Electronic thesis development at Cranfield University

Simon J. Bevan
Author: Simon Bevan is the Information Systems Manager at the Information and Library Service, Cranfield University, UK. E-mail: S.Bevan@Cranfield.ac.uk

Abstract

Purpose: To describe the issues involved in the introduction of mandatory submission of electronic theses at Cranfield University.

Design/methodology/approach: Background information on how the availability of e-theses has developed at Cranfield University is provided along with discussions and advice on issues such as the choice of software, thesis submission workflow and timeframes, particularly in relation to the publication of thesis-related articles. It also looks at metadata issues as well as both retrieval and usage of electronic theses. Finally it describes how the service has expanded from e-theses to other types of material and to the development and expansion of an institutional repository for Cranfield.

Findings: It is shown that there are a number of issues that will need to be addressed from the points of view of librarians, academic staff and registry staff and that one effective method of managing the process is to set up a working group with all stakeholders in the process. There is a clear need for administrative procedures to be discussed in detail and a recognition that the time involved in changing regulations may be significant.

Practical implications: It is clear that most of the issues that have arisen at Cranfield as outlined in the paper will be mirrored at other institutions that are considering the same changes and so those institutions looking at the area of e-thesis submission may gain some useful insights.

Originality/value: This paper provides useful advice on the issues that will arise as institutions go through the process of introducing the mandatory submission of electronic theses.

Category: Case study

Keywords: Electronic thesis submission; E-theses; ETD; DSpace; Institutional repository; Academic libraries; UK

No. of words: 5,116

1. Introduction

The major recent impetus in the development of electronic theses (e-theses) in the UK has come about via funding made available by the Joint Information Systems
Committee (JISC). Looking at the strategic aims of JISC, it is clear that access to e-theses is relevant in a number of areas, in particular improving the effectiveness of scholarly communication, promoting wider participation (perhaps by encouraging a new generation of students interested in creating theses of a non-traditional, non-linear nature), and sustaining a dynamic world class research sector. If research is available more easily, more widely and more quickly, then this will have a positive effect on dynamism in the research environment.

2. Electronic theses at Cranfield University - background

As an almost wholly postgraduate institution, the creation of, and access to, new research is of central importance at Cranfield University. The vast majority of students are engaged in the creation or development of new research – over 90% of students are postgraduates and consequently have to submit a thesis. As a consequence, for a number of years the Library and Information Service (LIS) has been involved in the development of electronic thesis availability both locally and nationally.

Cranfield was a founder member in the mid-1990s of the University Theses Online Group (UTOG), a loosely assorted group of librarians from a number of very different higher education institutions in the UK interested in widening access to this research material. The group was not funded in any way until later in its life (when very limited funding was made available) but it served as an essential first step in encouraging communication and in the forging of relationships that would later become important (Lomax, 1997).

Between 1994 and 1997 the Cranfield LIS, as a partner in the European Initiative in Library and Information in Aerospace (EURILIA) project, was involved in a limited thesis scanning project (http://www.cordis.lu/libraries/en/projects/eurilia.html). Funded under the European Commission’s Action Programme for Libraries the project developed a Z39.50 client to allow the searching of partners’ different online library catalogue systems (O’Flaherty, 1995). Where a relevant thesis record was retrieved from the Cranfield catalogue, users could click a link allowing them to access the full text of over 200 scanned PhD or prize-winning MSc theses.

Cranfield LIS was also involved in a pilot project in 2001 in collaboration with Index to Theses, which tested a mechanism for the uploading of both thesis metadata and full text. This was a limited project that provided very useful experience for all involved and although successful, has not been developed into a service.

It was in 2002 that JISC, realising that theses provided a largely untapped wealth of information, funded three projects relevant to the development of e-theses in the UK under the Focus on Access to Institutional Resources (FAIR) programme (http://www.jisc.ac.uk/index.cfm?name=circular_1_02). These were:
- Electronic Theses (http://www2.rgu.ac.uk/library/e-theses.htm) led by The Robert Gordon’s University (RGU) in Aberdeen. This project involves a consortium of organisations developed initially through contact within UTOG;
- Theses Alive (http://www.thesesalive.ac.uk/) at Edinburgh University (MacColl, 2002, Andrew, 2004);
• Daedalus project (http://www.lib.gla.ac.uk/daedalus/) at Glasgow University
  (Ashworth et al., 2004). Daedalus is mainly a project on Institutional Repositories, but there is a theses ‘strand’.
Cranfield University, along with the University of Aberdeen, the British Library and the University of London is a member of the consortium of the Electronic Theses project. It had been keen to move to electronic thesis submission and it was felt that by being involved in a collaborative project and developing local procedures at the same time, each would benefit.

3. Why e-theses?

In general terms there may be a number of local reasons why universities wish to move into the development of, and provide access to, electronic theses collections. Space has always been at a premium in many libraries, although with the development of e-journals and with library policies of providing access to electronic only journals, as well as the gradual increasing access to electronic books, space is being made available elsewhere.

It is unlikely to be the main driver and it may not be the case at many institutions, but moving to e-thesis submission may allow the overhaul, rationalisation and streamlining of administrative thesis submission processes within university registries.

The main reason for moving to e-theses at Cranfield is to try and ensure that a huge untapped information resource is made available more widely. In the past, theses have sat on the shelves gathering dust and since cataloguing may have been limited or of poor quality, these resources were all but invisible to the researcher. Suleman and others (2001), of Virginia Tech in the US, describe the development of the Networked Digital Library of Theses and Dissertations (NDLTD) which is a collaborative effort of universities around the world to promote creating, archiving, distributing and accessing Electronic Theses and Dissertations (ETDs).

Other arguments given for e-theses development include the idea that such a development will benefit graduate education in terms of graduates understanding the developing world of electronic publishing, copyright, intellectual property and so on. Another important push is that students may wish to develop new ways of creating access to their research – they may wish to create an electronic work for which there could be no paper equivalent. This may affect all kinds of disciplines from the arts (e.g. performance, or artistic design) to engineering with the embedding of computer programs or models.

Although not one of the main reasons, an important by-product of the creation and availability of electronic theses at Cranfield has been the input into the development of the University’s institutional repository.

4. Setting up an e-theses collection at Cranfield University

4.1 First steps
The initial steps at Cranfield started with desk research, looking at various Web sites in the US, where a number of universities have been accepting electronic theses for many years. There is also useful information on the various ETD conference Web sites at:

- Brigham Young University, Utah in 2002 (http://ce.byu.edu/cw/etd2002/)
- Humboldt University, Berlin in 2003 (http://www.hu-berlin.de/etd2003/)
- University of Kentucky in 2004 (http://www.uky.edu/ETD/ETD2004/).

These Web sites have links to papers describing all aspects of the development of electronic theses.

The next step was to make contact with stakeholders at the University and to draft a feasibility study for submission to Teaching Committee. The proposal was accepted in February 2002 and an invitation given to return to the same committee in November 2002 where complete proposals and submission recommendations would be considered.

As soon as the initial proposal was accepted, the LIS took on the further responsibility of forming a Working Group to investigate all aspects of e-thesis submission in detail. Clearly the membership of such a group was important to its success and to its authority. It was felt that it was essential that a number of stakeholders were involved.

As the holder of theses and the service that had the aim of further disseminating university research, the LIS took on the lead role of chairing the group. Other services involved were:

- the Registry, the owners of the submission process;
- the Computer Centre for its expertise in training, software and format advice;
- Cranfield Press for its knowledge of publishing, scanning and so on.

So with central services well represented and with input from the President of the Students Association, the final and most important representation and input would be provided by a member of academic staff on the Working Group.

One of the most important factors was to ensure that the group worked within certain terms of reference. At a couple of stages there was a danger of discussions moving into more general areas of thesis submission and it was felt that the role of the group should be kept relatively narrow. Suggestions on changing or rationalising bibliographic reference formats, for example, were felt to be outside the remit of the group. Obviously ‘terms of reference’ can be defined at the start and so if an institution wishes to look at all areas of submission, then this may be an opportunity to re-engineer the whole process, as long as this is made clear to the reporting committees at the outset. Where an issue was felt to be important but impossible to answer in the working group, it was referred back to Faculty Boards for discussion.

4.2 Software decisions

It may well be important to address the software question as a first step. At Cranfield, the initial idea was to take the two most mature open source, Open Archives Initiative, Protocol for Metadata Harvesting (OAI-PMH) compliant systems available at the time – the Virginia Tech (VT) e-theses software and the GNU e-prints software from
Southampton - and compare them with what was at the time a new IR software solution, DSpace, developed by the Massachusetts Institute of Technology (MIT) and Hewlett Packard (HP). Each system has different advantages and disadvantages depending on institutional requirements. In the end, at Cranfield, there was simply no time to undertake such a piece of detailed research as had been planned and so the decision was based on a number of core issues.

The VT software was designed for e-theses. At Cranfield it had always been a possibility (and part way through the planning process it became a certainty) that it would be necessary in the future to think about adding other types of material to the system, so clearly software designed for a single type of resource was not an ideal solution. DSpace quite clearly had a significant amount of funded development behind it, and it was perceived to be a more advanced system than the other main contender – GNU e-prints. The preservation issue had also been considered in the development of DSpace. Just as Cranfield was making a local decision on which to choose, independently, the FAIR projects had all largely come to a similar conclusion, and although in a couple of cases there was a wider remit to look at other solutions, DSpace was clearly the leading contender and the favoured choice by the technical teams.

Jones (2004) reports on the similarities and differences discovered between the VT software and DSpace as part of the Theses Alive project, and Nixon (2003) describes experiences in using Southampton’s e-prints software and DSpace in the Daedalus project.

4.3 Workflow

At Cranfield one of the decisions at the outset was that the LIS would be responsible for loading electronic thesis metadata to the server (reflecting the fact that the LIS is currently responsible for cataloguing theses on the library management system). Direct student upload to a server was discussed within the working group (of the sort trialled in the Index to Theses pilot). It was assumed that this would be the eventual method of receiving electronic theses, but at this first stage, prior to the adaptation of the workflow within the DSpace software, it was felt that the Library should have responsibility for cataloguing.

A related question is: at what stage is the e-thesis to be submitted? Should it be submitted up front prior to the marking, or should it be required at the end of the process once all corrections have been completed. The difficulty with the first method is that in the vast majority of cases, the copy of the thesis submitted is different from the copy of the thesis that is eventually passed. If there are different versions how do you manage version control? It was felt that at this stage, the need for version control would simply add to the complexity of a new process. One comment made during discussions was that there was simply no utility in having a copy that wasn’t the final copy (assuming that academic staff were still going to mark paper copies of theses).

4.4 Timeframe and article publishing

What is the proposed timeframe for making the theses live and accessible? Relevant here is the issue of the publication of thesis-based articles. There is US evidence that
publishers rarely refuse publication of an article based on an electronically available thesis.

In an article on the issues of whether Web availability of theses counted as prior publication, McMillan (2001) reports that the conclusions of a survey carried out at Virginia Tech found that “100% of those who had successfully published had not had any problems getting published because their theses or dissertations were online and readily available on the Internet”. Certainly, according to surveys carried out by Seamans (2001) of Virginia Tech and Dalton (2000) of the University of Windsor, Canada, some publishers do see ETDs as constituting ‘prior publication’, although in Dalton’s case only 14% stated they would not publish works derived from ETDs, while 82% of Seamans’ respondents said that they would be willing to publish such work. Interestingly, some publishers (Elsevier, Academic Press) encourage their authors to link from their journal articles to personal Web sites.

Text from the specification for electronic thesis submission for graduate students at Johns Hopkins University (http://etd.jhu.edu/etdpublic/howto/publisher.html) states: “it is the extremely rare dissertation that is published as a book without major revisions. Most authors spend several years rewriting and developing the ideas and argument in their dissertation. The more your book manuscript differs from the dissertation, the less it will matter to the publisher whether or not the original dissertation is available electronically. Where the dividing line is at which the original dissertation becomes irrelevant is a matter of judgment for each individual publisher, but it is probably wise to assume that by making your dissertation widely available electronically you are somewhat raising the threshold of acceptance for that later book manuscript.”

4.5 Cataloguing and metadata

It is likely that, as well as loading electronic theses into a separate repository (whether just for theses, or for more general institutional material), a library will also wish to continue to make them available within the standard library catalogue – this is certainly the case at Cranfield. Given that the most sensible approach would be to download from one to the other, a decision has to be made whether to initially create a record on the catalogue or in the repository. In the Cranfield case, and probably for all other DSpace users, it makes more sense to create the DSpace record first, because, as part of the creation process, DSpace creates a persistent URL in the form of a ‘handle’. This decision will mean harvesting metadata from DSpace for loading into the library management system used at Cranfield, namely the Sirsi Unicorn system.

A significant outcome from the Electronic Theses project led by RGU was the creation of a core set of metadata for e-theses (Electronic Theses, 2002). Other fields are also specified as being optional. Clearly it is essential that fields for both catalogue and repository systems should be borne in mind when creating the standard record.

How will metadata be collected if LIS staff are to be undertaking the cataloguing processes? At Cranfield, the idea of the creation of a ‘thesis documentation page’ as shown in Figure 1, came from experience of collecting technical reports.
Take in Figure 1

Figure 1 Thesis documentation page

<table>
<thead>
<tr>
<th>Thesis documentation page</th>
<th>Cranfield University</th>
</tr>
</thead>
<tbody>
<tr>
<td>This form is designed to bring together all of the metadata that describes the Cranfield thesis. All fields need to be completed and this form should appear as the final page of the thesis.</td>
<td></td>
</tr>
</tbody>
</table>

1. Thesis author

2. Student number

3. Thesis title and subtitle

4. Year of submission

5. Level of thesis (PhD, MSc, MBA etc.)

6. Supervisor(s)

7. School

8. Department

9. Abstract

10. Keywords

The thesis documentation page brings together all descriptive data about the thesis. Use of such forms will help the cataloguer to avoid having to do the equivalent of leafing through an electronic document trying to find all required cataloguing.
information and will have the added benefit of students being asked to supply free-text subject keyword terms. If libraries are considering the use of more controlled subject terms, then this is something that needs to be considered further, i.e. who will provide these controlled terms, could a relevant thesaurus be made available for students to use?

4.6 Plagiarism, format checking, copyright

Where a thesis is made available electronically, it will now be possible to run it through plagiarism software. This is something that an institution will wish to do far in advance of the final award, so once again would need to be fitted into the workflow. Cranfield is just beginning to look at integrating this process although it is more likely to be done during marking by academic staff (where academic staff may have asked the student for their own e-copy of a thesis to help facilitate this process). Another issue - it is important to ensure that the workflow for restricted theses is effective in ensuring that such works remain out of the public domain until the expiry of the restriction.

At some stage in the process the thesis will need to be checked for formatting (such as references format, margins). A cursory check (cursory because students should have been made aware of the legal issues already) for third party copyright material may be worthwhile here. At Cranfield, these issues can be addressed by looking at the paper copy submitted up front, but clearly if a different submission model is chosen, then this will need to be considered. In terms of permissions for making the thesis widely available, in common with a number of other universities, Cranfield holds the copyright on student theses and so the licensing issues are less onerous than they may be at other institutions.

Is there a process by which problems with PDF conversion can be identified? The solution at Cranfield was to request a copy in the native word-processing format, so that where a converted file failed to open or had pages missing, then a new conversion could be undertaken after the student had left the University. In fact a brief pilot undertaken at Cranfield showed that technical issues were unlikely to be problematical.

Perhaps the easiest method of filling an e-thesis repository is to undertake a retrospective scanning process. Often this is difficult (or rather, more expensive) where the thesis has been bound, but there are some institutions (for example the New Jersey Institute of Technology) that scan new, unbound theses. This is one method of ensuring that the thesis that was submitted, marked and corrected, is the same as the electronic copy that is made available. At Cranfield, most of the e-theses available were acquired in electronic format, but in a few cases, theses were also scanned.

4.7 Timescales

Cranfield first started to look seriously at the development of e-theses submission processes in late 2001 and finally had the Senate minutes confirming the rule change in July 2004. It is common sense that those responsible for steering through such changes need to be aware of which official committees need to be consulted and persuaded of the merits of such a change. It almost goes without saying that it is also
important to consider and develop unofficial channels and personal contacts in order to ensure academic support.

There also needs to be an acceptance that, even though there is general agreement about the concepts, there is liable to be lengthy discussion about the detail, and a need to re-visit the same committees throughout the process. It is almost inevitable that there will be delays in the progress - speed may well depend on the institutional culture and whether the idea has reached the collective psyche of the university.

4.8 Firm decisions

There were a number of agreed outcomes for Cranfield:

- The current thesis submission workflow would not change until right at the end of the process when students would be mandated to submit an electronic copy of their amended and corrected thesis. Students will continue to submit three paper copies.
- Submission will be mandatory for all research degrees (PhD, DBA, EngD, MPhil, Masters by research).
- Initially students will be advised to submit to the Registry a PDF copy of their thesis on CD-ROM (i.e. something physical, therefore easier to see and manage).
- Library staff will be responsible for submitting e-theses to the server.

The Library is due to report back to the various committees on any issues or difficulties that are raised.

4.9 Usage

One problem with the availability of usage statistics is that when you really need them, at the start of the process when they would be really useful for advocacy, they just don’t exist! At Cranfield, usage figures quoted were from institutions in the US. At Virginia Tech, e-theses were 100x more likely to be circulated than the paper equivalent (Moxley, 2001); at West Virginia University access to theses increased a staggering 145,000% - from 813 issues in 1998/99 to 1,181,111 accesses in 2003 (Hagen, 2003).

High use of an e-thesis server may encourage academics to engage further with e-prints and the development of institutional repositories, particularly where it can be illustrated that material within a DSpace system can be retrieved not just via the ‘standard’ Google but also by 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). In Figure 2 a search of Google for information on “humic rich waters” results in a match with a thesis in the Cranfield University system as shown in Figure 2.

Figure 2. Google search showing retrieval of Cranfield thesis
By clicking on the link from within Google, a user is taken to the item record within DSpace at Cranfield as shown in Figure 3.
5. Institutional repository

Another issue that needs to be considered (probably early in the e-thesis submission change process) is whether the institution is to hold e-theses alone, or whether the same system will be developed into an institutional repository (IR) with a wider collection development policy. At Cranfield e-theses were used as the first set of material for QuePrints@Cranfield the University IR. Access to e-theses can be used as part of the advocacy for populating a more general e-prints repository. At the moment (early 2005), QuePrints@Cranfield contains 47 electronic theses obtained voluntarily from students. Once loaded, and once a persistent URL has been generated, alumni students are notified that an e-copy of their work is available. Also at this stage, an e-mail is sent to thesis supervisors so that they too are aware that the work is available and can be linked from their own Web page. As part of this e-mail, there is also continuation text describing the development of the University IR, text which encourages the supervisors to consider providing their own pre-prints for addition to the server. Whether this works or not immediately is not necessarily the issue – it is designed as a method of advocacy so that when Cranfield really does start to pursue staff more seriously for papers, they have at least been made aware of local developments – again this is the acceptance that advocacy is a gradual process.

Hopefully authors of e-theses who eventually join the academic staff will already understand the benefits provided by e-theses and will be more fully engaged with the idea of IRs.

6. Main issues for academic staff

There were a number of issues that arose in official committees that had to be addressed.

- One issue that has arisen in many institutions is a concern over quality of non-research degree theses. This issue also arose at the UTOG conference held at UCL in 1997 when an academic made the point that many Masters level theses, even if they were original, were only the beginnings of research and were probably not worth making generally available. The decision at Cranfield was to concentrate on research degree theses and to revisit Masters level theses once the initial change had been running for a period.

- One issue that could have been important was one of Intellectual Property Rights (IPR), and the issue of giving away research. One question that was asked during the process was whether the possibility of making electronic theses available would lead to an increase in the number of theses that were restricted.

- As aforementioned, the serious issue of plagiarism was also debated in a number of meetings although generally it was accepted that the availability of an electronic copy of a thesis is more likely to ensure that plagiarism is more easily detectable.
• An issue accepted right at the beginning of the process was that for the foreseeable future academics will only mark theses on paper and this puts clear limitations on the submission workflow.

• Another central issue was the definition of the ‘copy of record’. Clearly if the University Library is no longer collecting paper theses then this is not an issue, but where the University still wishes to collect, bind and shelve paper copies of theses, in addition to making electronic copies available, a decision needs to be made as to which is the ‘copy of record’.

• Related to this (and again it depends on collection policies as to whether this is a problem), how do you ensure that the paper copy sitting on the shelf is exactly the same as the electronic copy available on the server? This is quite a difficult issue if you are dealing with paper and electronic copies separately. It is yet to be seen whether this is likely to be a problem at Cranfield but there are a number of safeguards in place. First, a student will be required to sign a form to confirm that the two are the same. Second, it is hoped that at the least corrections can be checked between the two copies. The third and possibly most important safeguard is that it simply isn’t in the students’ interest to get this wrong. Clearly where this is done by mistake, there needs to be some method by which differences can be identified. Although this is not drawn to the attention of users specifically, there is currently a message on QuePrints@Cranfield saying that: “In all cases the bound paper thesis is the copy of record”.

7. Conclusions

In conclusion, the choice at Cranfield was to take a pragmatic stepped approach which it is hoped will change and develop further in the future. The important aspect was to ensure that the concept of submission of electronic theses was accepted as a general principle. The path itself was one of least academic resistance in that the submission and marking process for academic staff did not change at all and for the Registry only changed slightly at the very end of the process (apart from the need to amend one or two official letters). Theses are initially provided by students in the same way, and they are marked and corrected in the same way. The copy of record remains the paper copy. In the future it may well be the case that a model will be developed such that a thesis is submitted electronically, copies are printed for marking, corrected and a final copy is submitted again and loaded to the server by the student. It is useful right at the start to consider where you wish to get to and so when developing initial proposals Cranfield has also been developing a methodology for moving from submission of e-theses to e-submission of theses. It is likely that over the next few years submission processes will be developed and refined and so ensuring that the infrastructure can deal with these future developments is essential.

References
(All URLs in this paper were checked 8th January 2005)


Electronic Theses (2002), Metadata Core Set for ETDs. Available at: http://www2.rgu.ac.uk/library/guidelines/metadata.html


Vol. 7 No.9. Available at: http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html