Easy To Do
A brief history of federated harvesting in Australia

Debbie Campbell
Director, Coordination Support Branch, National Library of Australia

Keywords: OAI, Open Archives Initiative, Protocol for Metadata Harvesting, OAICat, ADTP (Australasian Digital Theses Program)

ABSTRACT

This paper charts a brief history of the development of metadata aggregations in Australia and discusses how the dispersal of metadata over the Web has transformed library services from a single institutional outlook to delivery on a national scale. Rich applications of metadata have been able to demonstrate a return on investment by being reused in different contexts, exceeding the expectations of their creators.

Services including subject gateways and portals, which feature federated metadata collections created in a collaborative framework, have relied on the availability of open source software to establish themselves. Use of open source software in collaborative projects allows libraries to share the pain of the development and delivery of services. In turn, the open source movement has strengthened because of this approach. Whether this is sustainable in the longer term is still an open question.

While the Dublin Core Metadata Initiative is 10 years old, Australia changed its position from that of participant in the Initiative to that of leader quite soon after. A combination of the far-sighted thinking of then university librarian at the University of New South Wales, Marian Bate, the MetaWeb project led by the Distributed Systems Technology Centre, and the impending creation of digital theses in the Australian Higher Education sector culminated in the Australian Digital Theses Project. The Project, begun in 1998, is one success story - it became a Program in 2000.

But the ADT Program is not a one-off success. New services based on the federated harvesting of metadata continue to be developed. The paper explores how the application of metadata in a collaborative environment has provided new resources for education and research.

1. Introduction

The launch of the Dublin Core Metadata Initiative 10 years ago was integral to the development of collaborative, federated harvesting services (Dublin Core Metadata Initiative, 1995). An early implementer was the Australian Digital Theses Project, which sought to gather metadata records describing digital theses from distributed universities, and store them in a centralised metadata repository.

The new ADT service was based on the work of the MetaWeb Project (1997). Marian Bate, then university librarian at the University of New South Wales, saw the potential of the harvesting concept as a way of resolving the resource-intensive cataloguing processes which surrounded 'electronic' resources. The Santa Fe convention emerged just a few months later. (Santa Fe, 2000).

2. What do we mean by federated harvesting ?

Federated harvesting is an automated technique used to gather metadata records from distributed record repositories, usually on a regular basis. The technique, developed by the Open Archives Initiative and renamed the Protocol for Metadata Harvesting, manages harvesting of new and updated records (OAI, 1999).

An older technique, Harvest Control Lists, requires a total refresh of the harvested records each time, and does not scale well in both the creation process or the harvesting process. Repository files containing more than a few thousand records aggregated into a database of more than one million records are much more efficiently harvested using the date-driven algorithm provided by the OAI protocol.

The result of using either technique is an up-to-date aggregation of metadata containing links to digital content. Both are possible using open source software. The OAICat software, maintained by the Online Computer Library Center (OCLC), allows a repository to be harvested as well as to operate as a harvester itself.

Federated harvesting has evolved to keep up with user expectations of Web services. The technology is relatively straightforward to implement. In fact, it may be too easy to do, because many metadata-based aggregations are appearing as new Web portals. Portal managers have a responsibility to consider other services which may provide a similar aggregation, to minimise the number of competing portals and therefore user confusion. The National Library is reviewing its own portals in this context, by examining the role of a trusted aggregator.

3. What do we mean by trusted aggregations ?

3.1 Libraries Australia

A trusted aggregation is a file of records produced by a known service provider. For example, Libraries Australia, which sources its records from libraries and other collection managers around the country, has provided the National Bibliographic Database since 1981 (Libraries Australia, 2004). The National Library defines a trusted aggregator as a service which brokers metadata exchange. The exchange exists in two forms: harvesting of metadata to use in a "specialised service", or "contribution of metadata to a trusted aggregator without having to have direct relationships with every information provider." As long as the trusted aggregators are easily identified, then service providers should be able to rely on their brokerage services (NLA, 2005).

3.2 PictureAustralia, MusicAustralia

The National Library of Australia is a trusted aggregator. In addition to Libraries Australia, the success of several national services is based on federated harvesting. These include PictureAustralia (launched in 2000), MusicAustralia (launched in 2005), and the ARROW Discovery Service (launched in 2005) prove this.

PictureAustralia has more than 40 cultural institutional participants, and is not restricted to libraries. Museums, galleries, and universities have also contributed their records to the collaboration. PictureAustralia has switched to using the OAICat software for larger agencies, whose collections generally number several hundred thousand images.

Similarly, MusicAustralia has been able to encourage participation from music-centric agencies such as the Australian Music Centre. This aggregator service takes a slightly different approach for sourcing its records. Participants describe their resources using the Metadata Object Description Schema (MODS, 2002). Upon harvesting by the National Library, MODS is converted to MARC21 and loaded into the National Bibliographic Database. These records are then re-selected for presentation in the MusicAustralia service. Using this process, more than 144,000 music information resources are made available in a purpose-built portal. The choice of MODS has enabled collections not described in MARC21 to be included in a trusted aggregation.

3.3 The ARROW Discovery Service

The Australian Research repositories Online to the World (ARROW) Project has a specific remit to provide federated access to scholarly research outputs from all of Australia's universities. The ARROW Discovery Service harvests records from individual universities' institutional repositories using the OAI Protocol for Metadata Harvesting (ARROW, 2005). It in turn, is harvestable using the same protocol.

The ARROW Discovery Service is also harvesting metadata records from the Australasian Digital Theses Program (for Australian outputs only). This is important because ADT Program has a selection policy which confines itself to graduate research theses. ARROW on the other hand accepts all types of theses, including thesis by coursework. Duplication of records is expected to be modest until all universities are participating in both services. But the ADT records form an important component of the Service because many universities are still to establish their institutional repositories.

Duplication may not become the issue expected. If universities clearly delineate the discovery of theses by earmarking research theses for ADT only, and other theses for ARROW only, which is easily achievable using the OAI protocol, then each Service can achieve its full potential.

4. The case of the Australian thesis

4.1 Records available

In June 2005, a quick search showed that there were records for approximately 160,000 theses of Australian origin in Libraries Australia. These represent both print and digital theses. All of Australia's universities have supplied records to the National Bibliographic Database for theses, and a lot of other agencies as well, including the Commonwealth Scientific and Industrial Research Organisation (Rajaptirana, 2002). It is worth noting that not all digital theses have been catalogued, but this gap is closing as the digital services increase awareness of their existence.

The Australasian Digital Theses Program does not provide coverage for all of the nation's universities (18 out of the 40 have abstained from participation to date), so in order to address that, the ADTP search service is being augmented by all thesis records from Libraries Australia. This exemplifies the trusted aggregator exchange. The exchange will increase the number of records in this purpose-built service from 3,730 (as at June 2005) to more than 160,000.

The thesis is changing its nature as new digital file formats are exploited. Two theses well-known in Australia for their format as Web sites are **Flight of Ducks** and **Milkbar**, both of which are captured in Australia's digital archive, PANDORA. As part of this service, both theses have been catalogued into the National Bibliographic Database and are therefore discoverable via Libraries Australia.

Ignoring the gaps in the intellectual record caused by disinterest in deposit or lack of metadata creation, the richness of availability of theses through multiple freely accessible portals means that branding and brand awareness are issues for all of the services.

4.2 How are Australian theses discoverable ?

How does a searcher identify a trusted aggregation to find the authoritative source of an Australian thesis ? It is reasonable to assume that a searcher will use their portal of choice to find a thesis. Trust won't necessarily be judged on who is hosting the search service. Rather, it will be the search engine most reliably providing an answer to the question at hand. For a lot of searchers, this will be Google. Academics may prefer Google Scholar.

For academics who deposit their theses into their institutional repository retrospectively, and for those universities using their institutional repositories to store theses, the ARROW Discovery Service will provide a persistent, reliable discovery mechanism for theses amongst other digital scholarly research outputs. The Service is also brokering discovery via other national portals such as the Australian Academic Research Library Information Network (AARLIN) and EdNA Online.

However, this doesn't mean an academic won't expect to find an example of their own work in a regular search engine. International services including OAIster and Google have been approached to harvest ARROW Discovery Service records (including those for theses) using the OAI Protocol for Metadata Harvesting. As part of the brokering arrangement, OAIster supplies records to Yahoo. These exchanges are always brokered to retain Australian branding.

For those users who only want to focus on thesis output, the Australasian Digital Theses Program and its international equivalent, the Networked Digital Library of Theses and Dissertations (NDLTD) will be the portals of choice. The NDLTD is expecting to facilitate access to the full text of theses. Some Australian universities have already chosen to block access to full text, but their metadata records will still be available. The choice of NDLTD reinforces the trusted aggregator exchange.

National-scale services are best placed to resolve differences in encoding practices and identify digital theses orphaned from thesis services (including Milkbar and Flight of Ducks). The ability to update record details, particularly system-generated encodings via a global change process, provides for some robustness. Similarly, search protocols can be deployed to bring together records for theses from disparate services.

National-scale services have the capacity to try different approaches to discovery. The outline above shows that they can also be complementary, thereby making Australian theses as discoverable as possible.

5. Conclusion

Progress since the original germination of an idea by Marian Bate in 1997 has been considerable. Using simple but efficient protocols has allowed ultimate discoverability of the research outputs of a nation, and exemplifies that original vision.

References

Dublin Core Metadata Initiative, 1995. Retrieved June 17, 2005, from http://dublincore.org

The MetaWeb Project (1997-1998), Retrieved June 17, 2005, from www.dstc.edu.au/Research/Projects/metaweb/

The Santa Fe Convention for the Open Archives Initiative, 2000. Retrieved 17 June 2005 from www.openarchives.org/meetings/SantaFe1999/sfc_entry.htm

Open Archives Initiative, 1999. Retrieved June 17, 2005, from www.openarchives.org

OAICat, OCLC. Retrieved June 17, 2005, from www.oclc.org/research/software/oai/cat.htm

Libraries Australia, 2004. Retrieved June 17, 2005, from http://librariesaustralia.nla.gov.au

PictureAustralia, 2000. Retrieved June 17, 2005, from www.pictureaustralia.org

MusicAustralia, 2005. Retrieved June 17, 2005, from www.musicaustralia.org

Metadata Object Description Schema, 2002. Retrieved June 17, 2005, from www.loc.gov/standards/mods/

ARROW Discovery Service, 2005. Retrieved June 17, 2005, from http://search.arrow.edu.au. The ARROW Project, 2004. Retrieved June 17, 2005 from www.arrow.edu.au

Rajapatirana, B. 2002 Survey of Australian Theses Contributions to the National Bibliographic Database. Retrieved 17 June, 2005, from www.nla.gov.au/kinetica/austthesessurvey.html

Flight of Ducks. Retrieved June 17, 2005, from http://pandora.nla.gov.au/tep/10245

Milkbar. Retrieved June 17, 2005, from http://pandora.nla.gov.au/tep/13196

PANDORA, Preserving and Accessing Networked Documentary Resources of Australia, 1997. Retrieved June 17, 2005, from http://pandora.nla.gov.au

Australian Academic Research Library Information Network (AARLIN). Retrieved June 17, 2005, from www.aarlin.edu.au

EdNA Online. Retrieved June 17, 2005, from www.edna.edu.au

OAIster. Retrieved June 17, 2005, from http://oaister.umdl.umich.edu/o/oaister/

National Library of Australia Digital Object Repository. Retrieved June 17, 2005, from www.nla.gov.au/digicoll/oai/

Yahoo. Retrieved June 17, 2005 from http://search.yahoo.com.au

Google. Retrieved June 17, 2005, from www.google.com.au

Google Scholar. Retrieved June 17, 2005, http://scholar.google.com/

Networked Digital Library of Theses and Dissertations (NDLTD). Retrieved June 17,2005, from www.ndltd.org/ Full text access discussed at http://outgoing.typepad.com/outgoing/2005/06/scirus_and_else.html