ADT ProQuest Collaboration: a case study of a Library/Vendor alliance

Mary Anne Kennan, Tony Cargnelutti
ADT Program
University of New South Wales Library, Sydney

Helen Keyes, Paul Jensen
Griffith University Library, Brisbane

Austin McLean
ProQuest Information and Learning

Keywords: Digital theses, Information Management, Partnerships and Collaboration, Rights acquisition, Rights management, Open access, Australian research

ABSTRACT

The Australian Digital Theses (ADT) Program is collaborating with a commercial company, ProQuest Information and Learning (PQIL), to investigate practical solutions and service options to further the ADT's goals of open access to Australian research through world wide metadata exposure. Using a pilot project structure, the ADT is investigating the potential of a partnership with PQIL in the following areas:

- The scanning and digitisation of retrospective theses
- The performance of aspects of copyright checking (checking the thesis for attachments or other identifiable material that may be under another copyright)
- The provision of access to PQIL's Digital Commons Deposit software customised to ADT specifications for both current (born digital) and retrospective deposit
- The provision of a full text trial repository housing retrospectively scanned and born digital theses (plus metadata - in a PQIL proprietary format or output as Dublin Core)

This session will review the key issues for PQIL and the ADT in entering into the partnership, discuss the philosophical and logistical issues that had to be overcome, and provide detail of the progress of the project. One participating ADT institution, Griffith University, will reflect on challenges, benefits and outcomes of the project from their perspective, and impacts on its local ADT operations

1. INTRODUCTION

The Australian Digital Thesis (ADT) Program is a consortium of institutions managed on behalf of the Council of Australian University Librarians [CAUL] by The University of New South Wales Library. Since 1998 ADT has been seeking to increase open access to Australian-created theses. ProQuest Information and Learning (PQIL) under its University Microfilms International (UMI) imprint has been publishing theses since 1938. The two organisations were seen to have expertise in complimentary areas and aimed to leverage these in an exploratory pilot project funded by the Australian Government's Department of Education Science and Training [DEST] Systemic Infrastructure Initiative [SII] funding program through its Australian Research Information Infrastructure Committee [ARIIC]. The purpose of the pilot was to explore possible new models for the deposit, hosting, archiving and use of theses in an open access environment.

The pilot had two main initiatives:

- With regard to retrospective theses the project sought:
  - To complement and expand the existing ADT Program by digitising retrospective titles in an open access environment.
  - To explore the copyright issues with regard to the digitization of retrospective theses, to ensure both open access and protection of authors' rights.
- With regard to current theses the project sought:
  - To create an application in the form of a deposit form customised to the Australian context for the acquisition of rights and deposit of current born digital theses.
  - To showcase theses in existing databases and in an open access portal to facilitate world-wide access and dissemination.
  - To investigate and recommend standards for preservation of born digital theses and explore the mix of paper, microfilm and digital formats needed for 21st Century preservation.

Interested institutional members of CAUL met with PQIL representatives in January 2005. The proposed activities for the pilot were outlined and a project plan developed. Initially nine institutions expressed an interest; with seven institutions going on as active participants in the trial.

2. RETROSPECTIVE CONVERSION

2.1. Scanning and digitisation of retrospective theses

The umbrella ADT-ARIIC project provided the funds to cover retro-conversion costs at participating institutions. A number of options for retro-conversion were examined:

- Sending theses to PQIL in the USA to take advantage of their experience and efficiencies in digitisation processes
- Outsourcing to a commercial scanner in Australia
- Scanning within the library/institution

Initially participating sites were reluctant to send theses to PQIL in the USA. Most libraries do not have duplicate copies of theses and felt the risk of transportation was high. However, PQIL and their couriers have an exemplary safety record, the costs of outsourcing locally was high, and there was still a risk involved in transportation locally, so PQIL became an option for some institutions. A mix of methods was therefore tested. Theses were returned from ProQuest in an average of 8 weeks and were broken up and named according to ADT protocols. Unfortunately they were not OCR'ed, meaning that ADT sites had to run OCR software over them to enable book marking, accessibility and cutting and pasting of abstracts, titles etc. into the deposit software.

2.2. Copyright issues

Generally authors are the owners of the copyright to their thesis, and their permission would be required to republish the thesis and make it freely available on the web. Initially UNSW agreed to work with the UNSW Legal Office and Copyright Officer to develop generic letters and permission forms for the project. However, as the project progressed it became apparent each site had different copyright and thesis deposit procedures and regulations, and there was unlikely to be one generic solution. Further, while it may have been easier to seek approval for inclusion in all aspects of the project at once, UNSW Legal was of the opinion that authors should be advised of, and give permission for, each issue separately.

Issues identified that needed resolution included:

- Authors providing permission for their theses to be available in the project need to be made aware of the implications, i.e. that that thesis will be freely available on the web through ADT, and that the project is a joint one with a commercial partner who may include the thesis in a commercial product such as PQDT, and to give permission for each part separately.
- Where third party copyright (e.g. survey instruments, images) is included in theses, permission is required from the owners of the third party copyright for inclusion, and for the publishing of the thesis to the web. The author is responsible for obtaining this permission, otherwise the material is to be removed and an explanatory note inserted prior to digitisation. As time was limited during the trial, and for privacy and other reasons sites were unable to give ProQuest author contact details, the participating sites generally managed the permissions or removal of third party copyright material.

Logistically, for retrospective conversion, many authors were difficult to trace for permission (See Table 1).

Table 1 – Copyright permission

Institution	Copyright permission requests sent	Granted	No response	Bounced emails	Not granted	% success (rounded)
Adelaide	79	26	50	-	3	33
Griffith	500	*154	333	--	13	31
Swinburne	44	18	26	-	-	41
Melbourne	40	5	33	2	0	13
UNISA	326	*94	215	17	-	29
UNSW	129	*61	68	7	2	47
UWS	22	5	15	1	1	23

* When more permissions were received than required, institutions selected theses on a variety of bases, such as the lack of third party copyright material, order in which permissions were received etc.

3. CURRENT THESES

3.1. Customised deposit form

It was originally envisaged that PQIL would provide one Digital Commons deposit software interface at the ADT level which would be customised for the Australian environment for acquisitions of rights, generation of metadata, and deposit. Participating members collaborated online, working with a document provided by Griffith University, to produce a list of changes that would be required to the existing PQIL deposit form for the generation of metadata. As the project progressed it became apparent that PQIL would be able to develop separate deposit software customised for each participating site.

Actual deposit of theses proved to be a more vexed issue. PQIL had made the assumption that sites would be utilizing PQIL's customised Online Deposit software to load into an ADT full text site (which utilises Digital Commons Institutional Repository (DCIR) software) and that it would be relatively easy for them to arrange simultaneous loading into both the ADT and PQIL. This assumption was incorrect. ADT repositories are not only not DCIR but also not standard, and back end deposit needs to be customised for each site. The participating sites had also assumed that the PQIL software would automatically deposit the theses concurrently into the appropriate ADT repository and the PQIL repository. As the project progressed PQIL advised that in the absence of DCIR, participating sites would be required to build an import filter or manually load the data into ADT. This requirement was not explicit at the beginning of the project, and most sites elected not to develop software for the trial but to manually load into ADT after trialling the PQIL deposit form. This decision would be re-evaluated by sites if the project progresses beyond the pilot as PQIL has indicated that it can build an importing feature if required.

UNSW and Griffith, in early trials of the deposit form, found the process to be difficult and the instructions unclear. Part of the problem was that the permissions needed to be set at

ProQuest, and they were not all set up in the same way. Users not only saw screens different to those in the manual, but also different screens from each other. Issues encountered with the deposit form were numerous, most resolved over the period of the trial. The issues left included the drop down menus of degree names and subject categories which most participants felt while adequate for a trial would need further "Australianising" if the software was adopted.

3.2. Repositories

The initial plan was to showcase theses in existing ADT and PQDT databases to facilitate world-wide access and dissemination. As the project developed PQIL offered to also provide a trial of open access through their Digital Commons Repository portal. This provided the opportunity for the sites to test the Digital Commons Repository and compare it with existing repositories utilised by the ADT.

Initial feedback indicated that areas of satisfaction were the clean look of the repository pages, and the open architecture possibilities. Test metadata was successfully harvested via Open Archives Inititative (OAI)[1] harvesting protocol. Matching and deposit into the ADT shared metadata repository has not been tested as the infrastructure is not yet complete. Areas of dissatisfaction included the layout of the display, the inability of the institution to control the display, and limitations such as the requirement for theses to be delivered in three documents or less.

3.3. Preservation

At the outset of the project it became apparent that the plan to investigate and recommend standards for preservation of born digital theses and explore the mix of paper, microfilm and digital formats needed for 21st Century preservation was beyond the timing, scope and budget of this project. It also required more collaborative face-to-face discussion than the logistics of this project provided. As an interim measure ProQuest are creating microform of all page based material from the project, both from born digital and retrospectively scanned theses. The microform will be housed in the ProQuest vaults in perpetuity. Not all participants saw this as necessary or desirable, however, UNSW is actively exploring this as an alternative to compulsory deposit of a paper copy.

3.4. Other issues

Copyright was also an issue at the current thesis level. While most ADT submission forms have statements regarding originality, authenticity, copyright and third party copyright, it was decided not to include these in the trial submission form as a) sites were not allowing authors to make their own submissions directly so b) authors would have already made statements regarding originality etc. separately. If sites proceed with the Digital Commons Deposit option, then this issue will need to be addressed again.

4. PHILOSOPHICAL AND LOGISTICAL ISSUES

This project represents an interesting collaboration between a commercial vendor and a number of public universities operating in various Australian centres with open access policies under the loose umbrella of the ADT Program. ADT operates in a distributed environment with a commitment to an open access model with a few simple core standard protocols and maximum institutional flexibility. PQIL operates in a commercial environment using proprietary standards. It was apparent that there are differences in philosophy that need to be overcome. At times there was mutual misunderstanding. The project highlighted the importance of clarification and explication of the separate knowledge each had about their environment and products.

The pilot was exploratory and therefore could not be fully scoped. Initial timelines needed constant review as the expected unforeseen issues were understood and resolved. For example, the original scope included, then did not include, then again included a trial repository and individual site deposit forms. Many sites were still working on resolving the copyright issues for both the retrospective and current theses, when it was originally planned for the project to have been completed. This exploratory nature of the project meant that it was difficult to estimate realistic timelines.

Logistically, many of the site coordinators were attempting to find time for the pilot while also performing their permanent jobs. The Project Manager was only able to provide a maximum 10 hours per week. This lack of dedicated time meant that some sites had to pull out of the pilot, and others made slower progress than they would have preferred. The nature of the pilot also made communication difficult. Communications tended to be from the Project Manager to all the sites, but return emails just to the Project Manager. In many cases this was appropriate as the issues were individual to each site, but communication may have been enhanced by a formal collaborative communication space. The need for this is illustrated by several misunderstandings that grew out of lack of communication. For example extensive discussion between the project manager and ProQuest regarding third party copyright clearance was not made explicit to participants until very late in the process. Further the broad geographic spread of participants (around Australia and the USA) meant that most communication was done by email which is often not the best medium for clearing up misunderstandings. In addition time differences between continents, combined with the part-time nature of the Project Manager and site administrators, meant that often a simple problem could take days, even weeks, to resolve.

5. THE GRIFFITH UNIVERSITY PERSPECTIVE

5.1. Introduction

Griffith University was one of the seven foundation institutions in ADT in 1997. In early February 2005, the University decided to take part in the ADT-PQIL pilot. It presented an opportunity to investigate improved deposit processes and repository management solutions for the ADT Program as a whole, with a specific focus on PQIL's versatile Digital Commons repository product (This product is OAI-PMH compliant – which will create new discovery paths via Google and OAI[2]). Developing an in-depth understanding of the technical, logistical and legal issues involved in digitising theses was also of interest. The opportunity to increase the visibility and usefulness of the research output of the University to the U.S. and worldwide market, through digitising archived theses, was also very appealing. Finally, there was curiosity as to what sort of relationships might be made and sustained between the open-access ADT (and ETD movement as a whole) and PQIL, a commercial corporation.

5.2. Planning and Investigations

5.2.1. Pilot Plan

For Griffith, work on the pilot commenced in February 2005. The two main elements of the pilot were retrospective theses digitisation, and prospective theses deposit software investigation. Within those elements our objectives were:

- to retrospectively digitise a number of theses
- to investigate associated copyright issues
- to test a customised version of PQIL's DC software to suit our institution's requirements
- to upload the theses to the ADT[3], ADT trial repository[4], and PQ/UMI[5] .

Griffith pilot staff included a pilot officer, the ADT program co-ordinator and section Team Leader. We determined the hardware we required, the number of theses to digitise, the investigations to be carried out, the type of evaluation to be done at the conclusion of the

pilot, and a pilot timeline. Our expectations were that this pilot would be simple and quick (the initial estimated total pilot time was 9 weeks).

5.2.2. Preliminary Testing

In March 2005, each participating institution was required to review the PQ/UMI Online Submission Form for Theses and Dissertations[6] and perform some preliminary testing of the software. We compared the PQ/UMI form's treatment of metadata elements and features to the ADT form's treatment. Although much of the same information is collected by the two forms, the ADT and PQ/UMI applications were used for different ends. The ADT deposit software is used to create metadata and send attached files to a local institutional server in order to publish the thesis to the local university site, and the metadata is harvested into a central repository to create a distributed full text collection of theses. Students do not incur any costs in this process. The PQ software, however, is used for students to deposit a file to their institution's graduate office to be reviewed and then sent on to ProQuest to upload to ProQuest Digital Dissertations for a charge. Another issue we needed to point out at this stage was that at Griffith, students are not able to self-submit, which ensures efficient, consistent and correct entry of metadata. We would therefore be reluctant to use a public deposit form.

5.2.3. Options for Scanning and Digitisation

Five different methods were considered for scanning and digitisation, including scanning in-house at the library or through Griffith Digitisation and Distribution team; sending theses to the University print service; outsourcing; and sending theses to the U.S. for scanning by PQIL. We decided to send all of our pilot theses to PQIL for scanning, once our concerns about certain issues were addressed, including security whilst the theses were in transit, which party would be checking for and removing copyright-infringing material, whether scanning and courier costs would be covered by the pilot, and how the scanned files would be sent to us for uploading.

5.2.4. Copyright Issues

To scan and digitise a thesis which has previously only been made available in print form, we needed to look at how to obtain authors' permission to digitise their theses for inclusion in an open access environment, and the protection of authors' rights. The release currently used by the Research and Higher Degrees office was not sufficient for the pilot's purposes. The Griffith Copyright Officer advised us on how to address these issues, especially with regard to the rights of the copyright owner.

The use of material included in theses which has been taken from another source also had to be addressed. It was determined that both PQIL and submitting institutions would be involved in performing copyright checking for copyright-infringing material.

5.2.5. Formulation of a Letter and Permission Form to be sent to Theses Authors

We decided to communicate with authors about their potential involvement in the pilot by a hard copy letter with a permission form for the authors to complete. In early June, we commenced the mail out to 500 authors. Our response rate was 25.2% before the cut-off date of 1 July, and the total response rate at the time of writing was 33.4% (See table 2).

Table 2 Response/Acceptance rates to mail out					
Responses before cutoff date	126	25.2%	Acceptances before cutoff date	115	23%
Responses after cutoff date	41	8.2%	Acceptances after cutoff date	39	7.8%
Total responses	167	33.4%	Total acceptances	154	30.8%

The response rate for a previous retrospective pilot at Griffith (in 1999-2000) was approximately 10%.

5.3. Scanning and Digitisation

In selecting the theses which we sent to PQIL, our preference was for theses for which the author stated that there was no material taken from another source, the thesis was not bound (for a better quality scan), those that agreed to uploading to both ADT and PQIL, and those for which the permission forms were received earlier. In early July we dispatched 50 theses to PQIL offices in the U.S. for scanning. Estimated turnaround time was 6-8 weeks. We decided that the most efficient method for file transport was for institutions to download the files from PQIL via FTP. We received the first 2 batches of PDFs (16 theses' PDFs) via FTP in late August (7 weeks after we dispatched the paper copies).

5.4. Upload of Theses to the Trial Repository and to ADT

5.4.1 Training of Staff in Use of Software; Formulation of Procedures for Uploading

In July, pilot staff participated in a teleconference training session led by PQIL staff in the U.S. on the use of the DC customised software. Some issues regarding entry of metadata into the customised submission form raised during training and the software trial period were: how to accommodate for earlier dates of theses and the range of degree names used by Australian universities, PDF file naming, how to include supplementary files, and problems encountered with permissions and access. Generally support was good; however at times responses and follow-ups were slow.

We also needed to investigate how best to prepare the metadata and PDFs for upload, including preparing a text version of the abstract from an image, setting security, and ensuring file sizes were appropriate (i.e. that they met ADT standards, and that the files were not so large that they would take too long for users to download for viewing).

5.4.2 Testing of Customised Digital Commons Software

The estimated timeline for testing of software was 2 months. We began testing once we received the first batch of PDFs from PQIL in late August, and at the time of writing (mid-September) we were still testing and uploading to the trial repository and ADT.

During the trial period, the files of the scanned theses were FTPd from the PQIL server to a local PC. The PDFs and a text version of the abstract for each scanned thesis were then prepared and uploaded to ADT. The theses were then uploaded to the trial repository and holdings added to the library's online catalogue. Upload to ADT was straightforward, however as the scanned file sizes were very large (ranging between 4-10 MB); we had to split files into more acceptable pieces. This resulted in many titles having numerous PDFs attached, which could be difficult for users to manage. When uploading to the trial repository using the customised DC software, we experienced problems with access to certain options on the submission form. There were also issues with how the form would accommodate the multiple large PDFs we needed to upload. However, at the time of writing we had been able to upload approximately 20 theses, and the results, which can be accessed on the trial repository website[7], appear to be very user-friendly.

5.5. Conclusion

5.5.1 Challenges

This pilot posed a number of challenges. As time went on, the nature of the pilot changed, and its complexity increased. The pilot changed from a software testing exercise to a complex and ongoing investigation of many issues, including:

- *Quality control:* We were committed to the maintenance of the quality of the service that we are already offering in terms of providing digital access to research and higher degree theses. Until testing is complete and the trial period is at its end, we will not be able to make a true comparison between the viability of the use of the customised DC software, and the current situation (use of Apache software to upload all of Griffith's digitised research and higher degree theses to ADT), which has been functioning very well since the ADT Program began.
- *Communication problems:* The difficulty of communication between geographically distant participants led to delays and misunderstandings.
- *Preparation:* Much of the pilot time was taken up with investigation of issues that may have been more efficiently performed by a smaller, dedicated group before involving participating institutions.
- *Changes:* There were changes in what was being offered to participating institutions, which affected progress.

This ongoing investigation of issues meant that there were numerous resulting timeline extensions (and therefore the pilot budget grew accordingly).

5.5.2 Outcomes

Griffith has gained a great deal from being involved in this pilot, in particular:

- the copyright investigation which we undertook provided us with useful information for future planning
- we have gained a knowledge of the processes involved in digitising retrospective theses
- we have expanded our repository of digital theses
- we have gained increased exposure of some of our important archived research
- we have increased the profile of the University on a world-wide stage.

5.5.3 Implications for Griffith University

If Griffith adopts the new models for the deposit and hosting of our theses, there would be a number of implications for the University. PQIL have indicated that in the near future, they will investigate the application of plagiarism-detecting software to their database of digital theses. This would be optional for universities which submit theses to PQ/UMI. Griffith would need to investigate further, as it would have implications for the process of lodgement of theses through the Research and Higher Degrees office of the University and the conferral of degrees to students.

Additionally, if Griffith were to enter into an arrangement with PQIL for deposit and hosting of theses, we would need to revise our processes for deposit, including changes to the Lodgement form currently used, in order to notify higher degree graduates of information about the upload of their theses to PQ/UMI. Furthermore, there would be charges for uploading theses to the PQ site, which would have a great impact on the Library's budget. Presently, there is no charge to institutions using the ADT site for the uploading of theses.

5.5.4 The Future of Digital Theses at Griffith University

As testing has not been completed at the time of writing, the future of digital theses at Griffith is somewhat unclear. What is clear, however, is that there will always be a demand for open access to our research generated by research and higher degree students at Griffith University.

What remains uncertain is how we make that research available to the community. Any changes we make to the processes and software we use for the digitisation of theses must produce as good quality results as we presently get, if not better. The new service would have to offer outstanding technical support and greater exposure and accessibility

to our digitised theses, particularly as a major consideration in the decision to change processes would also be that the commercial option comes at a cost.

6. CONCLUSIONS

The project has been an interesting learning experience for ProQuest, the ADT and the participating sites. It illustrates that a cooperative project that utilises the different core competencies of libraries and publishers can succeed, and may be a precursor to future collaborative projects.

---

[1] http://www.openarchives.org/

[2] From Digital Commons@ website: http://www.il.proquest.com/umi/digitalcommons

[3] http://adt.caul.edu.au

[4] http://ariic.library.unsw.edu.au/

[5] http://dissertations.umi.com/

[6] http://dissertations.umi.com/students.html

[7] http://ariic.library.unsw.edu.au/griffith/