Moving Electronic Theses from ETD-db to EPrints: The Best of Both Worlds

A Project Briefing Presented at the
CNI Spring 2010 Membership Meeting
Baltimore, MD
April 12, 2010

Betsy Coles
Technical Manager, Digital Library Systems
California Institute of Technology
bcoles@library.caltech.edu

Katherine Johnson
CODA Coordinator and Metadata Librarian
California Institute of Technology
kjohnson@library.caltech.edu

ABSTRACT

Caltech Library Services' first digital archive came online in April 2001 using EPrints software from the University of Southampton. We began collecting electronic theses early as well: voluntary deposit of Ph.D. theses began in 2001, and became mandatory in July 2002. The thesis collection was hosted on the ETD-db software platform developed at Virginia Tech. By 2008 it became clear that we needed to consolidate our repository platforms. We decided to move our electronic theses to EPrints Version 3, the platform in use for our institutional repository. We did not, however, want to lose the many unique features of the Virginia Tech ETD-db software, such as thesis-specific workflow, the ability for staff to communicate with thesis authors via email from within the ETD-db interface, and fine-grained, thesis-specific access controls. We also wanted to add new features that were not available in either platform, such as tracking the progress of a thesis through the complex local approval and release process, and the ability to store related documents, such as signed thesis forms and permissions letters, with the thesis but in a "dark" area of the record visible only to repository administrators.

This briefing explains what was involved in the transition from ETD-db to EPrints for our thesis collection and how the Caltech Library took advantage of the flexibility of the EPrints platform to meet our requirements. It also suggests ways that other institutions may be able to adopt and build on what we've done, and why EPrints 3 may be a good electronic thesis repository solution for other institutions.

The Caltech Library's collection of digital archives (CODA) has a long history. CODA grew out of a longstanding commitment to scholarly communication and open access. Since its inception the CODA project has always been part of the Library's core mission, and almost everyone is involved in some way. There is no separate CODA department or function within the Library, nor is there any special funding—CODA support comes from the general operating budget.

Our first CODA archive, consisting of Computer Science Technical Reports, came online in April 2001. This archive used the EPrints repository software developed at the University of Southampton. Our first

electronic thesis was deposited in July of the same year into a separate Caltech ETD repository which used the ETD-db software developed at Virginia Tech. We chose ETD-db over EPrints for archiving theses in 2001 because ETD-db was designed from the start for theses and dissertations. It included many attractive features to support thesis workflow and metadata, as well as the ability to withhold or restrict public access to files pending patent applications or journal publication.

In the early days of electronic theses at Caltech, we had to overcome some not-unexpected skepticism about the electronic format from faculty and even students, so a full transition to electronic theses was accomplished in a series of steps. Electronic submission was voluntary at first and our old system of paper submission remained in place for those who preferred it. The initial period of voluntary participation demonstrated that electronic thesis submission was not only feasible but popular. It also gave faculty time to study and evaluate the process and approve the next phase, which was mandatory electronic submission for all Ph.D. candidates beginning in July 2002.

The library has continued to print and bind a copy of each thesis for preservation purposes, although the electronic version is now the official copy of record and we are committed to its preservation as well. Since students may include material that can't be reproduced on paper (video, software, data files, etc.), having the electronic version as the copy of record allows the most complete version of the thesis to serve as the official one. And having the library print a paper copy from the electronic submission has had the salutary side effect of freeing us from dealing with discrepancies between electronic and paper versions – a situation which used to arise fairly often when we asked students to submit printed copies!

Shortly after bringing up the Caltech ETD-db repository, we began a program of retrospective conversion of older theses, which are scanned and submitted to the thesis database by library staff. Each year we receive between 180 and 250 current theses and add, on average, 500 older theses. As of April 2010, our thesis repository contained 5,518 theses and dissertations, of which 1,433 were born digital and submitted by their authors. The database is heavily used: in March 2010, the Caltech thesis repository website received 19,000 visits from more than 14,000 unique visitors. More than 22,000 document files (.pdf, .ps, or .doc) were downloaded along with about 4,300 supplemental files including video, data, and software.

Our electronic thesis program has been an unqualified success in almost all respects. However, an assessment in late 2008 concluded that some changes were needed. Caltech Library Services was operating with reduced staff and very limited resources. Maintaining two different repository software platforms, one for theses (ETD-db) and one for everything else (EPrints), led to unproductive and duplicative effort. We felt that there could be more forward development if we could consolidate our repositories on a single platform. Also, the ETD-db software was growing old and lacked many of the features of modern repository software.

Development of CaltechAUTHORS, our EPrints institutional repository for materials other than theses, had of course been ongoing as well, and by 2008 it contained more than 11,000 items. In July 2008 CaltechAUTHORS was upgraded from EPrints version 2 to EPrints 3, and we were delighted with the results. EPrints 3 offers many sophisticated and attractive features, including a plugin architecture that

lends itself to extension and customization; a web-based management interface; an active development team at Southampton and an active, contributing user community worldwide; and a contract services organization (EPrints Services) with whose work we were familiar from earlier projects. We began to investigate whether we could move our theses to EPrints 3 without sacrificing the features of ETD-db we had come to value.

In addition to technical platform-migration issues, we hoped that a migration would provide an opportunity to rework and improve the instructions given to students submitting theses, and to coordinate those instructions with the many on- and off-campus organizations which are involved in the granting of degrees, such as the academic departments and the Graduate Studies Office. In the case of on-campus groups, such as the Graduate Office, we did not want to disturb our painstakingly arrived-at relationships. On the other hand, for submission of theses to Proquest/UMI, we hoped to be able to take advantage of automation in new ways. We also wanted to streamline the handling of theses by library staff. With these workflow and user-level goals in mind, we began by creating a system-independent description of thesis workflow at Caltech.

The specific technical goals of the proposed migration were to:

- Retain useful thesis-specific features of ETD-db –
  - Thesis-specific metadata fields and search capabilities (committee, major/minor options, etc.)
  - Ability to communicate with authors via email, from within the system interface
  - Special limited-access categories for thesis materials (restricted, withheld), at the file level as well as the record level
- Convert metadata to an EPrints schema and migrate our thesis files without loss or distortion of any kind
- Add brand-new features
  - New metadata elements including advisor(s), major and minor fields of study, Graduate Studies office approval date, degree-award date, funders, references, internal notes, and others
  - Ability to store and identify related documents (permissions, signed thesis forms, source files in formats such as LaTeX, etc.) along with each thesis, but in a hidden part of the record
  - Additional capability for automatically generated emails for status updates
  - Expanded ability to track theses through the complex approval process

With the assistance of staff from EPrints Services, we determined that EPrints 3 could meet all of these technical needs, and we made the decision to migrate our electronic thesis collection from ETD-db to EPrints 3.

Of course, the same limitations on resources that constrained our day-to-day operations also limited the staff time available to perform a major system migration. Therefore we settled on a combination of

local and outsourced development to complete the project. We contracted with EPrints Services in England to do the following programming and customization:

- Metadata conversion scripts
- Metadata and data migration scripts
- New email trigger function in EPrints
- New complex data structure for degree-granting departments
- New functionality to accommodate "hidden" documents, e.g. permissions letters

Other activities that required intimate knowledge of local practices and/or involvement of metadata and public services staff were carried out locally:

- Metadata analysis and modifications to EPrints standard metadata scheme for theses
- Modification and customization of the user interface
  - Public display of metadata
  - Submission screens
  - New browse and search options (by advisor, field of study, etc.)
- Customization of the system "workflow" (movement of theses through the stages of deposit and approval)
- Migration of persistent URLS within the Caltech Library's locally developed PURL system
- Screen and help text customized to local needs
- Creation of a new web guide for student submitters
- New documentation for library staff

The project was officially begun in March 2009 and was scheduled to end in late August of the same year, a 6-month window. The actual completion date was one month delayed, to late September. Our original goal was to have the system up and running by the beginning of the new academic year. This was accomplished with a week to spare.

Various library staff at different levels were involved throughout the entire process. In the development phase, four individuals with major responsibilities in the digital repository and electronic theses efforts were included:

- the coordinator for the separate ETD-db repository, who is also the metadata coordinator for the EPrints repositories, and who served as the project manager for the migration
- the programmer/system administrator for the digital repositories
- the subject liaison librarian who had been most heavily involved in the CODA development
- the Metadata Services Group staff person responsible for processing submitted theses

Once the needs and requirements of the project had been defined and the contract with EPrints signed, the programmer/system administrator became the liaison between the two organizations, as she had all the requisite technical knowledge to see the contract work through to its completion.

EPrints Services staff in Southampton were fortunately available to work on our project within our desired time-frame. Their promptness in getting software changes back to us on time (and often ahead of schedule) allowed us to stay on track as well. The initial version of the migration script was ready for us by the beginning of June, and the vast majority of the development had been done by the end of June. We were thus able to spend our summer debugging and finalizing the software, developing the interface design, and updating the documentation both for end-users and staff.

Integration of the contracted code was done at Caltech, as was testing of all elements of the migration. A staged approach to testing the full migration process was used: starting in early July, we created several test repositories and converted progressively larger numbers of records with each test iteration. The final test involved a full-scale conversion and migration.

The final migration push began in earnest in the latter half of August. Initial training of library staff began at the same time. We wanted to minimize the number of "in-process" theses requiring migration, so it was important that the submission buffers on the old platform be as empty as possible before the final conversion. This required getting the Graduate Office to approve as many pending theses as possible.

As we got closer to the final project phases additional staff, including other public services librarians, were brought in to review what had been accomplished already, focusing somewhat on the functional aspects, but more on the interface itself: documentation and screen displays. The final testing phase was open to all library staff, and we received input from many staff ranging from the University Librarian to circulation staff.

The actual migration was done over a weekend in late September. Submission of theses by students and staff was frozen for 48 hours. The searchable public system continued to be available throughout. The actual export of data from ETD-db, conversion of metadata, and import to the new EPrints 3 repository, CaltechTHESIS, took less than 8 hours, although full-text indexing of the EPrints 3 database, which ran in the background, took two days to complete. As hoped, CaltechTHESIS was open for business on Monday morning.

Within hours, a graduate student submitted the first thesis under the new system. We, of course immediately emailed him, saying:

> "Congratulations! You are the first student to have deposited his thesis into the new CaltechTHESIS database. Would you mind giving us some feedback on your experience? Ease of use, problems encountered, confusion?"

His short and sweet response:

> "Thanks! I was wondering when this had changed, realized it must have been recent. I found the submission quite easy, it took me only a couple of minutes for the whole process."

Our experience in the months since we went live with CaltechTHESIS has confirmed our initial impression that we now have a modern, flexible system that provides a better user experience and smoother process for both students and library staff.

As is invariably true in the real world, we are not completely "done" with work on CaltechTHESIS. Currently underway or on the horizon are:

- Data cleanup remaining from the conversion (ongoing)
  - integrating some supplemental thesis files that were stored in a separate database under ETD-db
  - fixing problems related to HTML tagging in the abstract field
  - In the long term, adding missing metadata, such as advisors, to our new metadata fields
- Export plugin for Proquest/UMI's XML metadata format, to allow us to automate the process of sending theses to Proquest/UMI via FTP (currently being tested)
- ETD-MS format for OAI harvesting (in process)
- User interface tweaks and improvements (a never-ending task!)
- Ability to search on additional fields, including full-text status
- Upgrade to the recently released EPrints version 3.2 which includes many important new features, including an abstracted storage layer supporting Postgres and Oracle as well as cloud storage, a REST-style interface, major speed improvements, SWORD2 support, rewritten search and index functions, and innumerable user and management interface improvements.
- Upgrade to new faster hardware and Redhat 5 64-bit Linux operating system
- Implementation of available EPrints add-on features such as
  - the IRSTATS statistics module
  - the DROID/PRESERV plugin to support preservation status monitoring
  - deposit via the SWORD protocol
- A faculty bibliography widget, to complement the automatically-generated faculty bibliography html code that is already available for inclusion in faculty pages
- Perhaps most important: complete the task of documenting the migration in technical terms and uploading migration scripts into the EPrints wiki, so that others may make use of what we've done.

Of course, much of this work will apply to CaltechAUTHORS and our other repositories as well as to CaltechTHESIS; that's the beauty of a single platform.

Viewed from the perspective of six months after go-live, the migration project has definitely been a success. It was completed with reasonable expenditure of both funds and staff time. We confirmed our hope that Eprints 3 would be easy to customize and would serve as a robust platform for theses. We also found that first-rate support is available from the EPrints Services team. We have met our most important goal of modernizing our thesis platform without giving up any of the thesis-specific features that we had come to depend on, as well as almost all of our subsidiary goals:

- A better and more easily supported technical system
- Reduced staff time requirements for metadata group staff processing theses
- A better user experience for
    - Students submitting theses
    - Library staff processing theses
    - Searchers worldwide

The project represents the "best of both worlds" in at least two ways. First, we avoided the "buy vs. build" dilemma by contracting out specific parts of the migration development work to experts, while using our own resources where our local skill set could be put to best use and where local control was crucial to success.

Second, we now have, in EPrints3, a single, full-featured repository platform for all of our institutional materials while retaining the valuable thesis-specific features that our old ETD platform provided. We gain substantial efficiencies and also are well-placed to keep up with new services such as SWORD deposit and cloud data storage. We look forward to beginning our second decade of institutional repository management with a strong and flexible foundation.

LINKS

- This Presentation: http://resolver.caltech.edu/CaltechLIB:2010.001
- CaltechTHESIS – http://thesis.library.caltech.edu
- CaltechAUTHORS – http://authors.library.caltech.edu
- CODA – http://library.caltech.edu/digital
- Thesis workflow planning document: http://library.caltech.edu/etd/System_Independent_Thesis_Workflow.pdf
- Web guide for student submitters –http://libguides.caltech.edu/theses
- EPrints software – http://software.eprints.org
- EPrints Services – http://www.eprints.org/services/
- ETD-db software – http://scholar.lib.vt.edu/ETD-db/developer/index.shtml