ETD2005 Conference
"Two into one will go": combining two institutional repositories at University of Melbourne.

Nicki McLaurin-Smith
Director, Information Management Program, Information Division, University of Melbourne

Eve Young
Metadata Co-ordinator, Information Management Program, Information Division, University of Melbourne

Shirley Sullivan
Electronic Information Co-ordinator, Information Management Program, Information Division, University of Melbourne

Keywords: Theses, ADT, Open Archives Initiative, metadata, eprints.

ABSTRACT

University of Melbourne has been a participant in the Australian Digital Theses (ADT) Program since its inception in 1998 and has had an eprint repository for research output since 2002.

Technical problems meant that the University of Melbourne server was unavailable for deposit or viewing of theses over an extended period. This caused a lot of frustration for doctoral students wishing to submit theses. In 2004 a software solution emerged from the University of Tasmania whereby the theses could be loaded in UMER (The University of Melbourne Eprints Repository) and harvested by the ADT.

The paper will cover the redevelopment of the University of Melbourne's ADT Program with the help of staff from UNSW Library. Changes in workflow consequent upon the altered deposit requirements will be addressed, including scanning, cataloguing and Kinetica work. Reference will be made to legal issues and consultation with the School of Graduate Studies.

Benefits of the solution will be outlined. These include the advantages of OAI (Open Archives Initiative) compliance, such as increased exposure to theses through search engines like Google, and the improved statistical reporting provided by UMER's use of eprints.org software.

Lessons learned include the need to focus on easy technical solutions for users, development of simple digital rights management guidelines and the need to work with the academic community to build their awareness and understanding of the changes in scholarly communications.

The Information Division of the University of Melbourne will consider whether a single repository will be sufficient to meet the diverse requirements regarding formats and access conditions of the different communities, or whether an approach where multiple repositories will co-exist will better meet the University's needs for management of a wide range of digital formats.

1. INTRODUCTION

This paper will outline the move of University of Melbourne digital theses from the Australian Digital Theses (ADT) server to the University of Melbourne Eprint Repository (UMER). The history of the University's involvement in ADT will be briefly outlined, and the reasons and methodology behind the move discussed. The paper will conclude with exploration of some broad intellectual property and copyright issues and possible future scenarios at the University of Melbourne.

2. AUSTRALIAN DIGITAL THESES PROGRAM (ADT)

The ADT program was established in 2000, following a successful project funded by the Australian Research Council. The University of Melbourne was one of seven original participants. The program was based on the work of the international Networked Digital Library of Theses and Dissertations[1] using software developed by Virginia Tech specifically for digital theses but amended slightly to suit ADT requirements. The main aim of the program was to create a distributed national database of digital versions of theses[2] and make them freely accessible via the web (Lafferty 2005). For more detailed information on the ADT program see the website at http://www.library.unsw.edu.au/thesis/adt-ADT/info/aims.html and the articles by Cargnelutti (1999, 2004), Wells (2003) and Lafferty (2005).

In 2004, the ADT obtained Strategic Information Infrastructure funding through the Australian Department of Education, Science and Training (DEST) to expand and redevelop the ADT. The metadata repository's content will expand to become a comprehensive record for all Australian higher degree theses, whether in digital form or not. Bulk retrospective digitisation of all Australian theses is also planned, so the ADT will become the source for all Australian theses capable of being digitised.

The proposed extension to the ADT includes building a simple Open Archives Initiative (OAI) metadata harvesting protocol interface to the central ADT metadata repository. The central ADT repository harvests the metadata from the local web pages and provides a national search service on the metadata linking back to the local repositories.. The metadata schema is Dublin Core (DC), used by many universities and research institutes (Sale 2004). This use ensures interoperability with other emerging metadata standards such as IMS and LOM (Learning Objects Metadata), which may be used on campus in other contexts such as the Learning Management System (LMS) (Young 2005).

The accepted document format in ADT is Adobe Acrobat Portable Document Format (PDF) to ensure that the data is independent of the platform on which it is created and that a high quality printed version can be provided if needed. It is also the preferred format for preservation and ensures the integrity of the content. Since PDF is a pervasive format used heavily by governments and commercial publishers as well as libraries, libraries will not be alone is trying to solve future migration issues (McMillan 2004 p. 327)

But this format restriction also lead to limitations. It is now possible to create a thesis that contains multimedia elements and links to large data sets, for example, that in print format are impossible. (Andrewa 2004). Software packages for institutional repositories allow input of formats other than PDF, e.g., audio and video files.

3. UNIVERSITY OF MELBOURNE EPRINT REPOSITORY (UMER)

UMER is the University of Melbourne's Eprint Repository. It stores research output from the University staff and postgraduates. UMER includes a range of literary formats, including preprints, conference papers, technical reports and theses. The ADT accepts only theses submitted and passed for research higher degrees, but UMER accepts all theses, including those of masters by coursework and honours.

4. DISCUSSION

The ADT program is designed to improve knowledge of, and access to, Australian theses, both nationally and internationally (Lafferty 2005). Theses are a potentially rich source of primary research across many disciplines, but are underutilised through lack of exposure.

Research in the US has shown that the use of theses increases spectacularly with electronic access, ensuring a far greater national and international visibility. (Pioneering 2005) At Virginia Tech, e-theses were 100 times more likely to be circulated than the paper equivalent. The URL http://scholar.lib.vt.edu/theses/data/somefacts.html#logs shows the hits on Virginia Tech's theses server, including location of the requester as well as downloads, and compares these with the loan statistics from their print theses. Analysis reveals that use of these theses is not limited to educational establishments; over the past few years government, commercial and other organisations have consulted the works (Copeland 2004).

The existing ADT repository gets high use from national and international searchers. See the usage statistics at http://adt.caul.edu.au/usage/usage_200506.html. The improved interface of the expanded ADT will encourage greater visibility and use of Australian theses (Wells 2003). Demonstrated high use of an institutional repository encourages academics to submit their work, especially when they see that the work can be found through Google as readily as via the more scholarly search engines such as OAIster and Google Scholar (which limits searching specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research).

Around the turn of the millennium, the University of Melbourne merged the library and IT services, with an ensuing restructure and changes in areas of responsibility. Somehow during this process the ownership of the ADT program "fell through the cracks". No-one updated the hardware, and no-one kept up to date with changes in potential software solutions, such as developments associated with the OAI.

By June 2003 major technical problems began to appear. The server broke down and was out of warranty. Many of the problems were caused because the software supporting the University of Melbourne ADT server was still the original 1998 version[1]. The free software enabling students to digitize their theses was no longer available. Hackers were getting into the site, creating duplicate theses records, and maintenance files were unusable, so students could not load their files. Despite messages placed on the site alerting potential users to the fact that the site was out of order, postgraduates were still trying to load their theses and people were trying unsuccessfully to access theses.

5 METHODOLOGY

[1] Many of the other partners had moved on to updated versions

An ADT stakeholders group was established in June 2003. Members of the group were drawn from five departments of the Information Division – Information Resources Access, Client Services, Systems & IT Infrastructure, Business Management Services and Teaching, Learning & Research Support. Over successive meetings the group learned that University's dedicated ADT server was broken, postgraduate students were sending letters of complaint about not being able to load their theses onto ADT, and insufficient technical support was allocated to redress problems.

As an interim solution, the Information Division (ID) requested assistance from UNSW staff. Tony Cargnelutti[2] and Fred Piper[3] responded graciously. UNSW placed the University of Melbourne's ADT theses on its server and cleaned up the files and duplicates.

ID staff had been promoting UMER to staff and postgraduate students, and the group agreed to place new theses into UMER as an interim measure. Before long there were 104 theses in UMER. A clear and intuitive interface to ensure that authors are encouraged to submit content is important. (Copeland 2004) Postgraduates reported that they found it much easier to deposit their theses into UMER and had less need to call on University staff for support, compared to loading into ADT. Eve Young reported to the ID Executive that other academic libraries had also found the ADT software rather complicated to use, and were using other software solutions, for example Curtin University is using the ETD-db from Virginia Tech.

Each institutional member of the ADT has its own server containing the full text of theses, linking to an ADT webpage containing metadata about each thesis. While loading theses into UMER had solved the technical problems a new complication arose, as the University of Melbourne now had two repositories containing theses. After consideration of Young's report, the ID Executive agreed to use UMER as the sole University of Melbourne thesis repository, provided that the theses metadata could be harvested from UMER by the ADT, thereby reaffirming the University's commitment to the ADT and participation in the DEST-funded expansion project.

In October 2004, the University of Tasmania released software which extracts the metadata for research theses and formats it for ADT to harvest. The software is similar to that used by University of Western Australia. The group agreed to use the University of Tasmania software to enable UNSW to harvest the theses metadata from UMER for ADT (http://eprints.comp.utas.edu.au:81/archive/00000078/)

Andrew Gfrerer, from the Web Publishing Team, had been collaborating with University of Tasmania staff on development of its Eprints statistics package. He obtained their permission to use their shareable software. Eve contacted UNSW, who confirmed that the University of Tasmania solution would work for us. It was, therefore, decided that UNSW would return the cleaned 3.5 GB data of theses.

UMER accepts all theses types from University of Melbourne students, including Honours and Masters by Coursework. ADT accepts only PhD and Masters by research. In order for the ADT to harvest the subset required, the list of theses types in UMER was expanded to include:

[2] Manager, Online Services Dept.

[3] Manager, IT Support Unit

- Honours thesis
- Masters Coursework thesis
- Masters Research thesis
- PhD thesis
- Other Degree Thesis

Theses loaded into the ADT are required to conform to the approved ADT Program filename standard <http://www.library.unsw.edu.au/thesis//adt-ADT/info/filename.html> and need to be numbered using the standard. Theses had to be broken up into 2–4 MB size, and put into a relevant file standard. Eprints uploads the whole PDF as one large file. It was decided that the ADT filetypes would be retained in UMER, but to avoid confusion for readers, the chapter number was added to each filetype. eg. PDF (chapter7-9)

Technical services staff have assisted ID staff by massaging the ADT records to fit UMER fields in regards to thesis types, subjects and faculties, inserting the keywords, and deleting duplicates. UMER uses the University faculty structure for subject headings. Each ADT thesis added to UMER had to be manually checked to see to which faculty the thesis was submitted and to include keywords, a useful field in UMER. Cataloguing staff also had to work through the ADT titles to add the new UMER URLs to the catalogue records, which were then uploaded again into Kinetica, the Australian national bibliographical database.

There was some extra information in the ADT records that were not required in UMER, so in order not to lose information, such as name of supervisor, that information was retained in the UMER comments and suggestions field.

Other information could be bulk uploaded. Fields such as institution and date of input were automatically inserted. As there was no knowing when the ADT thesis was first added, all the theses were given the date of when the bulk uploading was made. After that work was completed, ID staff informed the UNSW that our theses were ready to be harvested by ADT.

6. COPYRIGHT AND LEGAL ISSUES

The main function of the depositors' declaration is to ensure that the depositor is the copyright owner, or has sought and gained permission to include any subsidiary material owned by third party copyright holders. The repository administrators need to safeguard against third-party copyright material being inadvertently deposited by clearly indicating that reasonable care has been taken to prevent such occurrences and that any work will be removed if it is found to violate any copyright or other rights of any person. (Andrew 2004b)

The second function of the depositors' declaration is for the authors and any other rights holders, to grant permission to the host institution to distribute copies of their theses via the internet. Copyright and access issues in the ADT are the responsibility of each participating institution. The wording on permission forms was changed to include both repositories.

7. CONCLUSION

There are many advantages to using a single digital library for maintaining all forms of research output. A single piece of software is easier to maintain than two; longer–term archiving is easier to arrange, registration of the digital repository with service providers (such as harvesters and search engines) needs to be done only once, and training of staff is simplified. Time and effort expended by programmers to add value to the basic software, such as the addition to UMER of a statistics package, is easier to justify to management when one large database serves a wide range of clientele. (Sale 2005). Cataloguing staff need to familiarise themselves with only one database and its functionality. An additional benefit is that the loading interface for UMER is decidedly easier for users than that for the ADT, as mentioned previously.

Using a simple software package such as eprints.org enabled staff to learn how to create and manage an institutional repository. Academic staff were enthused by its demonstrated ease of use and utility in data management. There is now a need to develop a more sophisticated database to sustain a wide range of formats and resources to meet the needs of different departments and research centres within the University.

Lessons learned from the projects include the need to focus on easy technical solutions for users, development of simple digital rights management guidelines and the need to work with the academic community to build their awareness and understanding of the changes in scholarly communications.

Discussions between the ID and the academic staff have revealed that many academic staff have an interest in repositories. Despite the perceived advantages of one all-embracing institutional repository, the Information Division of the University of Melbourne may need to reconsider whether a single repository will be sufficient to meet the diverse requirements regarding formats and access conditions of the different communities, or whether an approach where multiple repositories will co-exist will better meet the University's needs for management of a wide range of digital formats.

References

Andrew, T. (2004b). *Intellectual Property and Electronic Theses*, Retrieved June 15th 2005 from http://www.jisclegal.ac.uk/publications/ethesesandrew.htm.

Andrew, T. & MacColl, J. (2004a).*Theses Alive Final Report, December 2004*. Retrieved June 15th 2005 from http://www.thesesalive.ac.uk/archive/ThesesAliveFinalReport.pdf.

Cargnelutti, T, Piper, F. & Kealy, K. (1999) The Australian Digital Theses (ADT) Pilot Project: the trials, tribulations and (some) successes. In *Educause '99*. Retrieved June 15th 2005 from http://www.library.unsw.edu.au/%7Eeirg/cause99.html.

Cargnelutti, T. (2004). The Australian Digital Theses Program: a national collaborative distributed model. In E. A. Fox, S. Feizabadi & J. M. Moxley and C. R. Weisser (Eds) *Electronic Theses and Dissertations: a Sourcebook for Educators, Students and Librarians*. (pp. 355-359). New York, Marcel Dekker.

Copeland, S. & Penman, A. (2004). The development and promotion of electronic theses and dissertations (ETDS) within the UK. *New Review of Information Networking*. *10*, 19-32.

Lafferty, S, Edwards, J & Dovey, K. (2005). The Australian Digital Theses Program: a disruptive technology? In *Proceedings, Educause Australasia*, Auckland, 5-8 April 2005.

McMillan, G. (2004). Implementing ETD services in the library. In E. A. Fox, S. Feizabadi & J. M. Moxley and C.n R. Weisser (Eds) *Electronic Theses and Dissertations: a Sourcebook for Educators, Students and Librarians.* (pp. 319-329. New York, Marcel Dekker.

Pioneering electronic access to UK theses (2005)**.** Press Release :Theses unbound, 7 April 2005. Retrieved June 15th 2005 from http://www.jisc.ac.uk/index.cfm?name=pr_theses_abound_news_060405.

Sale, A. (2004) Australian Computer Science on the global map. Retrieved June 15th 2005 from http://eprints.comp.utas.edu.au:81/archive/00000056/.

Sale, A. (2005). De-unifying a digital library. *First Monday*, *10,* May. Retrieved June 15th 2005 from http://www.firstmonday.org/issues/issue10_5/sale/index.html.

Wells, A. (2003). *Australian Digital Theses Program: Expansion and Redevelopment. A proposal for funding by the DEST Research Information Infrastructure Framework for Australian higher education.* Retrieved June 15th 2005 from http://adt.caul.edu.au/info/arricProposal2003.pdf.

Young, E. & Hughes, B.(2005) *If we're not there yet, how far do we have to go ? A review of web metadata at the University of Melbourne* . Retrieved June 15th 2005 from http://eprints.unimelb.edu.au/archive/00000923/.

[1] http://www.ndltd.org See also Chapter 5 of Fox, Edward A, ed. 2004 Electronic Theses and Dissertations for a brief account of this endeavour.

[2] OgCr nqqdpt hu` kdms+mc L ` r sdq ax qdr d` qpg nmkx-