An ETD Submission System for the UK

John A. MacColl
University of Edinburgh, Moray House Library
Dalhousieland, St John Street
Edinburgh EH8 8AJ
john.maccoll@ed.ac.uk

Introduction

The ETD movement has not yet taken root in the UK. Nevertheless, activity in the field has been growing steadily over the past few years. UTOG, the UK Theses Online Group, was established several years ago, and was aware of the activities of the ETD movement, but also of the distinctive characteristics of thesis publishing in the UK, which have made it - so far - unsympathetic to the creation of an ETD culture. In 2001, UTOG commissioned a report into the digitisation of theses from the SELLIC[1] Project at the University of Edinburgh. In addition to considering retrospective digitisation, the report also looked at the issues involved in the production and management of 'born digital' theses, identified reasons for the UK to become properly involved in this initiative, and recommended that UK institutions join the Networked Digital Library of Theses and Dissertations (NDLTD). These ideas were taken forward by JISC's Scholarly Communications Group, and SELLIC was encouraged to submit a project proposal to the JISC Focus on Access to Intellectual Resources (FAIR) programme at the beginning of 2002, under the name *Theses Alive!* As a result, a two-year project was funded, based at Edinburgh, and began its work in a national effort to promote ETDs in November 2002. *Theses Alive!* is one of three projects in electronic theses funded by JISC. The others are *DAEDALUS*, led by Glasgow, and *Electronic Theses*, led by the Robert Gordon University. *Theses Alive!* is working with both of these.

Several UK universities, and, indeed, the British Library, have joined the NDLTD within the last year. The NDLTD is a realistic and pragmatic movement. It does not expect institutions to change everything overnight. There is a journey to be undertaken, the three basic steps of which are:

1. Ensure that the metadata for ETDs is available online.
2. Move to a hybrid SYSTEM of print and online TDs for a period of time.
3. Arrive at the point at which the online medium is the authorised medium for the production of TDs.

This journey is likely to take several years. In the *Theses Alive!* Project, we hope to take a number of universities to the second stage of this three-stage process.

Is metadata not enough?

Do we need thousands of electronic theses and dissertations clogging the arteries of the internet (or possibly the Grid, in the near future)? In the UK at least we have grown used to engaging with this literature on the basis of its proxies - the metadata for theses, which usually includes an abstract. We have the British Thesis Service (BTS)[2], run by the British Library and supported by most UK universities - though there are a few significant omissions of research universities from its ranks. The BTS takes copies of printed theses produced by its members, and makes microfilm copies of them for loan or sale via the British Library. This saves the individual institutions the trouble of responding to requests for sale or loan copies directly, and also means that the metadata is searchable via a common union catalogue. The British Library made a commitment at the end of 2001 to join the NDLTD, and is planning to convert its microfilm operations to a digitisation-based service, which it also hopes to apply retrospectively to its huge body of thesis literature. When this is achieved - and it must surely be a mammoth task - hundreds of thousands of UK-produced ETDs will find their way onto the NDLTD.

The UK also has a commercial metadata service, the *Index to Theses*[3] published by Expert Information Ltd, whose coverage does not map exactly onto that of the BTS, although the two services are examining ways of harmonising their operations. The ProQuest service, *Digital Dissertations*, based on the University Microfilms International (UMI) *Dissertations Abstracts* database, while it has wide international coverage, does not feature many UK theses, because the UK has been so well catered for by the BTS and *Index to Theses*. However, ProQuest has recently begun to target the UK in a strong promotional effort, recognising that the switch from microform to on-

1 Science & Engineering Library, Learning & Information Centre www.sellic.ed.ac.uk
2 www.bl.uk/services/document/brittheses.html
3 www.theses.com

line availability makes international competition within this market a sensible proposition.

The model, then, has been primarily a centralised one, and one which is based upon metadata. Those wishing to search the thesis literature would most commonly use the *Index to Theses,* and then order a sale or loan copy of the thesis they wish to consult either from the British Library, or direct from the university concerned, if it is not a member of the BTS.

One reason why the UK has perhaps not moved faster into the world of ETDs is because of this centralised model. Many universities have become used to a procedure involving the despatch of their theses to the British Library, and have considered the management of their theses an issue for the British Library rather than for themselves, shelving their own local copies of theses in closed access stack, and fetching copies out when requested for use on-campus. But we are all familiar with the limitations of microform as opposed to online dissemination, and the difficulty for the British Library's service has been the size of its operation, which makes switching from a microfilm to an online process costly and time-consuming. Nevertheless, such a switch is necessary for several reasons, the main one being that metadata is not enough. The now intuitive action, for the researcher using this literature, is to proceed from the metadata to the full-text at the instant they wish to.

We might consider this a measure of its responsiveness. Until the advent of ETDs, thesis literature was one of the least responsive. Once identified in an index of theses (of which there were several in themselves), the thesis had to be requested using the local interlibrary loan service, could take some time to arrive, and could then only be used in association with the discomfort of a microform reader machine. If a bound copy was loaned, usually it would be for restricted use in the borrower's library only. As the rest of the research literature becomes more and more available through aggregated ejournal services, offering instant access to sets of journals extending back now often to their origins, searchable in a variety of ways across a large online corpus, the thesis literature, by contrast, could appear antiquated and intractable.

For that reason, and because web sites are now so prominent in the communications of researchers among themselves, thesis literature has been moving online anyway, in an unmanaged way. There is nothing to prevent a student putting a copy of their thesis onto their own web server, or a departmental server. And so there is 'bottom-up' pressure to provide ETDs, more than pressure from the organisations whose research is being published. There is also evidence that students know about and use the NDLTD already. Recent figures on the use of the ETDs in the Virginia Tech archive indicate a high number of hits from the UK[4]

The metadata alone is not enough, and is patchy in any case, with the *Index to Theses* providing the best service. The BTS cannot be searched directly, but recommends that users use the *Index to Theses* or the SIGLE[5] service, or else the British Library Public Catalogue Books File.[6] This multiplicity of possible routes to finding only the metadata is one of the reasons why the thesis literature is not more sought after by researchers. Many are not prepared to fight their way through the unconnected metadata sources only to end up with a potentially long wait for an item which is likely to arrive in a difficult format.

A strategy for the UK

In developing a strategy for the development of ETDs in the UK, however, we do wish to start with metadata. *Theses Alive!* will aim to 'genericise' the metadata creation process for all UK theses and dissertations, in order to simplify the distribution of metadata while at the same time linking it to an ETD wherever possible. The model we plan is shown below.

4 See http://scholar.lib.vt.edu/theses/data/somefacts.html#logs
5 'System of Information for Grey Literature in Europe' www.kb.nl/infolev/eagle/frames.htm
6 http://blpc.bl.uk/

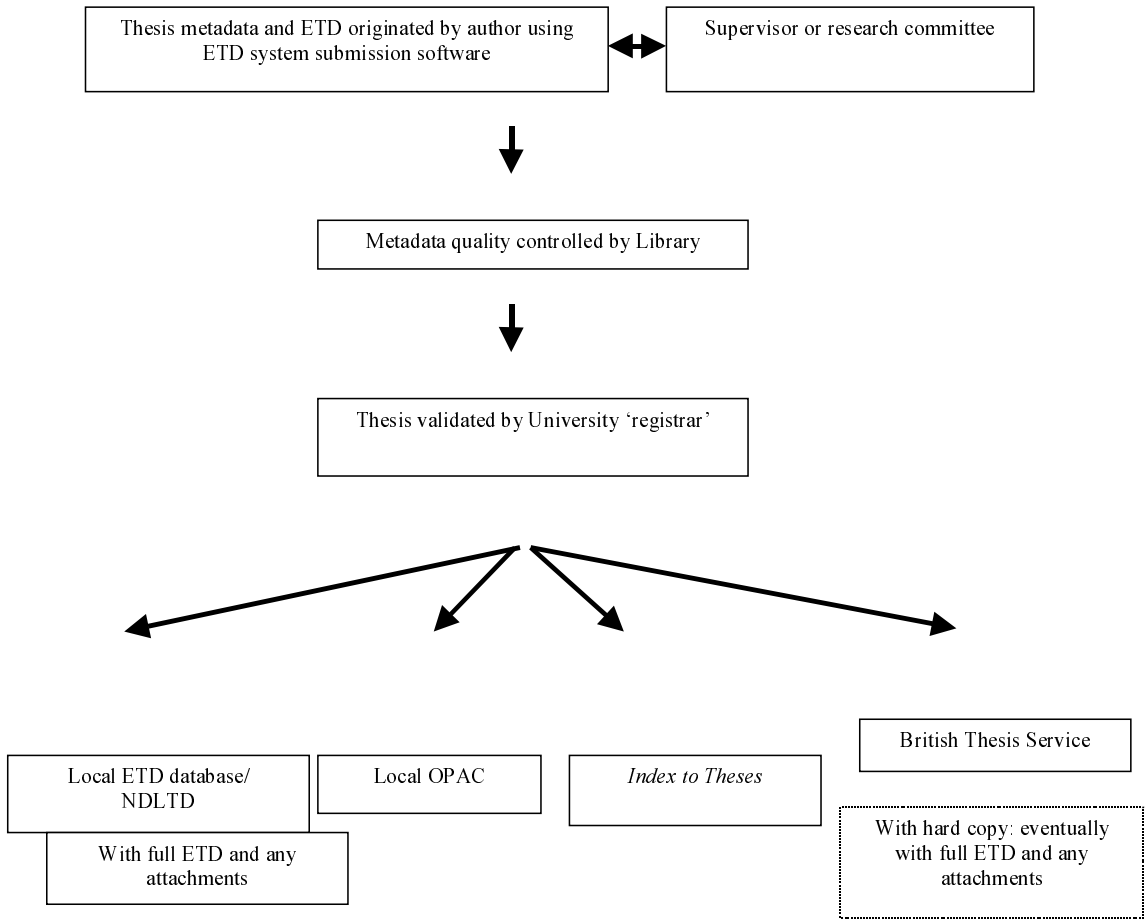

Figure 1: Theses Alive! Process Model

Using the submission software, the student creates their own metadata, which is quality-controlled by the Library. Also in the loop, inevitably, is the authority responsible for validating the approved thesis, which we have simply called the 'registrar' in the model. Interaction between student and supervisors goes on throughout the course of the degree programme. Finally, the SYSTEM outputs are the metadata, formatted as required for various agencies, and the ETD itself, which may be a PDF or other file format attachment, or may be part of the same XML file, plus any linked files.

Our expectation is that an XML schema - or perhaps a number of schemas - will be developed for UK theses and dissertations, possibly based upon schemas which already exist for use in another context. A schema will describe each thesis according to its various structural elements, and should support the export of metadata in all of the various formats required, while at the same time describing the full text of the thesis. In other words, PDF is not likely to be sufficient in the longer term. Using XML provides us with a non-proprietary format, with greater scope for database storage of deconstructed documents, greater search flexibility, and the possibility of preserving the 'raw' source of the document.

The more challenging task may be to find universities which are willing to allow ETDs to be created in their institutions, and to work with us in the *Theses Alive!* project, as pilot sites. We are not providing any funding for hardware for sites, but will support them with software installation, and will provide technical and advocacy support. We have just begun the process of soliciting interest from institutions willing to act as pilot sites, alongside the University of Edinburgh. We hope to have five or six of these, representing a mix of different university types in the UK, and providing both doctoral theses and Masters-level dissertations to the project.

Much of our work will be on the political and cultural changes needed in institutions in order to prepare them for the inevitable future context of ETDs. For some institutions, moving to an environment in which the electronic thesis or dissertation is the authoritative copy, the one which is preserved and used, may seem a huge step which is still years away. Even in the US, the numbers of institutions which have made provision for ETDs is still relatively small, though growing fast.

At this point, let us add a little more detail to the three-stage process described above. This is the strategy we wish to see adopted by individual universities in the UK,

and supported by *Theses Alive!*, at least as far as the second stage.

1. Genericise the metadata: this step is not a prerequisite, but it helps inasmuch as it implies the use of a single ETD as a structured digital object, and creates an 'ETD in waiting'. The ETD is there, but is restricted to non-public access only, by a SYSTEM administrator. This is the first goal.
2. Introduce a hybrid print and electronic TD publishing policy: very few institutions are likely to adopt ETDs outright, without running a parallel print and web service to begin with. During this stage, the print version remains the authority version for a period, but at a particular point, the roles swap over, and the ETD becomes the authority. This is the second goal.
3. It is then a fairly short step to the third stage, in which the electronic format is the required format for submission.

The challenge

Of course, there is a great deal of work to be done within institutions as they move through these stages. *Theses Alive!* will develop an ETD submission SYSTEM designed for use in the UK, but rolling it out for use by completing postgraduate students in universities will involve a lot of effort. In the US context, those universities which are already far advanced on the ETD path generally have achieved this by means of a collaboration involving four different players on campus - academic staff, administrators, library staff and IT staff. Of these four, the most important group is perhaps the administrators - those involved in the management of graduate education.

Most US universities have an organisation on campus called the 'graduate school', with a supporting infrastructure which is coherent and well-resourced. Few UK universities have 'graduate schools' as such, though they are growing in number. At Edinburgh, the University has recently created 21 Graduate Schools, with no overall point of contact. Postgraduate education in the UK is often still managed on a departmental basis. Being more fragmented, and less capable of achieving economies of scale across the postgraduate studies layer, may well make the task of engaging these administrators, the Deans of Graduate Schools, considerably more difficult in the UK. But without the support of senior university managers, the ability of an institution to move in the direction of requiring ETDs, or even encouraging a dual submission SYSTEM, is likely to be very much compromised. Certainly, the library cannot do it alone - nor the computer services department. Academics can lobby successfully, if they become convinced of the value of the initiative, but they might be content with achieving ETDs in their own department only - a partial solution which will not satisfy the library's desire for uniform access.

Once we have identified the contacts, the next task is to work with supervisors and theses authors themselves, who will use the *Theses Alive!* SYSTEM for the final stages of their thesis preparation and submission. We have a target of creating 500 ETDs across the project by the end of two years. We expect that in this, the first year, the pilot institutions will provide 20-30 submitting theses, and next year, with more time to prepare, and a more mature software product, we would hope for 70-80 per pilot partner. It is likely that, in addition to theses near to submission, which will provide the most effective test of the software, in each institution there will be a corpus of already completed ETDs, located in departmental repositories, and we will be happy to 'top up' our numbers with these, provided that we can satisfy the departments concerned about the necessary permissions. These will have a value in growing the overall corpus, and thereby assessing its value through the demand for access which will be monitored.

University staff will require training in order that they can offer training programmes to the students concerned. It is likely that these staff will include library staff, although other staff in a training role, from IT services or even academic staff training new postgraduates in research skills, may be the preferred training providers. Virginia Tech uses graduate students themselves, which clearly also has a number of advantages, though it would be a less common model in the UK. A major component of the training programme will be attention to Intellectual Property Rights (IPR). Students will require to be educated not only in their own rights in their theses or dissertations, but also of the need to clear rights for linked or embedded content. *Theses Alive!* will provide central support for the training programmes in pilot sites.

ETDs in institutional scholarly publishing

To date, the Project has evaluated three software packages for use in ETD submission management and storage: Virginia Tech's ETD-DB, eprints.org from the University of Southampton, and MIT's DSpace. After careful evaluation, we have decided that DSpace offers the richest functionality and the most satisfactory repository management and object preservation. Its latest release, which is overdue at the time of writing, will apparently provide further specific functionality in support of ETDs. Our intention is to develop an interface adapted to the needs of the UK university community, with support for the metadata output formats required in that context. To that end, we hope to work closely with DSpace colleagues at the University of Cambridge and in the DSpace Federation. We expect that this approach will fit well with the DSpace philosophy, which was described

thus by MacKenzie Smith and colleagues in a recent article:

'With the help of developers at other institutions that adopt DSpace under its open source license, we will work to add features and improve the different functions of the SYSTEM as we learn what users actually want, and how best to support such complex requirements as digital preservation and digital rights management.'[7]

We anticipate considerable benefits to our project from joining the DSpace community, which offers cooperative services of various kinds, as described in the DSpace@Cambridge Project Proposal:

'This includes functions such as ''virtual'' collections or publications distributed across several institutions, cross-institutional searching, and distributed services that federation members can take advantage of (e.g. data clean-up or enhancement, format migrations for preservation, and so on).'[8]

The cooperative approach will lend an energy to DSpace which should allow for rapid development, provided that SYSTEM take-up spreads quickly.

The SYSTEMs which handle eprints and ETDs, indeed, share a considerable amount of functionality. Much of the workflow is similar, though the sequencing is not quite the same. Both follow the same basic sequence of *preparation - submission - review - finished publication* - although ETDs obviously have much more in-progress review. A key difference will be in their readiness to join the corpus. A researcher using a search provider to query the research corpus is only likely to be interested in completed ETDs (and extremely unlikely to be able to find anything else), whereas they may be happy to search for eprints which include those submitted for publication but not yet published. An interesting area of overlap occurs in the case of research programmes which require or expect their students to publish an article or a number of articles in a peer-reviewed journal as part of their degree. It is not difficult to imagine a scholarly publishing SYSTEM which allowed students working in an ETD module to tap in to the functionality used by academics in eprint self-archiving and journal submission.

DSpace would appear to be particularly well-suited to material which has to be restricted, and ETDs will be frequently subject to restrictions and embargoes applied by their authors. Smith describes the capacity of the SYSTEM in this respect:

'For material that is restricted to local access, the item metadata is exposed to OAI harvesters but the SYSTEM will enforce the restriction when a user requests the associated bitstream(s).'[9]

Naturally we hope that as many ETDs as possible will be unrestricted in their full-text, but the reassurance that access can be embargoed for a period of time, or restricted by IP, will go a long way to persuading authors and supervisors to participate in our pilot project.

The FAIR programme has astutely recognised the common ground between ETD promotion and the promotion of eprints archiving projects, and this provides the opportunity to rationalise the promotion and advocacy work required in institutions which are seeking to develop a range of repositories at the same time. This is true of Edinburgh, which, like Glasgow, is participating both in an ETD project and an eprints project (SHERPA[10]) at the same time. Where JISC's vision has perhaps fallen short of that of the DSpace Federation, is in taking a view of institutional repositories as being primarily resources for research. DSpace is also likely to be used as the repository manager behind MIT's initiatives to husband and manage the digital objects used in learning, through its Open Knowledge Initiative (OKI)[11] and Open CourseWare (OCW)[12] projects. Inevitably, this is a requirement of the same universities which are adopting DSpace as a solution for their research publication archive needs. Its scope, therefore, built 'breadth-first', is sufficient to permit a transformation in the way universities manage their own generated knowledge, assuming that a transformation can of course occur within the ways in which academics work. This is likely to be the more difficult transformation to effect.

Conclusion

Theses Alive! is a project which many consider well overdue in the UK. The hard work of the UK Theses Online Group over the past few years now has an opportunity of bearing fruit. That work is now underway, and we hope that our project will assist the process of liberating the theses literature in the UK, making it available to the world, improving the research experience of students, and increasing access to knowledge for the benefit of research generally.

May 2003

7 Smith, M; Barton, M; Bass, M et al 'DSpace: an open source dynamic digital repository', D-Lib Magazine, 9 (1) January 2003 www.dlib.org/dlib/january03/smith/01smith.html

8 'DSpace@Cambridge: a project to extend DSpace into Cambridge University and the United Kingdom', July 2002 www.lib.cam.ac.uk/dspace/doc/proposal.htm

9 ibid.

10 See www.sherpa.ac.uk

11 See web.mit.edu/oki

12 See www.ocw.mit.edu