judith Pearce.

References

Australian Partnership for Sustainable Repositories (2007). ORCA Registry v1.0 documentation. http://www.apsr.edu.au/orca/orca_registry_1_0_documentation.pdf

Boston, Tony (2002). A practical approach to ensuring the persistence of digital collections at the National Library of Australia / by Tony Boston and Ninh Nguyen, May 2002. http://www.nla.gov.au/nla/staffpaper/2002/boston2.html

Cathro, Warwick (2006). The role of a national library in supporting research information infrastructure. IFLA Journal, Vol. 32(4): 333-339 (2006). http://www.ifla.org/V/iflaj/IFLA-Journal-4-2006.pdf

Hunter, Jane (2006). PANIC: an integrated approach to the preservation of composite digital objects using semantic web services / Jane Hunter and Sharmin Choudhury. *International Journal on Digital Libraries* 6(2):174-183. http://eprint.uq.edu.au/archive/00004572/

National Library of Australia. Library Labs wiki. https://wiki.nla.gov.au/display/LABS/Home

National Library of Australia. Managing Web Resources for Persistent Access. http://www.nla.gov.au/guidelines/persistence.html

Pearce, Judith (2005). New frameworks for resource discovery and delivery. http://www.nla.gov.au/nla/staffpaper/2005/pearce1.html

Pearce, Judith (2008). The Australian METS Profile: a journey about metadata / Judith Pearce, David Pearson, Megan Williams and Scott Yeadon. DLib Magazine, March/April 2008. http://www.dlib.org/dlib/march08/pearce/03pearce.html

Pearson, David (2007). 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

PILIN Project Home Page (2008). https://www.pilin.net.au/

Towards the Australian data commons: a proposal for an Australian National Data Service. October 2007.
http://www.pfc.org.au/twiki/pub/Main/Data/TowardstheAustralianDataCommons.pdf