

Smoothing the Transition to Mandatory Electronic Theses

Introduction

Electronic theses originated with the efforts of Ed Fox at Virginia Tech a decade ago. Today there is an international groundswell of participation from institutions of higher learning, including support from UNESCO. Electronic theses, in a globally networked environment, have many advantages over their print-on-paper origins:

- Ease of access for remote users.
- Discoverability from free search services, both general (e.g. Google) and targeted (e.g. NDLTD Union Catalog).
- Capacity to integrate rich multimedia.

Caltech committed to participation in the Networked Digital Library of Theses & Dissertations (NLTLD) in 1999. Following Commencement in June 2001, recent graduates were invited to contribute their theses to the fledgling CaltechETD database. With roughly two dozen electronic theses deposited voluntarily, the Graduate Dean, the Graduate Studies Committee (GSC) and the Caltech Library System (CLS) decided in April 2002 to move forward with regulations requiring graduate students to submit an electronic thesis as a condition of receiving their degree. For a detailed timeline, see Appendix.

CLS has chosen the ETD database (ETD-db) software, developed at Virginia Tech as a joint project between the Graduate School at Virginia Tech, the Digital Library and Archives (a division of the University Libraries), and the NLTLD. Access controls for protecting intellectual property rights played a key role in the selection of the ETD-db software. Massaging the user interface of the software, which is freely available, was just a portion of the groundwork required for launching the initiative.

Workflows in Technical Services and the Office of the Dean of Graduate Studies were reimagined. It was necessary to edit some existing documents and to create some forms from whole cloth to achieve a reasonable procedure. Some entirely new tasks and responsibilities for the subject liaison librarians arose as a result of the implementation, in addition to the new software maintenance and customization efforts for Library Information Technology staff.

Approached with good humor and good will, the administrative side of the house is now in order. A series of seminars offered by the campus' Digital Media Center have included library participation, focusing on the new electronic thesis requirements. Library seminars have been developed. These are offered daily for one week every three months, initially in late June 2002, to explain the processes and requirements, and to highlight appropriate resources available throughout campus, which will aid in the transition faced by the graduate students. Easing the burden on the graduate students, we've developed a

website with advice, guidance, requirements and templates. Templates are available in Microsoft Word and TeX, which implement all of the Graduate Office's structural requirements for theses.

CaltechETD is compliant with the Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH). OAI Services built using the OAI-PMH [e.g.; DP9 (<http://arc.cs.odu.edu:8080/dp9/index.jsp>) and the NDLTD Union Catalog Project (<http://oai.dlib.vt.edu/~etdunion/cgi-bin/index.pl>)] increase the discoverability of Caltech theses via standard web search engines (e.g.; Google) and allows for targeted federated searching in conjunction with other OAI repositories.

The evolution of electronic thesis project at Caltech has been rapid. We have been exploring issues of promotion to and support of the graduate students. We've begun accepting deposits of source files, in addition to the required presentation format, Adobe PDF. Thesis source files are kept in a separate archive to help future format conversion and migration efforts in a modest attempt to address the problem of very long term document preservation caused by format obsolescence.

Technical Decisions

We faced a number of technical choices and challenges in setting up our thesis repository. These included:

- Choice of format for included documents.
- Choice of OAI-compliant software for repository and user interface.
- The need to provide reliable, persistent URLs for use in citing included documents.
- Archiving issues, based on our commitment to providing access to electronic theses over the long term.

Document Formats

In common with many other institutions, we require that students submit their theses in Adobe Portable Document Format (PDF). This decision was based on the ubiquity of Adobe Acrobat Reader software, making it likely that users worldwide will be able to download and read the theses. It was also based on the fact that PDF, unlike HTML, is an excellent format for producing printed copies—something we expect users will continue to want to do. There are good tools available for producing PDF from the document formats our authors use to create their theses. For authors who use Microsoft Word, the Library acquired a campus site license for the Adobe Acrobat package, which is the primary tool for producing PDF from Word in a Windows or Macintosh environment. Tools for the Unix TeX and LaTeX environment, such as, *pdf_latex*, are available free as part of various TeX distributions.

For students who enhance their theses with multimedia files, we chose to accept the same list of file formats that is currently accepted by University Microfilms

(http://www.lib.umi.com/dissertations/about_etds). Students who wish to use other formats are encouraged to discuss their needs with us.

Repository Software

We had a number of requirements for the software to host our thesis database. The solution we selected needed to be:

- Free or cheap, and preferably open source
- Well supported and stable
- Easy to use, for both authors and searchers
- At least somewhat customizable, so we could add a Caltech “look”
- Allow limiting access to some theses, based on authors’ wishes
- Compliant with the Open Archive Initiative Protocol for Metadata Harvesting (<http://www.openarchives.org/>), to allow our repository to be part of various federated search services.

Based on these requirements, there appeared to be three candidate choices: the ETD-db software (<http://scholar.lib.vt.edu/ETD-db/developer>) developed at Virginia Tech as part of the NDLTD effort; the Eprints software (<http://www.eprints.org/>) available from Southampton University in the UK; and some kind of home-developed solution. We did not consider any commercial solutions.

The “home-grown” option was rejected early on. Although it would have allowed us to create a solution that conformed exactly to our needs, it was far too slow and labor-intensive. We needed to leverage the work of others to get our thesis database up and running in a reasonable amount of time.

The Eprints software from Southampton was an attractive option. We were already using it for a number of archives in our digital collections program (CODA; see <http://coda.caltech.edu>). The Eprints database is highly customizable and we had considerable experience with it. Eprints is OAI-compliant and open source and has a large and active group of users as well as a support commitment from its developers. However, it is oriented toward “free and open access” to materials and therefore lacks a key feature which we needed: the ability to restrict access of some theses to on-campus viewing only, or even to withhold a thesis completely for a short period for patent or proprietary purposes. Although Eprints is able to restrict access at the file level, authorization is based on an authenticated Eprints user account, whereas the ETD-db software authorizes access based on a valid host IP number. This allows us to restrict access to our campus domain, if necessary, without requiring user authentication. Our discussions with other campuses that had implemented ETDs convinced us that this was an essential feature.

In the end, we settled on the ETD-db software from Virginia Tech. This software is freely available and though the user group is not large, it is growing. In retrospect, we are happy with this choice. The thesis submission user interface for authors is not fancy but it is functional; we have had essentially no complaints about difficulties with

submission. Authors are able to specify access restrictions for all or part of their theses, if necessary. For library users, the software provides both a browse capability and a Boolean search capability. We have been able to use customization features to change the “look” of the pages to conform to our local library web style, although we have not been able to modify the functionality of the software to meet various suggestions for improvements that we have received.

OAI-PMH compliance was in development when we implemented the ETD-db and became available soon thereafter. We installed the OAI support and our theses are now included in several OAI-based federated searching services, including the Electronic Thesis/Dissertation OAI Union Catalog (<http://rocky.dlib.vt.edu/~etdunion/cgi-bin/index.pl>) and Old Dominion University’s ARC service (<http://arc.cs.odu.edu>). On the down side, migration to upgraded versions of the ETD-db software has been anything but simple.

Persistent URLs

The problems with short-lived URLs on the web are well known. We needed a way to provide a URL for each thesis that would be unchanging for the life of the document, so that it could be used reliably in links and citations. The native URLs produced by the ETD-db software (or any archiving software for that matter) are inherently unstable since they are constructed from the archive server’s hostname and file path names and identifiers peculiar to the software, e.g. (<http://etd.caltech.edu/etd/available/etd-07132001-180811/>). If we decided in future to move our repository to different software or a different server, these URLs will break.

To solve this problem, we have implemented a Persistent URL Resource Resolver (PURR). This allows us to assign each thesis a persistent URL constructed from the resolver server’s hostname and identifiers chosen by us (not the software). These elements will not change should we move the thesis archive to a different server or software platform. A persistent URL for a thesis record looks like this: (<http://resolver.caltech.edu/CaltechETD:etd-07132001-180811>). The persistent URL is featured prominently on each thesis’ home page, and is the recommended URL to use when citing the thesis.

When a user clicks on one of these persistent URLs, the PURR resolver consults its database and maps from the persistent URL to the current actual URL. Then it automatically redirects the user to the current location of the document. The PURR database is dynamically updated nightly, to keep it in sync with the actual URLs of the documents it tracks. More information about the PURR is available in the article by Ed Sponsler at (<http://resolver.library.caltech.edu/caltechLIB:2001.003>.)

Archival Issues

The issues involved in long-term archiving of electronic documents are thorny and the subject of much debate. While we as a library have a commitment to the long-term

maintenance of the electronic theses in our collection, we could not afford to wait until it becomes clear exactly how this is to be done. We know that the documents in our collection will need to evolve with technology. The required PDF format for theses is probably not an ideal long-term format; the specification has already undergone several revisions, and it is under the control of a corporation rather than an international standards body. However, we believe that the ubiquity of PDF in the current web environment guarantees that migration tools will be available when required.

One way we are addressing archival issues is by encouraging authors to submit their “source” documents, whether in Word, LaTeX, or FrameMaker, along with the PDF version. For older theses, which have been scanned and added to the ETD database, we store the TIFF page images. These source format documents are not part of the publicly available ETD-db, but are stored in a separate repository managed by the Eprints software. We expect that having the original source documents available may be of some help in the inevitable conversions that will be required in the future.

We are also taking some preliminary steps toward archiving theses in XML, which holds much greater promise as a long-term archiving format. There are currently nearly insurmountable problems in getting authors to produce their theses in XML, so our efforts are focusing on conversion. We are beginning a pilot project involving theses authored in LaTeX, which is a fairly structured markup language and lends itself well to conversion to XML. A small number of theses will be converted from LaTeX source to XML and archived in that format, along with PDF. We hope to use the knowledge gained from these experiments to expand our conversion program, eventually encompassing theses authored in Microsoft Word as well. A major point of possible contention was avoided by maintaining the print thesis requirement, acknowledging the long-term stability of the print record. Thus print is the copy of record, at least for now. This decision was facilitated by Caltech’s relatively small size and necessitated by the decision to move forward without an iron-clad format migration plan.

Training and Guiding

The Website – One Stop Service

The Caltech Electronic Theses website (<http://library.caltech.edu/etd/>) (Figure 1) was developed to meet a variety of needs, both informational and operational. From a single webpage, information is provided about the origin of the electronic thesis movement, how Caltech became involved in it, advice on formatting issues, the login for submission, and links for searching both the electronic thesis database and the online catalog. The Caltech Ph. D. Thesis Regulations (http://library.caltech.edu/etd/thesis_regulations.pdf) were updated by Daniel Taylor (CLS) to reflect the mandatory nature of the electronic thesis requirement and the availability of more current style guide editions, and then printed as PDF for easy web distribution.

The website was the centerpiece of the library’s presentation to the Graduate Studies Committee, a policy-making body. The Graduate Studies Committee’s approval was

necessary to implement the mandatory requirement. A brief demonstration of the website satisfied the faculty, the Graduate Dean, and the graduate student representative that the majority of probable hurdles had been identified and measures taken to mitigate their impact on students. One particularly thorny challenge has been sidestepped for the near term by continuing to designate the printed copy of the thesis as the version of record. As additional stumbling points are identified, the ETD Team works through the challenge with the graduate student, and then documents any generalizable lessons in the ETD FAQ (http://library.caltech.edu/etd/ETD_FAQ.htm). A commitment to this approach assuaged the lingering concerns of the graduate student representative. The only major change to the website since April 2002 was the addition of the Caltech ETD Statistics link in February 2003.

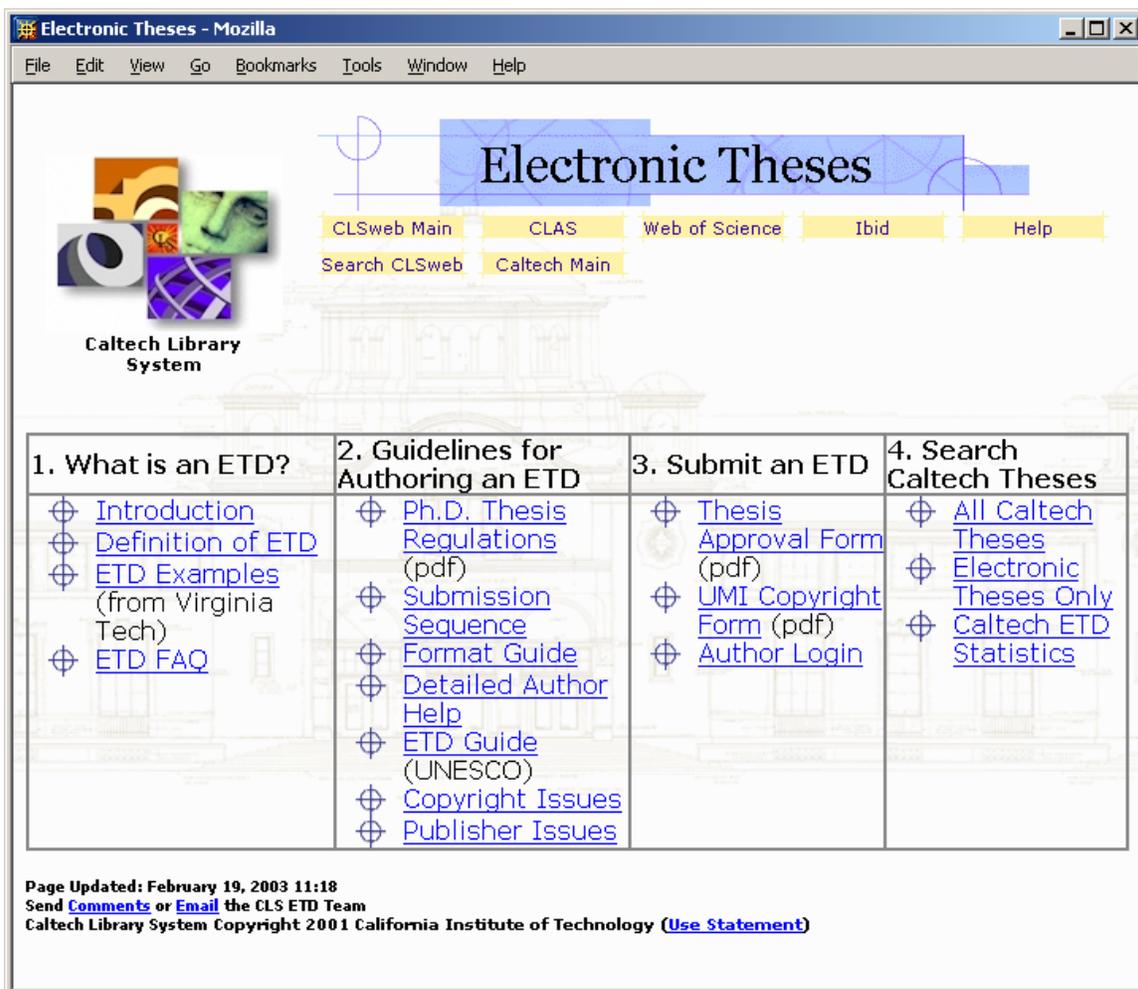


Figure 1: Caltech’s Electronic Theses website

Workshops

The library has hosted a series of classes (see Appendix. Caltech’s Chronology) to orient students to the policies of the Graduate Office and to the Caltech Electronic Theses

website. The classes also provide an opportunity to gather data on computer platforms and writing applications used in different disciplines and research groups. The class is designed to be very practical: what's needed and how to accomplish it. As such, significant time is devoted to creating PDFs from Word and TeX, the integration of Adobe Acrobat into the Microsoft Office suite during installation, font choices, and file sizes and naming strategies. Several strategies are demonstrated for generating PDF from source files and for choosing compression settings within the Acrobat Distiller software.

The classes feature a step-through of the ETD-db user registration, metadata generation, and file uploading mechanisms. A particular issue is that the abstract has to be in exceptionally plain vanilla ASCII. A standard portion of the metadata demonstration is to copy and paste an abstract from a thesis. The abstract chosen for the purpose uses a font with special glyphs to kern letters, employs a couple of Greek letters as scientific symbols, and has an unlauded o (ö), none of which reproduce accurately within the ETD-db metadata interface. Of course, all special characters display and print properly in the actual thesis PDF file.

Additional examples are displayed and discussed, both salutary and cautionary:

- Fonts that do not display well online.
- Multicolor detailed maps – easily displayed; plottable at full size – printable at page size.
- Huge single file theses with download times exceeding an hour at 56Kbs.
- Multiple versions – high resolution, lower resolution, color vs. B&W graphics.
- Named files, where appendices are listed first, because files display alphabetically.
- Numbered file names, where the author controls the sort order of file presentation.

Templates

Templates have been developed for implementing the thesis regulations in commonly used software formats. Kathy Johnson (CLS) created a Microsoft Word thesis template. Several candidate TeX style files were identified, posted on file servers. Librarians and a campus TeX expert evaluated the TeX style files. As luck would have it, Daniel Zimmerman's implementation was exemplary. Zimmerman has continued as a lecturer in the Computer Science department and has agreed to lend his expertise in evaluating proposed enhancements to the TeX style file. Adobe FrameMaker is a distant third among campus adherents, but George Panotopoulos was sufficiently excited by the concept that he created appendices in his thesis to fully model all of the sections defined in the requirements. Panotopoulos generated a FrameMaker template from his thesis and deposited it for distribution.

Sharing the Load

The library simply can not do everything. Fortunately, the library does not exist in a vacuum on campus. Caltech's Digital Media Center (DMC)

(<http://www.its.caltech.edu/its/digitalmedia/>) provides expertise and instruction in a variety of software products and media technologies. The DMC developed a workshop on PDF for Long Documents and Dissertations in the spring of 2002. Librarians were invited to participate in the workshop to emphasize the applicability of the material to the electronic thesis preparation process. Documentation from the DMC workshop is linked from the Caltech Electronic Theses website (<http://morel.caltech.edu/classes/pdfs/0612PDF%20for%20Long%20Documents.pdf>).

Copyright Permission Assistance

Electronic theses have a great deal more visibility than printed theses. As a result of the increased visibility, the long-standing requirement for authors to secure permission from third-party copyright holders has taken on even greater significance. Subject librarians have taken on the challenge of identifying the correct contact information at the various publishers. If a student has a copyright permissions question, they are encouraged to email the ETD Team or their subject librarian. Students have been uniformly relieved to know that they will have assistance in getting their request directly into the hands of the right person. Librarians benefit from developing a working knowledge of the rights contacts at publishers in specific subject areas making subsequent requests for assistance much easier to fulfill.

Impact/Effect of the Initiative

The graduate student community has taken ownership of the electronic thesis concept. Dean Kiewiet announced the change in policy by email to the graduate student population in May 2002. He wrote a brief article for the Graduate Student Council's June 2002 newsletter (http://www.its.caltech.edu/~gsc/newsletter/jun02_main.html - ETD) to further publicize the new policy and included a "leader board" of the 5 theses that had been accessed the most in the previous month. *The GSC Newsletter* has continued to run articles on electronic theses almost every month highlighting heavily accessed theses and popular keyword searches of the thesis database. The continuing presence of electronic theses related articles is a reflection of the interest and enthusiasm of the graduate student community for this new high visibility distribution mechanism for the fruits of their labor.

Critical Mass

Nothing succeeds like success. Conversely, an archive of electronic theses with a dearth of content is only attractive to highly motivated early adopters. A leak in a storage facility caused severe water damage to many circulating copies of theses from the 1960s. Rather than photocopying and binding new circulating copies from the archival copies, the decision was made by the library, the Graduate Office, and the Office of the General Counsel to scan the theses and mount PDF versions in CaltechETD. To date over 300 theses have been scanned. The General Counsel provided guidance with respect to copyright and access limitations for the older material. With over 400 theses and

dissertations, originally written between 1930 and 2003, CaltechETD is an attractive repository, both for continuing contributions from the campus and to organizations creating federated online document collections.

General Lessons

Planning is crucial. Identifying likely problems and questions, then working through the challenges before changing graduation requirements, went a long way toward allaying the concerns of all interested parties.

Don't get bogged down trying to get everything perfect before you start. Allow yourself room to learn and grow.

Don't be afraid to temporarily take on other responsibilities. Re-drafting the thesis guidelines within the library accelerated the process and allowed for accurate documentation of procedural changes on the first pass.

Trial balloons are welcome. A voluntary start-up phase allowed for low volume, highly motivated individuals to identify glitches. These are the people best able to handle the unexpected and to help find creative solutions.

Direct contact with decision makers is extremely helpful. ETD Team benefited from the Graduate Dean's acquaintance with individual librarians.

Promote advantages to the students and faculty. Graduate students spend 3-7+ years of hard work to arrive at the end product, a thesis. Let them know that electronic theses may be downloaded more than 100 times/month; print theses rarely see 2 uses/year. Professors get excited about making older research easily accessible.

Mission critical work. Theses are central to the mission of research universities. Making theses globally available is easily understood to be within the mandate of the institution and of the library. This implicit mandate creates an opportunity for constructive collaboration.

Offer support and reassurance. Creating templates and giving workshops allows the library staff to demonstrate an awareness of the students primary concerns.

A little effort can engender a lot of good will. Copyright clearances can be arduous for graduate students. Smoothing the way by offering to identify the optimal contact at various publishers is a tremendous service for the students. Librarians quickly amass a working knowledge of the proper contacts and develop methods of ferreting out contacts at publishers where they have not yet identified the rights person. An hour or less of librarian effort can save more than a day's worth of student frustration.

Be flexible and responsive. Create a team to support the ETD effort and be ready to assist students with whatever problems arise.

Appendix. Caltech's Chronology

- 3/99 Ed Fox, from Virginia Tech, made a presentation on electronic theses at Caltech. Arden Albee, Graduate Dean, committed to implementing an electronic thesis program at Caltech.
- 4/99 Caltech joined NDLTD.
- 4/01 ETD-db software loaded by CLS.
- 6/01 CaltechETD opened up to voluntary submissions. - Rod Kiewiet, the new Graduate Dean sent email to those who have just graduated.
- 11/01 Library's ETD Team created.
- 12/01 Preservation scanning of water-damaged theses begins.
- 4/02 Library made presentation to Graduate Studies Committee after having worked through policy and presentation issues.
- 6/02 Graduate Dean officially made ETD submission mandatory. Library staff offered series of classes - 50-60 students attended.
- 7/02 Mandatory submission policy begins. New Graduate Dean appointed.
- 9/02 Library classes - 20 students attend.
- 12/02 Scanned theses begin to be added to the database
- 1/03 71 "new" theses in CaltechETD.
300+ "old" scanned theses in CaltechETD.
Library classes -- 10 students attend.