Unifying ETD with open access repositories

Arthur Sale
Professor of Computing (Research) and Research Coordinator
School of Computing, University of Tasmania, Hobart

Keywords: institutional repositories, open access, unified archives, user feedback, open source software.

ABSTRACT

The fundamental proposition of this paper is to argue that ETD collections should be unified with institutional open access repositories, where they sit alongside other forms of research output. This change from separated collections confers the following benefits:

- Better accessibility and searchability, leading to greater impact and citation rates
- Better archival processes
- Value can be returned directly to the thesis authors to help in their research planning
- Easier provision of bibliometric information about access
- Lower operating costs through use of one software system, one data store, and reduced training
- Redundancy in Internet access paths.

The paper examines each of these claims and substantiates them. The practice of separate ETD collections may have arisen from paper-based practice, and the paper argues that this should be challenged.

An actual implementation of a unified ETD collection is also discussed, with actual data on its performance. Software was written that allowed the Australian Digital Theses Program (ADTP) to harvest research theses from a unified institutional repository running GNU Eprints. This software was tested with ADTP, and will be used by the University of Tasmania. The University of Melbourne has already adopted the software and serves up theses to ADTP from its own Eprints repository. The software is available to any university under a GNU open source license.

This implementation has allowed the measurement of accesses and a range of useful data, which are available to the thesis author. Examples of this type of data and the information that can be derived from it will be discussed. Since some theses are available on the Internet via ADT, the UTas repository, and the ARROW Discovery Service, comparative information on searchability can be demonstrated.

1 CENTRAL ISSUE

In the rest of this paper, *theses* will be used to denote theses for research degrees, dissertations or exegeses. The terms *university* and *institution* will be used interchangeably depending on the most common usage.

The tradition in paper-based libraries was and is for theses to be collected by the university in bound paper form but not published. Their availability for reading, borrowing or copying is advertised in the university's library catalogue, often in a named and segregated collection. Frequently, theses from a selection of universities were indexed and catalogued in a 'Union List of Theses'. These practices were sensible, since paper theses are generally

irreplaceable, and therefore demand tight control over access. These practices largely continue, and appear to have influenced the treatment of theses in an evolving digital world.

In Australia for example, digital copies of theses are held almost exclusively in local institutional data stores, separate from all other data (the separate collection theme). The Australian Digital Thesis Program (ADTP) provides gateway services for the totality of Australian theses and a few other institutions (the Union List theme).

The central issue of this paper is to argue that it is time to re-evaluate this practice, and that digital theses should be unified with all the other forms of research output from universities. In other words, they should be held in shared data stores and archived together. Separate gateway services can continue to be provided but need not be.

2 THESIS CREATION

Nearly all theses are now created digitally in all disciplines in a university. A small minority may include some non-digital information, such as a pasted-in photograph, the printed output from a scientific instrument, an original art work, or perhaps a handwritten musical score. There is therefore no technological or cost barrier to a university capturing the electronic form of a thesis, since it involves the student in no extra cost (or at most, the cost of a CD-R at $5.00). Exemptions and procedures could be put in place for the small minority of theses which contain non-electronic material.

Collecting an e-form of the thesis does require appropriate permission, as the copyright belongs to the graduate. However, universities have been getting around this for at least a century, by making the submission of a bound paper form of the thesis for the library a requirement of graduating. Similar rules could be applied to require an e-form for the university's ETD repository. It is therefore surprising how diffident many Australian universities are to take this step. No coherent arguments have been made to justify this caution, and students are far less worried about the issue. Not that they would stretch themselves to voluntarily supply an e-form, but if it was mandated they would happily comply (Swan & Brown, 2005).

Holding an e-form of a thesis with appropriate copyright license permissions immediately opens the way for universities to make their theses more accessible by others over the Internet. Parenthetically, it remains dubious whether paper submitted theses can be retrospectively digitized and supplied online on-demand without going back to the authors for permission. Some university libraries seem to assume that this was implicit in the paper-based copyright authorizations, but this assumption is legally dubious.

Finally, e-theses should be considered: theses designed for electronic viewing which cannot be reduced to paper, or which would lose impact if so reduced. Such theses may include animations, multi-media, software, or be organized in a hyper-linked fashion. Again, Australian universities seem to be slow in opening this door to the digital world. It is the PhD candidates who suffer as a consequence though inability to express themselves naturally.

3 UNIFICATION

3.1 Accessibility and searchability

There are two aspects to the impact of unification on accessibility and searchability. The importance of each depends on the particular scheme adopted for separate repositories.

Firstly a unified repository is likely to be better indexed, better visited by robots, and generally ranked higher by search engines than two separate repositories. This isn't necessarily so, but on balance of probability the doubling of work to make two repositories equally accessible makes uniformity less likely. In some cases, the outcome is quite clear; the local ETD repository is not visited at all by Internet robots. The converse is unlikely, since the prime purpose of institutional OA repositories is accessibility, and their data stores are usually much larger than ETD stores thereby warranting greater effort.

Secondly the treatment by search engines of gateways (such as ADTP, ARROW) and leaf repositories is quite different. Gateways carry only metadata about the thesis (similar to a catalogue), whereas leaf repositories hold both the metadata and the full text of the thesis. Good search engines such as Google have access to both the metadata and to the full text,

and use *both* to index the reference. The net result is that the leaf repository full-text often ranks higher in search engine result lists than gateway metadata.

3.2 Better archival processes

The provision of persistent URLs and a strategy to ensure that materials are planned to have continued online access are important features of both ETD and OA repositories. A unified repository ensures that what applies to one applies to both, and the outcome is likely to be better. In the UTas case, the option of annually copying material to the Tasmanian STORS archival/deposit repository is an example.

3.3 Value-adding services

The traditional thesis service provides no value-added return to the author after the thesis has been lodged. The thesis resides in the university library and any accesses are usually not known to the author. No further communication takes place or is planned for.

There is no reason why ETD services should be the same. The provision of information as to the readership and interest in the thesis is valuable information to the beginning researcher, and may well result in international linkages. These have been experienced in several of the cases of theses in the UTas server. As discussed later, the UTas server has also implemented value-added services which have been well-received by researchers.

Again, the issue of scale suggests that a unified service is likely to provide a better service. In addition lodging a thesis in a unified ETD/OA repository offers the opportunity to induct the beginning researcher into electronic dissemination, with major benefit to his/her future career. This aspect of research training is an important part of PhD study.

3.4 Bibliometric information

Online servers provide a wealth of bibliometric information. Counting 'reads' (or more accurately full-text 'downloads') is a quite simple exercise, as is automatically analyzing the resulting data. The relation of reads to citations is the subject of several recent research papers (see Brody, 2004). Certainly the availability of bibliometric information emerging from OA repositories is a significant research resource, and extending it to ETD repositories is clearly important.

To extend this further, a £151,000 research project in which the author is involved aims to track global access usage of OA repositories by search-source university and scientific journal or publisher. It would be a simple matter to extend this to theses with the degree-granting university regarded as the publisher.

3.5 Lower operating costs

The unification of ETD stores with OA repositories results in lower operating costs for institutions for the following reasons.

3.5.1 Software costs

Operating one set of software rather than two will obviously result in lower software support costs and lower costs of backup and updates, provided only that the integrated software used is one of the two systems that might otherwise be used. Evaluating the relative software acquisition costs is not as easy, but these in any case are usually negligible compared to the labour costs. Most institutional repositories run largely on open-source software. Of course, whether these costs are borne by the central ICT service, the central library, or elsewhere is irrelevant to the institution.

3.5.2 Hardware costs

Hardware costs are likely to be smaller, unless both applications (ETD and OA repository) run on a single machine (which is quite possible). Hardware costs are also small relative to the labour costs, even for large disk storage.

3.5.3 Training and user support

This is probably the largest cost component, being composed mainly of labour costs in support of ongoing content addition. Training will be simpler with one system rather than two, both initial and ongoing. It is more difficult to make such simple assertions about user support. It seems likely that supporting a single system will be less costly than two, but this ignores the possibility that one or both systems are tuned to make their content type easy to support. The experience at UTas is that this is not the case, and a unified system offers the possibility of support significant savings.

3.5.4 Management

The cost of managing two systems is likely to be larger than one. This is true whether the management is purely in-house supervision or involves university committees.

3.6 Redundancy of Internet access paths

In Australia, with a few exceptions, Internet accessibility of ETDs relies mainly on the Australian Digital Thesis Program, a metadata gateway. Failure of the ADTP server, less active promotion of its service to search facilities, or unawareness of the service, means that a thesis may not be found by searchers. The ADTP service for example is little-known to searchers outside Australia and New Zealand, and not well-known to searchers even within Australia.

A unified ETD/OA repository, on the other hand, can both feed data to the ADTP and be harvested directly by search engines. If one path/service is down, the others provides an access path at least to the metadata and probably the full-text. In addition the existence of alternate paths provides a market-force stimulus to specialized services like ADTP to strive to keep up with best practice. As with all market force situations, a service that does not evolve to keep up with user requirements will eventually disappear or become subsidized 'in the public interest'.

To take an example, PhD theses on the UTas and UniMelb servers are directly indexed by all the standard search engines, but in addition they are harvested by both ADTP and the ARROW Discovery Service which are themselves harvested by search engines. This gives multiple opportunities for searchers to find a particular thesis.

4 UTAS ADTP FEEDER

4.1 Eprints server

The University of Tasmania operates an institutional server for its research output which has been operational since June 2004. The server is in process of transition to a major university service. The server runs on the GNU Eprints software, which is OAI-PMH standards-compliant. The server contains published journal and conference papers, PhD and other research theses, and First Class Honours theses. It is registered with and harvested by all probable search engines that a searcher might use.

4.2 Interface software

Given the existence of an Australian ETD gateway, it was important to make sure that the content of the UTas server was harvested by ADTP. This posed two problems:

(a) ADTP does not at present use an OAI-PMH standard form of harvesting (which Eprints provides), with XML as the harvesting interface. Instead, it relies on an HTML robot that explores a subsite through hyperlinks, and uploads to its database data from the HTML metatags in all the documents it finds.

(b) ADTP does not want to see all the documents in the server; it only wants to see the metadata for each research degree thesis.

To resolve these issues, a small piece of software was written by Sale (2005) as a *cron* job, running once per day (the frequency can be reduced). This software (*EprintsToADT*) scans all documents in the database and for each that satisfies the criteria of both being a thesis *and* being a research degree thesis, writes a page containing the metadata in non-visible HTML

metatags in a directory (*ADT*). Subsequently it also writes an index page containing links to all the generated pages into the same directory.

The ADTP harvester is given the URL of the index page, and from there it finds all the metadata pages in the ADTP format (and no others) and it indexes them in a daily basis. Care was required to ensure that the harvester could not escape from the ADT directory through unwanted links, otherwise it would traverse throughout the Eprints server, trying fruitlessly and mistakenly to index all the pages in the institutional repository. The software was tested in conjunction with ADTP, and after initial debugging ran flawlessly.

4.3 Take-up

The University of Tasmania has taken a cautious approach to this software and decided to consider implementing it after the launch of its university-wide Eprints service (estimated for mid 2005). However, a significant number of trial theses were uploaded to both Eprints and the local ADTP repository, allowing for evaluations.

The University of Melbourne, however, decided to adopt it and over late 2004 worked to close their previous ETD repository and transfer the theses to their Eprints server. In January 2005, the University of Melbourne was regularly serving up theses to ADTP and continues to do so. In the meantime, the theses are of course being served also to Google and ARROW amongst others though more usual pathways.

5 VALUE-ADDING FOR THESIS AUTHORS

A variety of value-added services are provided to thesis authors on the UTas server through in-house written software (which is also available as open source software for others). These are a direct consequence of piggybacking off the main effort to support an open access repository. The purpose of providing value-added services (for the repository operator) is to increase user satisfaction and increase take-up of the service. The authors gain the following benefits.

5.1 Increased impact

Each thesis is exposed to the Internet from the leaf repository, indexed by all the major search engines, and can be expected to be read much more often then paper theses lodged with libraries. This translates directly into impact, since the thesis is more likely to be cited. Of course this apples to specialized ETD repositories and gateways as well, provided their accessibility is as great as for OA repositories, which is not usually the case. In addition, the thesis can be viewed in the context of other research output in the same field and university, thereby building a coherent picture of the local research strengths.

5.2 Accessible monitoring

Any author can access the statistics related to their own thesis, at any time. These statistics provide data on the country of origin of the viewers, a time series analysis, and whether the viewers accessed the abstract or the full-text of the thesis. Some authors are known to collect this data for promotion purposes or research impact studies. Surges in interest provide a wealth of useful information and can sometimes be linked to new conference presentations or article publications, not necessarily by the author. For example, a UTas author recently gave a paper in the USA, and one month later there was a marked increase in the download rates for his other papers, including ones that had been on the server for a year. See for example:

http://eprints.comp.utas.edu.au:81/es/index.php?action=show_detail_eprint;id=59;year=0;month=0;range=4w.

It is assumed that the events are linked, which is borne out by the country of origin of the requests (largely USA at the time of writing).

5.3 Research planning

The data mentioned in 5.2 can indicate a level of interest in the topic to a researcher, and thus may assist them in planning their future research activity or publications. Comparative data on accesses can lead to improved self-practices in keyword and metadata choices.

5.4 Website maintenance

Several staff have adopted the practice of providing links to the Eprints full-text of their papers/thesis on their personal website, thus facilitating searcher lookup. In the University of Tasmania such links are also provided on the University's central research data server WARP. Other staff upload every old paper if they get a request for a reprint/copy.

The more adventurous have simply replaced the lists of publications on their personal home pages by a 'search link' on the eprint server. This ensures that the publications list on their page is always up to date (provided the Eprints uploads are up to date). Papers that were published pre-Eprints or deliberately excluded from the archive for good reason still need to be listed on the website.

6 CONCLUSIONS

- The time has arrived to integrate Thesis Repositories with general Open Access Repositories in all universities.
- Provision of an e-form of a thesis should be made mandatory (or routinized at least) in all universities (Swan & Brown, 2005).
- Training in the presentation of and uploading of e-documents to OA repositories (including papers and theses) should be built in to the generic skills to be acquired by PhD candidates.
- Value-adding services need to be devised and provided to thesis authors.

7 REFERENCES

Brody, T (2004). Citation Analysis in the Open Access World. *Interactive Media International*. Retrieved from http://www.ecs.soton.ac.uk/~harnad/Temp/timIMI.htm, 15 June 2005.

Brody, T. and Harnad, S. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine 10*(6).

Sale, AHJ (2005). De-unifying a digital library. *First Monday*, *10*(5), 2 May 2005. Retrieved from http://www.firstmonday.org/issues/issue10_5/sale/index.html, 15 June 2005.

Swan, A and Brown, S (2005). Open access self-archiving: An author study. Technical Report, Joint Information Systems Committee (JISC), UK FE and HE funding councils. Retrieved from http://cogprints.org/4385/ and http://www.ecs.soton.ac.uk/~harnad/Temp/alma-amst.pdf, 15 June 2005.