Developing a Common Submission System for ETDs in the Texas Digital Library

Adam Mikeal, Tim Brace, Scott Phillips, John Leggett, Mark McFarland

Texas A&M University Libraries, The University of Texas Libraries

Abstract

The Texas Digital Library is a consortium of universities organized to provide a single digital infrastructure for the scholarly activities of Texas universities. The four current Association of Research Libraries (ARL) universities and their systems comprise more than 40 campuses, 375,000 students, 30,000 faculty, and 100,000 staff; while non-ARL institutions represent another sizable addition in both students and faculty. TDL's principal collection is currently its federated collection of ETDs from three of the major institutions; The University of Texas, Texas A&M University, and Texas Tech University. Since the ARL institutions in Texas alone produce over 4,000 ETDs per year, the growth potential for a single state-wide repository is significant.

To facilitate the creation of this federated collection, the schools agreed upon a common metadata standard represented by a MODS XML schema. Although this creates a baseline for metadata consistency, there exists ambiguity within the interpretation of the schema that creates usability and interoperability challenges. Name resolution issues are not addressed by the schema, and certain descriptive metadata elements need consistency in format and level of significance so that common repository functionality will operate intuitively across the collection.

It was determined that a common ingestion point for ETDs was needed to collect metadata in a consistent, authoritative manner. A working group was formed that consisted of representatives from five universities, and a state-wide survey of the state of ETDs was conducted, with varied levels of engagement with ETDs reported. Many issues were identified, including policy questions such as open access publishing, copyright considerations and the collection of release authorizations, the role of infrastructure development such as a Shibboleth federation for authentication, and interoperability with third-party publishers such as UMI. ETD workflows at six schools were analyzed, and a meta-workflow was identified with three stages: *ingest*, *verification*, and *publication*. It was decided that Shibboleth would be used for authentication and identity management within the application.

This paper reports on the results of the survey, and describes the system and submission workflow that was developed as a consequence. A functional prototype of the ingest stage has been built, and a full prototype with Shibboleth integration is slated for completion in May of 2007. Demonstrators of the application are expected to be deployed in fall of 2007 at three schools.

Introduction

In 2005, four Association of Research Libraries (ARL) universities—The University of Texas, Texas A&M University, The University of Houston, and Texas Tech University—came together to form a unified digital infrastructure for scholarly activity in the state of Texas. Together, these four universities and their systems comprise more than 40 campuses, 375,000 students, 30,000 faculty, and over 100,000 staff. Dealing with digital collections at this scale presents unique challenges; these four universities produce over 4,000 theses and dissertations a year.

An enormous amount of intellectual capital exists between these institutions that is not readily available to users across the State (Bush 1945); the Texas Digital Library (TDL) was created to meet that need. Its charge is to serve as the "center of excellence for the creation, curation, and preservation of digital scholarly information for the State" (Leggett 2006). Electronic Theses and Dissertations have played an important role in TDL since its formation, and its largest collection is currently the federated collection of ETDs from three of the original institutions; The University of Texas, Texas A&M University, and Texas Tech University. The collection continues to grow, with the recent addition of University of Texas at Arlington, and more are expected to follow in the near future.

Background

A decision was made very early in the organization of TDL to focus on the development of a federated collection of ETDs, as the two largest member institutions already had a working ETD system, and were publishing their ETDs independently. To facilitate the creation of the collection, the four member schools agreed upon a common metadata standard for ETDs, expressed as a MODS XML schema (Surratt 2006). This schema created a baseline for metadata collection and dissemination, and allowed the schools to federate the collections into a single access point. At the same time, it allowed for a degree of consistency in the presentation of the records; by conforming to the MODS schema, the schools were guaranteeing that certain fields—like the degree discipline and degree date—were present and would appear in a certain format.

As beneficial as this first step was, there were still areas of divergence among the records collected from the schools, usually due to ambiguities present within the interpretation of the schema. This created usability and interoperability challenges, and made basic repository tasks, like browsing and searching the collection, more difficult than necessary. Name resolution issues are not addressed by the current iteration of the MODS schema, nor are there any controlled vocabularies for fields such as discipline or major. Additionally, some of the descriptive metadata elements need consistency in format or level of significance in order to provide intuitive functionality across the entire collection.

The Common Submission System

It was determined that in order to support our goal of a single, unified collection of ETDs that was both usable and scalable a single ingestion point was needed in order to collect the metadata in a consistent, authoritative manner. In spring of 2006, a working group was formed with members from six different universities across the state. Its charge was to identify the issues and policies involved with ETD workflows in the member institutions, and make recommendations to the team charged with the development of the actual application.

The first task of the working group was a state-wide survey to identify the current state of ETDs at the participating institutions. Each group member interfaced with the appropriate staff on their campus to perform the necessary research; workflow documents in story format were then produced for each school that described the processes currently in place (or in one instance, planned for implementation in the near future). These workflow stories were used to identify a baseline workflow that could describe the process from all institutions to at least some degree; this is described in the following section. Beyond these high-level workflow descriptions, a questionnaire was distributed with specific questions related to ETD policies (see Appendix A); the goal of these documents was to ascertain the level of diversity with regard to processes and practices concerning doctoral and masters theses at the various participating institutions.

Baseline workflow

Through the workflow stories gathered during this process, a baseline ETD workflow was identified that was able to express each institution's current or planned practices to some degree of accuracy (Figure 1).

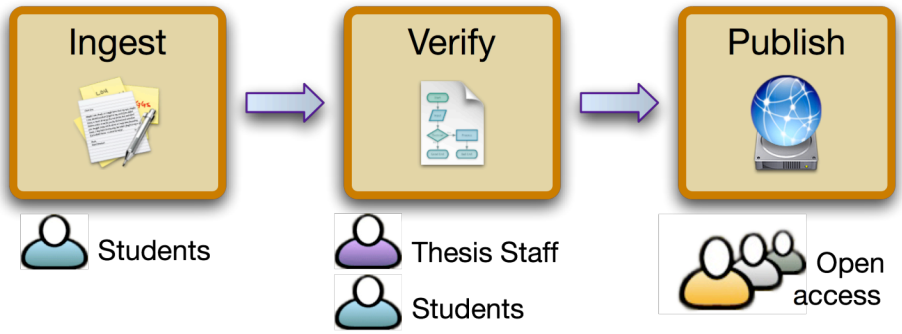

Figure 1: Baseline workflow visualization.

While every institution had variability within their workflow, and all schools did not necessarily implement each detail step instituted by every other school. This workflow is available as Appendix B, and is comprised of three very basic stages:

> **Ingest**. The *ingestion* stage describes the workflow undertaken by the student to submit her thesis to the university. In this stage, the student will need to access the web application that accepts the document, authenticate their identity and validate their ability to submit to the site, enter the metadata for the document, and finally upload the thesis or dissertation itself, plus any supplementary files that might need to be attached to it. Any additional parameters or questions, such as a copyright release forms or submittal to UMI can also be handled at this stage.
>
> **Verify**. The *verification* stage deals with the iterative process that involves authorized staff at the institution and the author. The staff must have the ability to filter, browse, and search the submitted ETDs across a variety of criteria; and the interaction between the student and the staff may be completely logged and easily searchable through the interface. There is a wide variety of practices and implementations at this stage, and each university will need a system that can adapt to it's own individual needs.
>
> **Publish**. *Publication* is the final stage. Once designated as approved by the authorized staff, and if an embargo has not been requested, the file will be deposited in an open-access repository. A record of the ETD may be added to the home institution's library catalog, and possibly to a distributed or external catalog system, such as WorldCat, as well. Other recipients of the ETD, as designated by the student or by policy of the home institution (for example, UMI), will receive their copies of the ETD at this point.

Requirements analysis

At the same time the surveys and questionnaires were being completed by the working group members, TDL leadership met to discuss the technical requirements for a common submission system. Decisions were made on both technical and policy levels regarding the scope of the application, decisions regarding copyright and access control issues, and which schools would be the first participant in the demonstration system.

Scope Analysis

The most immediate issue involved TDLs level of engagement with each of the three main stages of the baseline ETD workflow. The lowest level of engagement would be leaving each institution to manage the ingest and verification stages independently, and offer TDL as the final recipient and publisher for the ETDs. Another model would have seen TDL providing a web-based ingest stage, handing the documents back to the home institution for processing through the verification stage, and then receiving them back again for final deposit and publication. However, neither of these models adequately addressed the original needs of metadata authority and consistency.

Ultimately, it was decided to implement a fully functional ETD workflow system—from ingestion, through verification, on to final publication. This monolithic, all-encompassing system assures that the metadata in question is never outside our control, and allows us to enforce consistent standards and conventions on the metadata fields that could otherwise allow for ambiguities. It is this model that was ultimately chosen for implementation. Similar systems—such as the ones developed and deployed in the US by UMI and OhioLink—have been proven successful in defining a set of procedures usable by various institutions.

Author's Rights and Access Control

For the student, the most concerning aspect of any submission system are the rights they will be required to release to the publisher (in this case, to TDL). There are many issues involved, and dealing with original works created in an academic environment only makes the situation less clear. Will the author be allowed to decline submission into the system, or demand individualized access restrictions? How will the student indicate when a publication hold is required for patent or journal copyright reasons? Will TDL offer a publication solution that is restricted to a specific scope, such as users within the state or a particular campus?

TDL has a practical and philosophical alignment with the concepts of open access, and the decision was made that any submission through the TDL system would require the acceptance of a non-exclusive license that allows TDL to publish the work in an open manner in perpetuity. Since the license is non-exclusive, the author's rights are fully preserved, and they are free to publish or sell their work with a third-party dissertation distributor if they so choose. Experience has shown that if mandatory open access is a policy at the student's institution (as it is at The University of Texas, for example), the great majority of students accept this as part of their graduation requirements (Jewell 2006).

Journal and patent holds will be handled by an embargo system, where the student can flag a thesis or dissertation for delayed publication, and the institutional staff that manage the verification stage will evaluate and process that request. Eventually, however, without continued action by the student to preserve the embargo, the work will ultimately revert to a published state.

Deployment Schedule

Since the University of Texas and Texas A&M provide the majority of ETDs produced in the state of Texas, it was decided that the initial system would be built to meet the needs of these two institutions and their sizable graduate student population. Under the assumption that these two schools contained the greatest diversity in their student needs and offerings, other institutions should be able to use the resulting system with little to no modification of their existing processes.

For those institutions already producing ETDs, our goal was to make the transition to the TDL system as transparent and painless as possible. For those not yet producing ETDs, it is hoped that the system will prove easy and convenient enough to motivate their migration away from paper documents, thus easing their entrance into the ETD community.

System Implementation

Based on the workflow analysis, policy decisions, and the federated nature of TDL's organization, three primary requirements were identified that dictated the architecture of the system: 1) a robust and usable interface that required minimal training for the verification staff, and no training for students; 2) a secure, integrated and scalable authentication mechanism for system access, and 3) the ability to contribute the system back to the community as a turnkey application. To address these architectural requirements, we selected two open source software solutions: the Manakin/DSpace digital repository application, and Internet2's Shibboleth authentication middleware.

User Interface

Because this system will be accessed by the majority of its users only once, it is imperative that the usability of the interface is a high priority (Shneiderman 1997). Students under stressful deadlines will often not have the luxury of training, or the tendency to read documentation. Our design incorporates the contextual directions embedded into the interface in question (Figure 2). In addition, graduate office staff responsible for the verification stage of the ETD workflow are geographically dispersed throughout the state, which makes the training process challenging, and sometimes infeasible.

Because UT and TAMU were already using DSpace as the digital repository for their existing ETD collections, it was the logical economic decision to extend this platform for the common submission application. Manakin is a DSpace project that provides the ability to easily modify the look-and-feel of individual repository collections (Phillips 2007). Manakin uses two primary mechanisms to allow for this type of customization: Themes and Aspects.

Themes stylize the look-and-feel of a specific collection or entire repository, and are distributed as self-contained packages. Themes may integrate with existing websites, visualize metadata, and otherwise change the interface. Aspects are interactive extensions to DSpace that provide new features for the digital repository. Aspects may provide functionality such as specialized searches or custom workflows. For the common submission system, we created both an Aspect and a Theme; the Theme modified the look-and-feel of the system to provide for the unique interface needs shown in Figure 1. The Aspect implemented two of the three main stages of the ETD workflow: the ingest stage used by the students, and the verification stage used by the graduate office staff.

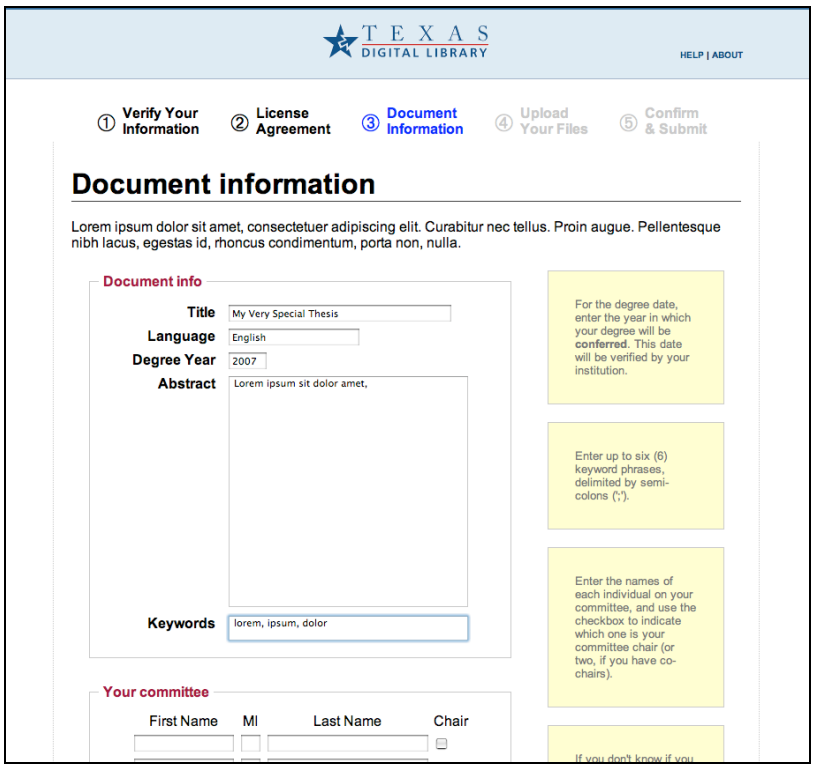

Figure 2: Screenshot of prototype interface for student submission application showing embedded help text.

Distributed Authentication

Like any large web-based application, authentication and identity management are necessary components. It was very quickly determined that a centralized identity system that required users to create new, single-use accounts was not feasible. TDL simply does no have the staffing resources to manage the volume of identities—for both students and staff—that would be created. Since each institution is already providing authentication services and identity management for their campus, a federated solution that could leverage this existing architecture was decided to be the best approach.

Shibboleth is a standards-based, open source authentication middleware that provides web-based Single SignOn. An Internet2 initiative, Shibboleth allows complete separation of applications from account management, through the introduction of *service providers* and *identity providers*. Service providers are end-user applications such as a digital repository or ETD submission system, while identity providers manage user accounts and provide authentication. Both types of providers are organized into a federation, which provides three main benefits:

> **Federated identity management**. By tapping into the pre-existing account databases maintained by each member school, TDL was freed from duplicating the large task of account management across the state. Additionally, each member school retains full control over the accounts for which they are responsible, which serves to further protect the students' privacy.

Secure metadata. By creating a trusted relationship between TDL and each member institution, we were able to access information about each user—official name, email, school and department, etc.—that has already been vetted to some degree by university staff, and can be accepted with a much higher level of assurance than data that is simply collected from the user online.

Flexibility. The architectural structure of Shibboleth was a good fit for TDL, as it allows for the addition of new identity providers at any time without any disruption of the existing authentication mechanisms. Additionally, each school is free to design and scale its own identity management system in the manner it sees fit; it can disregard any interoperability issues with the rest of the federation other than basic Shibboleth compliance.

Having decided on Shibboleth as the distributed authentication system, there were several significant implementation decisions to be made:

Federation membership. TDL established the first state-wide Shibboleth federation in the state of Texas. Unlike the initiatives in place in Europe and Australia, the United States has been slow to develop authentication infrastructure for higher education. The InCommon federation was evaluated as an option, but rejected because of its stringent membership requirements. The difficulty associated with meeting those requirements was determined to be infeasible for many of the smaller schools in the state of Texas.

General access identities. In order to provide authentication services to schools not yet members of the TDL Shibboleth federation, a fall-back option is available through a general access identity provider operated by TDL staff.

User Metadata. The metadata schemas used in the TDL Shibboleth federation are based on standard identity schemas, one of which is the eduCause eduPerson standard. However, there were several key data points still not available; specifically, the student's major and major code, and their graduation date. To address this issue, a new schema was created that defined the additional fields.

Turnkey Application

There is high demand for a system that provides an end-to-end turnkey solution for ETDs, from ingest, through verification, to publication. This demand is evident through the existence of commercial providers and the various non-profit initiatives aimed at addressing this problem, each providing a different level of engagement with the three workflow stages. The TDL Common Submission System will provide a highly-integrated application that is tailor-made for the unique needs of ETDs.

Manakin offers several new tools to be able to build a more modular interface for the repository. Using the previously mentioned DSpace extensions called Aspects, new functionality can be added to the repository. These extensions are self-contained and can be easily shared between repositories. The TDL Common Submission System, based on Manakin's architecture, will be easily distributable and adaptable to the ETD workflow challenges present at many other academic institutions. Once the application has been tested and deployed within the state of Texas, all source code, documentation, and training materials will be made publicly available under an open source license.

Conclusion

This paper has described the results of the workflow analysis and site surveys performed by the Texas Digital Library's ETD Working Group, and introduced the application that was designed as a result of that work. We discussed the policy decisions that arose during the planning stages, and the archtecture decisions that were made in order to consider those requirements: Shibboleth and Manakin/DSpace.

The TDL Common Submission System is currently under active development by staff at TDL, Texas A&M, and The University of Texas. A functional prototype of the ingest stage has already been built, and a fully developed ingest prototype with Shibboleth integration will be completed in May 2007. Demonstrators of the complete application—ingest, verification, and publication—are expected to be deployed in a testing capacity at both Texas A&M and The University of Texas in the fall of 2007. After one or more semesters of concurrent deployment with the existing systems, a full deployment is slated for 2008.

References

Bush, Vannevar. "As We May Think". *The Atlantic Monthly*. July 1945, 101-108.

Jewell, Christine, Lynn Judge, William Oldfield, and Lisa Tomalty-Crans. "Required Open Access to ETDs: Technical, logistical, and philosophical implications". In *Proceedings of the 9th International Symposium on Electronic Theses and Dissertations*. June 7—10, 2006. Quebec City, Canada.

Leggett, John, Mark McFarland and Drew Racine. "The Texas Digital Library: A Business Case". Prepared for and published by the Texas Digital Library, July 2005, revised July 2006.

Phillips, Scott, Cody Green, Alexey Maslov, Adam Mikeal, and John Leggett. "Introducing Manakin: Overview and Architecture". In *Proceedings of the 2nd International Conference on Open Repositories*. January 23—26, 2007. San Antonio, TX, USA.

Schneiderman, Ben. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Publishing Company, 1997.

Surratt, Brian. "MODS Meets Manakin: Innovations in the Texas Digital Library's Thesis and Dissertation Collection". In *Proceedings of the 9th International Symposium on Electronic Theses and Dissertations*. June 7—10, 2006. Quebec City, Canada.

Appendix A: ETD questionnaire

1. Do your students currently turn in electronic
    a. Dissertations?
    b. Masters theses?
    c. Masters reports?
    d. Any other type of document, such as a "Record of Study"?

2. Are any of the above mandatory? If so, which ones?

3. Do you currently use an online submission system for your ETDs?

4. If so, is it
    a. Home grown
    b. UMI/Proquest
    c. Other vendor (please specify)

5. Do you send your dissertations to UMI/ProQuest?

6. Do you send your Masters theses (or any other documents) to UMI/ProQuest?

7. What place in your overall graduation workflow is occupied by the electronic component of thesis/dissertation submission?

8. If there is no electronic component to the thesis/dissertation, is there any electronic component at all?

9. Do you have a central office (e.g., a Graduate Studies office) that processes all the dissertations and theses/reports for your institution, or is it handled at the department level (or somewhere else)?

10. What type of review / corrections process is used by your university, and who is responsible for its implementation?

11. Is any editing (however minor) of the submitted document done by university staff, or must all document changes be made by the student?

12. How are the theses/dissertations made available to the public? Physical access? A digital repository?

13. When is the access granted in relation to the overall document workflow? In a batch, or a trickle?

Appendix B: Baseline ETD Workflow

1. Ingest stage
    a. Obtaining ETD metadata from student
    b. Obtaining thesis / dissertation from student
    c. Obtaining supplementary files from student
    d. Obtaining copyright / release authorization from student

2. Verification stage
    a. Gathering all required documentation by OGS / student
    b. Verifying metadata
        i. Descriptive metadata for ETD
        ii. Committee members/chairs
    c. Iterative proofing process with OGS / student
        i. OGS makes corrections
        ii. Corrections sent to student
        iii. Student reflects changes in original document
        iv. Original document replaced with new copy
    d. ETD marked as *cleared*

3. Publish stage
    a. ETD w/ proper metadata migrated to public repository
        i. Appropriate access is assigned based on release authorization
        ii. Embargoed ETDs are potentially left in a holding area and not migrated to the repository at all
    b. Records for each ETD added to the library catalog
        i. Optional feedback loop with cataloging for further data correction (corrections made to the catalog are reflected in the repository)
    c. Copy (complete or metadata only) is moved to TDL central repository