Date : 07/06/2006

The Role of a National Library in Supporting Research Information Infrastructure

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

Meeting:	**155 Information Technology with National Libraries with Academic and Research Libraries and Knowledge Management**
Simultaneous Interpretation:	Yes

WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL
20-24 August 2006, Seoul, Korea
http://www.ifla.org/IV/ifla72/index.htm

ABSTRACT

In Australia, the National Library is collaborating with the higher education sector as part of a national program to improve the nation's research information infrastructure. These activities have provided a focus for the National Library in its engagement with the university community.

Among other things, the National Library has:

- *worked to digitise research resources, especially those required to support research in the humanities;*
- *developed integrated discovery services such as the national union catalogue and a service providing search access to all Australian university repositories;*
- *actively participated in three research information infrastructure projects; and*
- *worked with partners to develop solutions to the problem of sustaining university repositories to support long term access.*

One of the research information infrastructure projects is Project ARROW (Australian Research Repositories Online to the World) which is developing a solution for institutional repositories in collaboration with a commercial vendor. The

National Library's principal contribution has been to build a national search service covering all Australian institutional repositories, thus improving the discovery infrastructure.

In another project, the Australian Partnership for Sustainable Repositories (APSR), the National Library is sharing its expertise in digital preservation. Among other things, the Library is helping the project to develop a sound approach to assessing the obsolescence risk of file formats, is advising on a strategy for including preservation metadata in the repositories, and is seeking to influence the future development of open source repository software to make use of preservation metadata.

INTRODUCTION

This paper is based on the experience of the National Library of Australia during the past three years in working with universities and the federal Department of Education, Science and Training to improve research information infrastructure in Australia.

In Australia in 2001 the federal Department of Education, Science and Training launched the Systemic Infrastructure Initiative [1] as part of the Government's "Backing Australia's Ability" Program. The Initiative was aimed at improving the effectiveness of Australian research by developing "research infrastructure" services. Such services comprise categories such as:

- collaborative support services (services needed by distributed research teams such as peer-to-peer data sharing, simulation and visualisation tools, and collaborative annotation tools);
- middleware services (such as federated authentication and authorisation, and digital rights management) which will support more streamlined access to the resources that each researcher is entitled to use; and
- services which allow institutions to manage, make accessible and preserve the outputs of research, such as research papers and data sets.

This third category represents what is termed, in this paper, "research information infrastructure". A more formal definition is "the set of services that support the discovery and management of research resources and research outputs by the current and future research community". These services are often developed collaboratively and are intended to benefit the whole information access system. Examples of research information infrastructure include:

- content, such as electronic texts, which support research in the humanities;
- services used by researchers to discover resources to support their research;
- institutional repositories which manage research outputs; and
- digital curation and similar services which ensure long term access to research outputs.

There are many providers of such services. In Australia, there is not yet a fully organised process for achieving inter-operability between such services, or for identifying and dealing with overlaps and gaps. However, there is a formal process to

provide advice to the Department of Education, Science and Training on the future development of research information infrastructure. The Department has established an advisory committee (the Australian Research Information Infrastructure Committee) which includes representation from university libraries, academics, the Australian Research Council, and the National Library [2].

Many services, publicly funded to support research, have the potential to deliver benefits to the general public. The Internet itself is the most prominent and obvious example of "research infrastructure" which has evolved into public information infrastructure. The public, no less than the researcher, requires rapid and easy access to information resources that they are entitled to use. In other words, there is an overlap between "research infrastructure" and "national information infrastructure". The National Library of Australia is interested in identifying and strengthening these synergies.

The National Library has a longstanding record of support for information infrastructure. Its enabling legislation gives the Library a mandate to build a national collection, to make it available in the national interest, to provide a record of Australian publishing and other bibliographic services, and to "cooperate in library matters". In acting on this mandate over the years, the Library has developed some key components of the national information infrastructure. These activities support underlying objectives such as:

- to provide online access to a greater range of content;
- to provide easy-to-use services to discover and access information content, including the collections of Australian libraries;
- to ensure that future users will be able to access online content; and
- to give people efficient access to the online resources that they are entitled to use.

Examples of the National Library's recent activities in pursuit of these objectives are set out below.

DEVELOPING DIGITAL CONTENT

The development of digital content by libraries and other services has a clear ability to support research of all kinds. It has a particular benefit in the humanities, as was explained in a recent submission to the Department of Education, Science and Training [3]:

Humanities research data … has to be collected, primarily from non-digital sources, and its collection has traditionally been a matter of individual labour, closely integrated with the processes of annotation and commentary that transform data into information. The labour-intensive processes by which humanities data is collected make it expensive to obtain. It is, however, highly reusable, and in digital form it can be preserved and benefit from a vastly increased capacity for circulation.

Humanities researchers obtain much of their research data from the collections of libraries, archives and museums. A growing proportion of these collections, including

a range of historical and cultural records, are being digitised, but the pace of these activities has been constrained by resource limitations. It is possible to identify many additional projects which could dramatically expand the resource available to the humanities and social science researcher.

An example identified in Australia is the digitisation of a collection of significant newspapers covering a period of 150 years from the early 19th century. Such a project, which would cost around A$3 million, would enable researchers to search across the full text of all digitised newspapers. Similar projects have been funded in the United Kingdom by the Joint Information Systems Committee (JISC) and in the United States by the National Endowment for the Humanities. A national digital newspaper database could:

- support biographical and historical research, enabling researchers to locate relevant newspaper articles for more efficiently than at present, and thus enabling them to focus more on unpublished information sources;
- provide an invaluable resource for longitudinal cultural studies, including media studies; and
- could support longitudinal research in certain scientific fields such as ecology and climate change.

In 2005 the National Library partnered with the Australian National University (ANU) and other universities to prepare a funding bid for such a project. The funding bid was not successful, but the Library remains convinced that such a service would be a valuable component of research information infrastructure. It is therefore continuing to explore mechanisms through which the project might be advanced.

An online searchable newspaper service provides a good example of infrastructure which could benefit both the research community and the general public, including historians, family historians and a wide range of other users.

Funding the development of digital research resources from a public source will overcome the access restrictions which occur when public domain works are digitised by the commercial sector. In the latter case, even though the works are in the public domain in their printed form, the licenses restrict use of the digitised versions to authorized, fee-paying users.

DISCOVERY SERVICES

Researchers and the general public need easy-to-use services to discover and access information content, including the collections of Australian libraries. In its *Directions Statement* for 2006 to 2008, the National Library defined one of its major undertakings as "to enhance learning and knowledge creation by further simplifying and integrating services that allow our users to find and get material" [4].

There is a significant challenge in improving the power, ease of use, and level of integration of the available discovery and access services, mainly because the discovery landscape is currently quite complex. Mechanisms for discovering information resources include:

- Google, including Google Scholar;
- Institutional and collaborative portals;
- Library catalogues and union catalogues;
- Specialised discovery services, including those which provide a central point of access to multiple institutional repositories;
- Indexes, databases and electronic journal aggregation services, subscribed to by university libraries, that support access to journal articles, conference papers, and similar resources; and
- Subject gateways.

There is a need for collaborative action to simplify this complex discovery landscape and to ensure that the various services are inter-operable. There is also an argument for subsidising the costs of these services as part of the national research information infrastructure.

One approach to simplifying the discovery landscape is to enhance the role of union catalogues. By aggregating metadata, union catalogues are well placed to aggregate both supply and demand, thus increasing the chance that a relatively little-used resource will be discovered by somebody for whom it is relevant. In addition, union catalogues are well placed to seed metadata on behalf of libraries to public search engines such as Google, increasing the exposure of library collections to researchers and the public.

In recent years, a number of union catalogue services have moved to a more open business model. For example, OCLC's "Open WorldCat" program is making the data in the WorldCat database freely available to web users via popular Internet search, bibliographic and bookselling sites. With a similar motivation, the National Library of Australia introduced free access to the Australian union catalogue (known as *Libraries Australia*) early in 2006. In addition, the Library has seeded metadata from the union catalogue to Google, along similar lines to Open WorldCat. This will help bring the content of Australian library collections to the attention of users who might have overlooked some library catalogues as discovery pathways.

In addition to *Libraries Australia*, the National Library has developed number of other freely accessible federated discovery services, namely PictureAustralia, MusicAustralia, the Register of Australian Archives & Manuscripts, and the ARROW Discovery Service.

The ARROW Discovery Service [5] provides federated search access to the content of Australia's university repositories. The service regularly harvests metadata from these repositories, using the OAI Protocol, and aggregates it into a database hosted by the National Library. A search will result in an integrated display presenting relevant items in various repositories, and a click on a selected result will take the user to the corresponding item in the local repository. To date, the service has harvested over 20,000 metadata records from 13 institutional repositories, including repositories based on ePrints, DSpace, Fedora (Fez and VITAL applications) and ProQuest's Digital Commons. The aggregated metadata will be made available for other services

(including international disciplinary based services) to harvest using the OAI Protocol.

ARCHIVING AND CURATION

Our definition of research information infrastructure referred to "the current and future research community". It is important to ensure that future researchers are able to access the online resources which are being created and archived today.

The National Library is undertaking several activities to achieve this. For the past ten years, in collaboration with the state libraries and other partners, it has developed PANDORA, an archive of selected Australian web sites and online publications. During 2005, the Library undertook the complementary activity of commissioning and analysing a comprehensive capture of the Australian web domain. These web archives will provide a rich resource for future researchers to mine and analyse.

Ensuring future access to such content will require a range of concerted measures beyond merely archiving the content. These measures will depend in part on gaining a better understanding of the file formats which comprise the content, the obsolescence risks associated with each format, and the costs of dealing with obsolescence through format migration and other techniques. Among the content requiring sustainable access are the primary and secondary outputs of research.

The primary outputs include data sets, images, video files and sound recordings generated as the "raw outputs" of research. The secondary outputs include books, pre-prints, journal articles, conference papers, theses, technical reports, unpublished papers and web sites which interpret and summarise the research findings.

The secondary outputs have of course traditionally been preserved, in printed form, in our library collections. However the majority of these outputs are now available in electronic form through electronic journals, personal web sites of academics, and institutional and departmental web sites. This change has raised issues about how these documents should best be managed for long term access.

The primary outputs are typically managed by academics on faculty servers, or (in the case of some research disciplines) by national and international data centres. The long term preservation of this data is threatened by inadequate data management practices, by the steady emergence of new file formats and by the technological obsolescence which accompanies this process [6].

Some of these primary outputs have long term value. These may include:

- quality research in the humanities;
- social science research, including statistical data, where future time series analysis is likely to be beneficial;
- epidemiology in medical research;
- ecological studies of particular regions; and
- most geoscience and meteorological data.

In 2003, the United Kingdom's "e-science curation report" [7] noted that:

- researchers have low awareness of data longevity issues;
- there are no procedures in place to encourage researchers to work in partnership with curators;
- funding of data curation tends to be short term funding, which is antithetical to the long term nature and needs of data curation;
- where retention of data is a requirement set by funding bodies, most researchers said that this requirement was not funded; and
- there was no government level strategy for data stewardship to which researchers and administrators can refer.

The position in other countries is likely to be similar.

Institutional repositories

The development of institutional repositories has, at least in part, been a response to the need for improved infrastructure to support academics in managing their research outputs.

In Australia, a number of universities have been using the Southampton E-prints software for the last couple of years. Some others, most notably the Australian National University, are implementing DSpace as their repository solution.

The ARROW Project (Australian Research Repositories Online to the World), led by Monash University, was funded in 2003 by the Australian Department of Education, Science and Training. ARROW has supported the development and deployment of the VITAL software from VTLS Inc., which is based on the open source Fedora software [8].

As a partner in ARROW, the National Library of Australia has been trialling the VITAL software. Though the Library already has a digital services architecture which supports both the PANDORA Archive and the digital content which it creates through its digitisation workflows, the Library does not currently have a satisfactory solution for a third ingestion method, in which digital content is submitted or deposited by publishers. Trialling the VITAL software will give the Library an opportunity to evaluate a potential solution for this ingestion method, and an opportunity to use this information in a reassessment of the medium term future of its digital collection architecture.

Another Australian institutional repository project is the Australian Partnership for Sustainable Repositories (APSR). This project, led by the Australian National University (ANU), aims to develop demonstrator repositories and support continuity and sustainability of digital collections, including research data sets [9]. The demonstrator repositories are being developed at ANU, University of Sydney and University of Queensland.

As a partner in APSR, the National Library is sharing the expertise on digital preservation issues that it gained through activities such as PANDORA. The Library is also helping the project to develop a sound approach to assessing the obsolescence risk of each file format represented in the APSR repositories, and is trialling a software tool that will alert repository managers of impending obsolescence. The Library is also advising on a strategy for including preservation metadata in the repositories, and will seek to influence the future development of open source repository software (such as DSpace and Fedora) to make use of preservation metadata. The aim of APSR is to make these tools and processes widely available to all those who are aiming to build sustainable collections of digital content.

National infrastructure to support curation

It is unlikely that institutional responses to the challenges of preserving research outputs, on their own, will be sufficient. National infrastructure support will also be needed.

In the United Kingdom, the e-Science curation report [7] called for long term funding for data curation, including funding of the Digital Curation Centre on a permanent basis. The Centre will deliver a number of outputs [10] including:

- an advisory service and help desk;
- repository guidelines, and a curation manual;
- a programme of research into curation issues; and
- testing and certification processes.

In the United States, the National Science Board in 2005 released a report entitled *Long-lived digital data collections: enabling research and education in the 21st century*. The report [11] called on the National Science Foundation to develop a clear technical and financial strategy relating the investment in building and maintaining data collections of long term value to the use made of those collections in supporting research and their significance to the wider community. It recommended that research proposals for activities that generate digital data should be required to include a data management plan. It also recommended that more be done to foster the skills of the data scientists who will be needed to support the management of these collections.

In Australia, the Prime Minister's Science, Engineering and Innovation Council (the Government's principal source of independent advice on issues in science, engineering and innovation) [12] has recently established a Working Group on Data for Science, which will report to the Government in December 2006. The Working Group will review the current approaches to the management of large amounts of scientific information and recommend a data management strategy "to ensure Australia's scientific sector provides benefits to the Australian economy, environment, and society". The National Library is represented on the Working Group.

ACCESS CONTROL SERVICES

Access control services ensure that users gain efficient access to external information resources, where they are entitled to such access. They support requirements such as federated authentication and authorisation, and digital rights management. These services form another important component of the national information infrastructure. The major Australian project aimed at developing improved access control services is the MAMS (Meta Access Management System) Project [13].

Under the federated authentication model that the MAMS Project is developing, a user seeking to gain access to any service that is part of the national information infrastructure would be referred back to a home institution for authentication there, and then redirected back to the service provider with a standard "security handle". The service provider would then use this security handle to query the home institution about the user's attributes. Based on these attributes, the service provider would then give the appropriate level of access rights to the user. All of these processes would occur through machine-to-machine transactions and would be managed by an "Access Control Federation" which will be established as a by-product of the MAMS Project.

The benefit of this approach is that the service provider will not need to set up and maintain its own user directory and authentication system. The service provider needs to know "is this person from an organisation that we trust?" and "does this person belong to a class that is entitled to use this service?" but does not need to maintain individual passwords or certificates, and does not need to know any more details about the user.

The National Library was invited to participate in the MAMS Project because of its work in the directory standards field. The Library will be modelling use cases that are relevant for library services, and will specify the mechanisms through which user attributes can be related to the service provider's policies. The Library will also pilot use of the MAMS tools in its own services, including *Libraries Australia*, in order to support federated authentication.

As with the other elements of information infrastructure, there is no reason why this approach could not be used beyond the research sector. For example, it could be a future mechanism supporting access by public library users to a national licensed set of online content.

CONCLUSIONS

We have defined research information infrastructure as "the set of services that support the discovery and management of research resources and research outputs by the current and future research community". In Australia, the federal Department of Education, Science and Training is funding improvements to the research information infrastructure.

The National Library of Australia has been a participant in these developments. In particular, the Library has:

- participated in the committee which has advised the Department on the future development of research information infrastructure;
- worked to digitise research resources, especially those required to support research in the humanities;
- developed integrated discovery services such as the national union catalogue and a service providing search access to all Australian university repositories;
- actively participated in three research information infrastructure projects;
- worked with partners to develop solutions to the problem of sustaining university repositories to support long term access;
- participated in a major national working group to develop a strategy for the management of Australia's scientific information; and
- undertaken work in the standards arena aimed at developing improved access control services.

These activities have provided a focus for the National Library in its engagement with the university community. In the process, the Library has gained an improved understanding of the challenges faced by our university partners, while also being able to share our own experience and skills in fields such as digital preservation. The Library is committed to continue collaborating with the higher education sector to improve the national research information infrastructure.

REFERENCES

1. Department of Education Science and Training 2001. *Additional funding for systemic research and research training infrastructure in universities*. http://backingaus.innovation.gov.au/2001/research/systemic2001.htm

2. Australian Research Information Infrastructure Committee (ARIIC). http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/australian_research_information_infrastructure_committee/default.htm

3. Maltby, Richard; Harris, Margaret and Eggert, Paul. NCRIS Strategic Roadmap exposure draft: a response. December 2005.

4. National Library of Australia. *Policy and planning*. http://www.nla.gov.au/policy/

5. ARROW Discovery. http://search.arrow.edu.au

6. Cathro, Warwick. Preserving the outputs of research. http://www.nla.gov.au/nla/staffpaper/2004/cathro1.html

7. Lord, Philip and MacDonald, Alison. E-Science curation report: prepared for the JISC Committee for the Support of Research by Philip Lord and Alison Macdonald. http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

8. ARROW: Australian Research Repositories Online to the World. http://www.arrow.edu.au/

9. Australian Partnership for Sustainable Repositories 2005. http://www.apsr.edu.au/

10. Burnhill, Peter. So who's that new kid on the block? DPC Forum: Digital preservation – the global context, 23 June 2004. http://www.dpconline.org/graphics/events/for040623.html

11. National Science Board (US). Long-lived digital data collections. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_1.pdf

12. The Prime Minister's Science, Engineering and Innovation Council (PMSEIC). http://www.dest.gov.au/sectors/science_innovation/science_agencies_committees/prime_ministers_science_engineering_innovation_council/

13. Meta-Access Management System (MAMS). http://www.melcoe.mq.edu.au/projects/MAMS/