

**ETD 2003: Next Steps – Electronic Theses  
and Dissertations Worldwide  
Humboldt University, Berlin, Germany, 21<sup>st</sup>-23<sup>rd</sup> May 2003  
Conference Report**

**Standards Committee Meeting**

This was the meeting of the NDLTD standards committee, which was also made open to conference delegates for their input. A number of issues were addressed:

*Subject Areas/Sets for OAI*

There are issues with both cross-language and cross-culture compatibility when deciding on which sets to place items. It is suggested that we use some of the more unusual Dublin Core elements that are provided to make this classification easier. In general, users cannot be expected to correctly implement a ‘controlled vocabulary’, so there are issues for librarians and for automatic set classification (NDLTD reports a 60-80% success in this area). In addition, it is known that transforming between ETD-MS and MARC 21 has various shortcomings due to some incompatibility between certain fields.

*PDF*

NDLTD are considering what steps will be taken to ensure that PDFs are still readable in years to come.

*Identifiers*

The pros and cons of technologies such as DOI and POI (using an OAI identifier to provide access), and OpenURL are briefly discussed. A number of suggestions and warnings were provided, including:

- use Crossref to handle your DOI, and allow them to do the work and distribute the cost;
- it is not possible to use completely separate identifiers for print and ETD – there will always be some crossover.

*Digital Rights Management*

This must be decided at the institutional level, but the general feeling is that if you don’t want people to read it, don’t put it in an open archive.

**Opening Session**

Following a welcome from the conference organisers, Axel Plathe from UNESCO introduced their policies on the subject: free flow of ideas; maintain, increase and diffuse knowledge, as well as noting that public domain information is not sufficiently well known, and that there are growing restrictions on availability and use. (See 2002 UNESCO Guide to ETDs; <http://etdguide.org/etdguide.pdf>). UNESCO’s sponsorship was strongly felt at the conference, which included many delegates from poorer countries, particularly in Africa, who were subsidised by UNESCO. The appetite of these countries for digital solutions to their resourcing problems was very plain, and reminded those of us from more affluent countries of why this effort is so important.

Joan Lippincott subsequently spoke about who needs to be involved in an ETD project: academic administrators, faculty, students, IT staff, and librarians. She also spoke about how

to cooperate not only with other universities, but with other departments within your institution.

A representative from AJLSM then demonstrated a new system 'Cyberdoc' or 'Cybertheses' built (in French) to handle thesis submission and archive. They plan to translate this in the near future into a number of different languages.

### **Technology and Workflow**

Stefan Anderson presented the ETD workflow for Uppsala University in Sweden. Among other things they require an electronic copy submitted three weeks before the defence date. Each thesis and summary must be posted online, and to ensure integrity they have a 'controlled form' for certain metadata elements.

Stefan Kramer from the Fielding Graduate Institute presented their online archive ([www.lib.umi.com/cr/fielding/main/library/dissertations](http://www.lib.umi.com/cr/fielding/main/library/dissertations)), and discussed the issues with having a distributed student base with varying amounts of bandwidth.

Rolf Rasche from ImageWare presented their software model for dealing with digital lending of print theses, which takes into account the copyright issues involved by bar-coding each lending and limiting access. It is essentially an extension of the ILL procedure.

Alice Keller from ETH Zurich Library spoke about integrating ETD procedures into your staff workflow. As a general rule, authors prepared to go online are prepared to provide PDFs, but it is necessary to ensure that the PDF is identical in content to the print version (which remains authoritative). In general it takes around six months after submission for an ETD to make it into the live archive, and librarians are principally responsible for metadata and archive.

### **National ETD Initiatives**

Julia Puig from the Miguel de Cervantes Digital Library (Alicante) introduced their online repository (<http://www.cervantesvirtual.com/>), which offers free access to theses from 33 institutions, mostly from Spain. Theses from other institutions are accepted where the subject matter is relevant (e.g. Hispanic culture).

Chris Rensleigh presented the various issues and processes involved in setting up an institutional ETD repository at the Rand Afrikaans University (South Africa). An initial pilot project over two years with a limited budget of R15000 (~\$2000) has produced a repository with currently 400 submitted theses.

Olga Lavrenova presented the Russian State Library's (<http://www.rsl.ru>) efforts to create a national ETD initiative: Open Russian Electronic Library (OREL), which currently contains 200 ETDs (mostly retrospectively scanned). The finished product has been running for eight months and has collected approximately 30 'born digital' theses.

### **Panel Session 2**

Representatives from Africa presented the DATAD project (<http://www.aau.org/datad>).

Susan Copeland from RGU introduced e-theses developments in the UK, with special reference to their E-Theses Project, the Glasgow DAEDELUS Project and the Edinburgh Theses Alive! Project. The UTOG group and JISC were also introduced in the role of driving these developments in the UK.

In a bold 'call to arms' to the conference, Gail McMillan from Virginia Tech focused on

preservation. She began by quoting from Shelley's *Ozymandias* to dramatic effect ('Nothing beside remains. Round the decay/Of that colossal wreck, boundless and bare/The lone and level sands stretch far away'). She spoke about the NDLTD's need to consider digital preservation more, stating that we not only need accessible digital libraries, but that we need to preserve them. Archiving should be a core activity of the NDLTD with collaborative persistent mirroring, with effective partnership between creators of documents and the stewards of those documents. We need to take the 'L' in 'NDLTD' (for 'Library') much more seriously, and aim to provide a copy of each member institution's ETD collection through collaborative persistent mirroring.

Ana Pavani then spoke about how to train your staff and students to deal with ETDs, with special reference to the people of Brazil with their vast diversity of educational status.

### **Special ETD Projects**

A group from MIT Libraries' DSpace team (<http://thesis.mit.edu/>) presented their proposed workflow for ETDs using their DSpace system. It begins with the student providing the file and the metadata and then the Faculty Advisor checking for correctness and approving submission before the library checks the patent hold database, edits the metadata and then either approves or holds the thesis. MIT is migrating its existing collection of ETDs (~250 born digital, and 8,000 scanned) into DSpace from the previous platform based upon the Dienst protocol. The team also introduced some additional functionality specific to ETDs that they are considering implementing, such as a virtual workspace, additional workflow configuration, support for the METS standard and integration with the student database. We also learned some information about the developing DSpace Federation, which currently consists of MIT, Cornell, Columbia, Ohio State, Rochester, Toronto, Washington and Cambridge.

Charles J. Greenberg presented the Yale Medicine Thesis Digital Library Project (<http://ymtdl.med.yale.edu/>) and included discussion on policy issues, project implementation and use from the Yale perspective. Particularly interesting was how this small graduate school project has raised the awareness of ETDs in the broader academic community at Yale. He described the difficulties they had faced in overcoming faculty and researcher fears about prior publication (which he called 'Ingelfinger thinking', after the prescriptive policy of the *New England Journal of Medicine* when under the editorship of Franz Ingelfinger).

Nishtha Anilkumar from the Physical Research Laboratory (India) detailed the major initiatives that the PRL Library are working towards, including creating and maintaining a Digital Library and E-Journal repository.

### **Technology and Workflow**

Uwe Miller from Humboldt University showed us that in a workflow the process for a document is non-linear, and circularities may evolve in the procedure. He also added that there are often time limits for step completion, and that there are various methods for allocating tasks (push/pull). He saw roles for the author, the referee, librarians, printing services and the document server manager in producing a sound workflow. He also advised on a pre-workflow set of procedures to prepare a document for smooth transition through the main workflow.

Horst Kastner introduced the idea of the Organisation Advancement of Structured Information Standards (OASIS), then Jean-Claude Guédon of Montreal further demonstrated the Cybertheses (Cyberdoc) system (<http://www.unige.ch/cyberdocuments/>). Mainly French-speaking, this system is becoming increasingly multilingual, and is particularly focused on providing e-resources to libraries in developing countries. Professor Guédon described it as

being at the ‘confluence’ of open source and open access currents.

### **Technical Issues**

Kathrin Schroeder presented various methods of using ‘persisting identifiers’, such as DOI, Handle Servers, PURL, and URNs. She made specific reference to URNs as these are being strongly considered/used by the German National Library.

William Nixon from Glasgow University presented their EPrints system (<http://eprints.gla.ac.uk/>) within the scope of the DAEDELUS project.

Matthias Schulz then showed how the open source nature of OpenOffice.org allows institutions to provide custom menus to deal with various ETD templates and options, allowing them to save documents in their custom XDiML format. They acknowledged that this approach suffers from some relatively serious problems.

### **ETD training methods**

Bettina Berendt described a large-scale empirical study of the marketing strategies of the document and publication server of Humboldt University (<http://www.edoc.hu-berlin.de>). Using an e-mail/web based questionnaire for wider dissemination (1180 users were targeted – but only 101 finished the questionnaire), various issues were investigated, including the experiences, assessments, plans and wishes of the two main stakeholders in the project- the submitting authors and the online readers. A general finding was the lack of publicity for the eDoc service. She gave some interesting information on the theses publishing culture in Germany. All theses have to be published in book form before a degree can be awarded. This can be costly for authors. All are also required to be deposited in the national library.

Guy Teasdale introduced the training methods that Laval University (Canada) implemented for its institutional ETD project (<http://www.theses.ulaval.ca>), which was officially launched in 2002. Since then, in-house training for over 200 doctoral students has been given via workshops and online training. The training was necessary because the system requires a strict system for document submission.

Gerald Eichler talked about the developmental concepts that T-Systems Nova are interested in. The presentation was not very relevant to ETD issues, though it did present an interesting argument for the use of active learning systems, and their relevance to industry.

In the last presentation of the session, John MacColl presented the Theses Alive! project (<http://www.thesesalive.ac.uk>) to the conference and also gave background information about the ETD movement (or lack of!) in the UK.

### **Panel Session 3**

The floor was opened up for members of the conference to put their questions to the panel of ETD experts. The primary topics covered were digital rights management and preservation, along with some calls for the NDLTD to provide a more supportive role to ETD projects.

### **Panel Session 4**

Peter Schirmbacher from Humboldt University introduced some of their services (<http://edoc.hu-berlin.de/> and <http://www.dini.de/>).

Shalini Urs then presented Vidyanidhi (which means ‘treasure-house of knowledge and scholarship’) - the Indian Digital Library of Electronic Theses, which has ~40,000 theses already in its database. She showed how it represents an extreme case of cross-language

compatibility problems because India has 18 official languages and >400 other live languages (many without a script). A large number of these are represented in the database. She reported that UNICODE is in fact 99% satisfactory for dealing with this.

Victor Sibirsky from Moldova introduced their ETD project (<http://www.mir.acad.md/>).

Finally, Andrew Wells from the University of New South Wales introduced their ETD project (<http://adt.caul.edu.au/>) where they use the VT-ETD software. Their database now has ~1,200 ETDs, and was based upon a business plan drawn up by the Council of Australian University Libraries (CAUL). All members contributed funding to provide the initiative with a stable foundation, and they have a commercial partner (ProQuest) also involved.

## **LaTeX and ETDs**

Gunter Torner explained what the pros and cons of converting your LaTeX files into PDF are: it is preferred by libraries, but you lose some semantic information and structure, and there are version problems whereby Acrobat will interchange symbols. Preserving the original mathematical source is not quite so important as preserving the original scientific information.

A representative from Duisburg-Essen went on to further explain that it is better to create your PDFs from LaTeX than it is from MS Word due to the way that layout changes are done in the transformation.

Hans Hagen introduced us to XML documents with specific reference to MathML. Although this language is not yet up to standard, it is still very much under development (<http://www.pragma-ade.com/> and <http://www.pragma-pod.com/>).

Thomas Fischer then urged libraries to keep their TeX-based documents as their base document rather than importing to PDF due to various preservation issues.

## **Library Issues and OAI activities**

Eva Muller introduced the DiVA portal (<http://publications.uu.se/portal/>) - a collaboration between five Swedish universities to provide an OAI compliant linked repository of ETDs. The main outcome was that a common agreement on interpretation of metadata standards and vocabularies are necessary for meaningful resource discovery.

Kimberly Douglas addressed the motivation and the impact of withholding theses from the voluntary Caltech ETD repository during the first year of its official launch (2001/2002). The repository attracted an approximate 20% submission rate from graduates. One-fifth of these submissions asked for a restricted access policy. The study found that advisor input was a significant factor in restriction.

Tim Brace, who represented the University of Texas at Austin (UTA), gave the final talk of the session. The presentation highlighted some of the problems that the university has faced since moving from conventional to electronic submission of theses in the summer of 2001. The main problems encountered were copyright issues, such as the use of embedded proprietary fonts, images within theses and conventional publishers' attitudes to prior publication. The 'take home' message from UTA is that they don't consider an ETD as conventional publishing; therefore the issues of 'embedded copyright' are bypassed if 'fair use' is implemented (although this does not apply to embedded proprietary fonts). They have good authority for this approach, since the libraries in the U Texas system have their own copyright attorney (Georgie Harper, who co-presented the session). They have reached this position after much dialogue with publishers (and have a letter from Elsevier to prove it). Theses and dissertations are considered more akin to teaching materials than to commercial

publications. Permissions are sought, but if they are difficult or expensive to acquire, fair use protects the web publication of embedded copyright material.

### **Closing Session**

Ed Fox of Virginia Tech, in many ways the father of the ETD movement, and Chair of the NDLTD, described some of the more recent projects being developed at Virginia Tech. Project MARIAN is developing a 'crawl and scrape' tool for identifying ETDs on web sites and creating a union catalogue of them. Work is also going on into assessing the value of the LOCKSS approach to digital preservation ('Lots of Copies Keeps Stuff Safe') within the ETD world.

### **Workshop: DTDs for ETDs**

This workshop briefly discussed the basics of DTDs, but the primary topic was in transformation and preservation. The talk rapidly moved on, and remained on, using templates to generate XML files which are compliant with your DTD. Much of this involved using OpenOffice.org to provide a customised set of saving templates.

*Theo Andrew, Richard Jones and John MacColl, **Theses Alive! Project***  
11 June 2003