Large-scale ETD repositories: A case study of a digital library application

Adam Mikeal, James Creel, Alexey Maslov, Scott Phillips, John Leggett, Mark McFarland[1]

Texas A&M University Libraries
5000 TAMU
College Station, TX USA
+1 (979) 862-3887
{adam, jcreel, alexey, scott, leggett}@library.tamu.edu

The University of Texas Libraries[1]
1 University Station S5400
Austin, TX USA
+1 (512) 495-4129
mcfarland@austin.utexas.edu

ABSTRACT

We describe the implementation of a statewide system for managing and preserving electronic theses and dissertations (ETDs) from Texas universities. We further explain the theoretical, technical and political issues that arose during the implementation of this system. These issues range from technical components developed by TDL—such as a customized workflow management application and adding OAI-ORE capabilities to DSpace—to human-centered issues such as stakeholder engagement and participation. Our experiences reflect the challenges, expected and unexpected, that others will face when attempting to build digital library applications to scale.

Categories and Subject Descriptors

H.3.7 [**Information Systems: Information Storage and Retrieval**]: Digital Libraries—*dissemination, standards, systems issues, user issues.*

General Terms: Design, Human Factors.

Keywords

Electronic theses and dissertations, scalable systems, digital library architecture, electronic document workflow

1. INTRODUCTION

Theses and dissertations have long been a primary scholarly product of universities. In recent years, there has been a growing trend of moving the workflow and management of these scholarly works into the digital realm [18, 22]. An electronic thesis and dissertation (ETD) system provides all necessary management services throughout the lifecycle of a student's thesis or dissertation, from initial submittal to the thesis office, through the iterative review and approval process, to the final publication in a digital repository [25]. The following strategies are essential to a robust ETD system:

- **Stakeholder participation:** to effectively engage the different stakeholders that must participate in any ETD effort;
- **Flexible architecture:** to adapt to the various roles and expectations of departments, colleges, universities, and university systems;
- **Scalability:** to handle the growing volume of records and network traffic that will coincide with the growth in the system;
- **Integratability:** to leverage universities' existing expertise and data, such as identity management or pre-existing repositories.

The Texas Digital Library (TDL) is a consortium of public and private institutions from across the state of Texas [16, 44]. Over the past three years, TDL has been engaged in a large-scale project to develop an ETD management and publication system for the state. We began this project by engaging CIOs, graduate deans, and library administrators from multiple schools in Texas. The technical implementation of the project then followed: we created a MODS profile to support ETDs [41]; established a statewide identity federation using Shibboleth [45]; built a federated collection for ETDs leveraging the DSpace repository platform [34]; developed a customized ETD submittal and management application using the Manakin interface framework [26]; and modified DSpace to enable automatically harvested collections using OAI-PMH and the new ORE protocol [30]. This system is currently in production at Texas A&M University, and is scheduled to be deployed at The University of Texas in Austin in fall of 2009.

Because of the physical size and population of Texas, the creation of statewide information systems presents unique challenges. Six public university systems exist in Texas, and together they comprise more than 40 campuses, nearly 400,000 students, and over 130,000 faculty and staff; The University of Texas alone processes nearly 1,400 ETDs every semester [26]. There are dozens of smaller public and private institutions that are not attached to one of the six systems, and Texas is second in the nation only to California in the total number of graduate theses and dissertations. Dealing with digital collections at this scale is difficult and requires a different approach from what is effective for a single university.

One of TDL's earliest initiatives was the previously mentioned federated collection of ETDs. Beginning with contributions from Texas A&M University and The University of Texas at Austin, the collection has been steadily growing in both number of records and participants [46]. The size and diversity of this collection introduced several challenges—such as inconsistent metadata and differences in storage and access methods—that

JCDL'09, June 15-19, 2009, Austin, Texas, USA.
Copyright 2009 ACM 978-1-60558-322-8/09/06... $5.00.

made offering effective tools and services across the collection difficult.

Addressing these challenges prompted the creation of a cohesive ETD system for statewide deployment. Such an effort fell squarely within TDL's mission, and cost sharing for such a project provides TDL members with clear technical and economic benefits [16, 24]. This paper presents issues we have encountered establishing this statewide ETD repository service and the lessons learned over the course of this multi-year project. Many of these issues are relevant to other digital library applications of similar scale and complexity.

2. ISSUES

A statewide ETD management and publication system must address many issues, both of a technical and administrative nature. Following are the most important of these issues and their resolutions in the TDL ETD repository system. We begin with the human-centered issues of stakeholder engagement and metadata, because these were foundational to the success of the technical components.

2.1 Stakeholder Participation

Multiple levels of engagement are necessary for such a system to succeed. Technical support from university CIOs is required for many of the infrastructure-level components, such as authentication. The Deans of graduate education must be involved in policies and in communicating necessary information to students and faculty. Finally, the tasks of archiving and preservation of theses and dissertations has historically fallen to the library [7, 21].

Policymakers at several levels must be engaged in any campus-level ETD effort, and introducing multiple campuses only results in more complexity. Many policy questions can be grouped into related categories: author's rights and licensing issues, workflow models, positions on open access, and the system's relationship to third-party publishers [1, 3].

Finally, the most affected stakeholder groups are the end-users of the resulting web applications. Students, administrative staff, and the general public will all interact with the repository in different ways [14, 23, 38]. Technical support and training must be addressed from the outset of the project to adequately support these groups.

The Texas Digital Library began to engage many of the necessary stakeholders early; long before development had begun on several of the critical components. In 2006, a working group was formed with members from six Texas universities. Its charge was to identify the issues and policies involved with ETD workflows in the various member institutions [25]. In 2007, TDL was awarded a National Leadership Grant from the Institute of Museum and Library Services (IMLS) [9]. This grant helped to coalesce necessary buy-in from the university administrations involved.

Developing effective end-user support is a continual process. TDL has organized its user support structure with the concept of *support tiers*: student support requests are handled by the originating institution, and issues from administrative staff are shared across the consortium through the TDL support infrastructure. TDL has also established a regular schedule of hands-on training classes for the administrative support staff.

A useful lesson learned from this experience is that care must be given to limit the scope of the policy analysis, at least during the initial stages of development. As a result of the working group's policy surveys, it became clear that focus should be limited to the two schools slated to be the system's initial users. Once the system was in place and stable, changes could be introduced as necessary to adapt to other school's needs.

2.2 Metadata Challenges

In a statewide effort, content is contributed from multiple institutions, and each institution may rely on different software solutions to manage their data. Metadata generated by these applications will have subtle differences that make creating effective tools and services across the combined data difficult, such as ambiguities in date encodings. Additionally, most existing metadata standards fail to provide for the unique properties of ETDs. For those that do, many fail to provide the level of specificity required to sufficiently describe the item. ETD-MS, published by the Networked Digital Library of Theses and Dissertations, is a Dublin Core-based metadata standard with a low barrier to entry and broad application, focusing on repository interoperability [27]. Because of this, there are several ETD-specific properties that create difficulties, such as disambiguating the roles assigned to multiple advisors.

Metadata was a primary focus for TDL at its inception in 2005. A working group was formed that consisted of members from six schools across the state. This group wrote a descriptive application profile for ETDs using the Metadata Object Description Schema (MODS) [37]. This profile was based on earlier work by the Networked Digital Library of Theses and Dissertations [27]. Version 1.0 of the *MODS Application Profile for Electronic Theses and Dissertations* was published in December of 2005 [41].

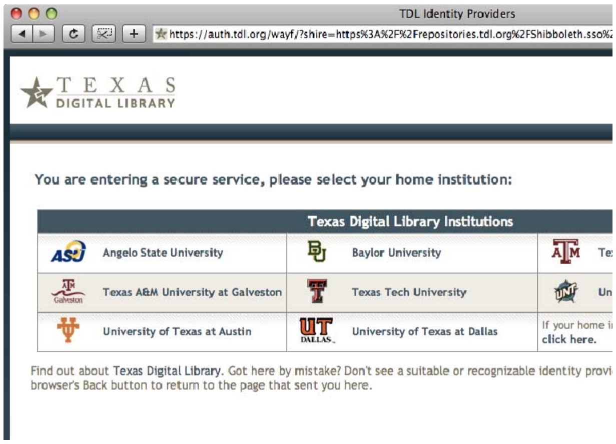
TDL schools participating in the federated ETD collection quickly adopted the new standard. However, ambiguities in the schema led to inconsistencies in the metadata collected from those schools. Additionally, the metadata encoded in the MODS profile required conversion to the more general ETD-MS standard for external collaboration, a process for which no standard mapping or convention existed. To address these concerns the working group released the *TDL Descriptive Metadata Guidelines for Electronic Theses and Dissertations* in June 2008. It contained an updated MODS profile, mappings to qualified Dublin Core and ETD-MS schemas, as well as recommendations for best practices. Under the new guidelines, several schools were forced to make minor modifications to the metadata in their existing ETD collections, a time consuming and delicate process [37].

Our experience with ETD metadata has impressed on us the lessons that the cost to change direction once established is high, and it is far better to spend an adequate amount of time during the early stages of the project and make informed decisions regarding schemas and metadata collection practices. Unfortunately, it is all too likely that even with a reasonable amount of preparation and foresight, there will be decisions and assumptions that must be revisited later. Careful planning may at best limit, but not prevent, those costs, so one must provide for metadata migration from the outset.

2.3 Identity Management

Authorization and identity management are critical components of any web-based application platform. Typically, each application maintains its own “island” of identities, forcing users to manage and remember additional credentials for every new service they use. This model presents several problems in a statewide service: the scale of the system presents significant management challenges, including the liability assumed when storing security credentials. Additionally, the usability of the system is greatly increased when users are not required to remember yet another username and password for what is often a single-use service.

The solution is to use the existing identity management infrastructure that most universities have already implemented. This lowers both the support and maintenance costs for the overall system. Such a distributed identity management system allows access to information that has been vetted by the university itself. Furthermore, data exchange eliminates the need for students to re-enter data they have already provided to their registrar, reducing the possibility of errors [12, 31]. However, by pushing the authorization task back down to the local institution, one is forced to tackle the issue of integrating data from various identity management systems. Each university may choose to implement their identity management systems in a variety of platforms, from Kerberos to LDAP to Microsoft’s Active Directory.



TDL Identity Providers

https://auth.tdl.org/wayf/?shire=https%3A%2F%2Frepositories.tdl.org%2FShibboleth.sso%2

TEXAS DIGITAL LIBRARY

You are entering a secure service, please select your home institution:

Texas Digital Library Institutions		
Angelo State University	Baylor University	Te
Texas A&M University at Galveston	Texas Tech University	Un
University of Texas at Austin	University of Texas at Dallas	If your home i click here.

Find out about Texas Digital Library. Got here by mistake? Don't see a suitable or recognizable identity provi
browser's Back button to return to the page that sent you here.

Figure 1: Screenshot of TDL's Shibboleth Federation

When TDL was formed, there was no statewide system to support distributed identity management. Shibboleth—an open-source middleware application—was chosen for several reasons: it provides a mechanism for sharing trusted data between services and identity providers; usability of the system is enhanced because it provides for a graphical identification process for the originating institution (see Figure 1); and as an Internet2 initiative, it was developed with higher-education use cases in mind [39]. TDL established the infrastructure for a statewide federation, starting with identity providers from Texas A&M University and The University of Texas.

Implementing a statewide Shibboleth federation presented several challenges for some universities. As the number of involved partners grew, many of the participating schools were found to lack the availability of technical resources required to implement and maintain the local components, and some schools lacked a single, centralized authentication system at the local level. This challenge was addressed with centralized training and direct assistance to individual schools. Centralized training was developed in conjunction with the UT System in the form of a one-day “ShibFest”. Technical staff from around the state were invited to participate in hands-on installation and configuration training. Individual technical support was provided to several member schools over the duration of the federation development period. The new statewide federation has added nine members over the past two years.

The lessons learned from this process encourage patience. Many different groups are involved in creating this type of cooperative effort. Coordination between distinct groups is a slow process, and will be measured in months and years, not weeks. Additionally, while in an ideal world this type of fundamental infrastructure would be a service provided to the library, it is often the case that no statewide entity exists with the capacity to coordinate among the various parties. This was the case in Texas, and so it fell to TDL to initiate the project. Now that the statewide Shibboleth federation is well underway, direct management is being handed off to the Lonestar Education And Research Network (LEARN), a consortium of Texas universities that provides networking infrastructure [20].

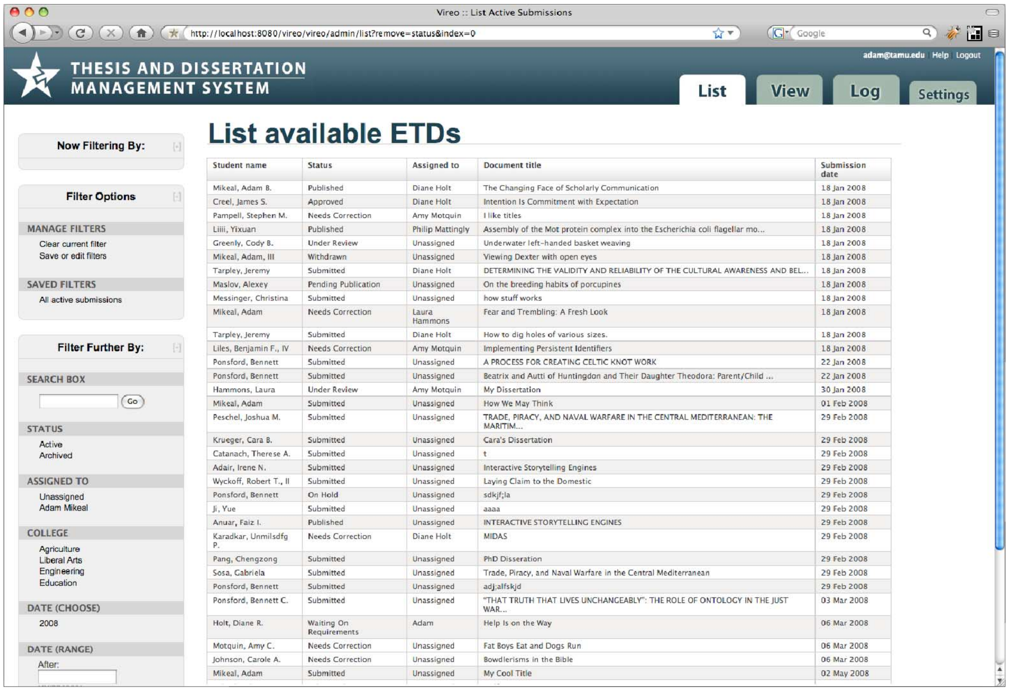
2.4 Document Workflow

Theses and dissertations have complex workflows, originating with a student, then moving to the staff member responsible for verification and sometimes validation against an external style standard. This verification step may be an iterative process, requiring several exchanges between the staff member and the student before the thesis office or graduate school is satisfied with the output and approves the student’s work [10].

Additionally, when establishing a statewide system, one will certainly encounter the added difficulty that each institution approaches this workflow from a different perspective, or engages the ETD process in a different manner. Some institutions will have a centralized system in a dedicated thesis office with many full-time employees; others will have a single employee in the graduate school that may be responsible for other duties; and still others will have a distributed system that pushes the responsibility for verification and approval out to the college or department level. Any workflow system selected must be able to support the existing workflow and policy choices of all institutions involved in the effort.

TDL’s greatest investment in programming effort has been toward a customized submission workflow application. TDL had already adopted DSpace as its repository platform, and analysis revealed that the standard submission workflow available in DSpace was not sufficient to meet the complex needs of the ETD submission process, which involved students, their faculty advisors, and staff from multiple divisions’ independent groups [42, 43]. Since DSpace’s newest interface layer provides for the ability to add customized functionality to the repository [32, 33, 34], TDL created *Vireo*, an ETD management and workflow application.

Since Vireo is built directly in the repository application stack, it allows for a very tight relationship between workflow management and the repository. This makes actions such as withdrawing or modifying items post-publication much more straightforward, and allows for all aspects of the system to be executed together in a single instance, if desired.


Vireo :: List Active Submissions
http://localhost:8080/vireo/vireo/admin/list?remove=status&index=0
Google
THESIS AND DISSERTATION
MANAGEMENT SYSTEM
adam@tamu.edu
Help
Logout
List
View
Log
Settings
Go

Now Filtering By:

Filter Options

MANAGE FILTERS
Clear current filter
Save or edit filters

SAVED FILTERS
All active submissions

Filter Further By:

SEARCH BOX

STATUS
Active
Archived

ASSIGNED TO
Unassigned
Adam Mikeal

COLLEGE
Agriculture
Liberal Arts
Engineering
Education

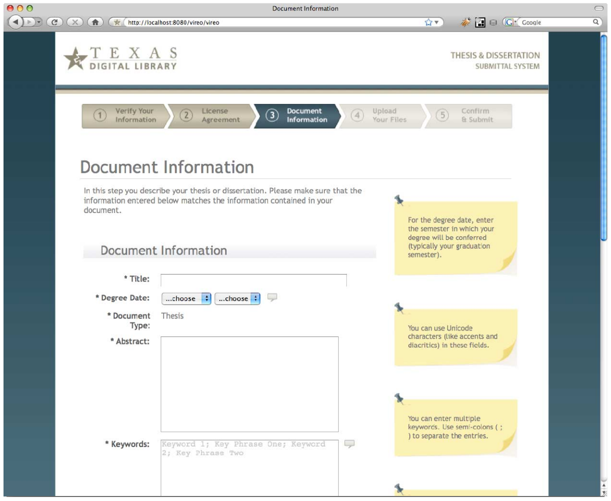
DATE (CHOOSE)
2008

DATE (RANGE)
After:

List available ETDs

Student name	Status	Assigned to	Document title	Submission date
Mikeal, Adam B.	Published	Diane Holt	The Changing Face of Scholarly Communication	18 Jan 2008
Creel, James S.	Approved	Diane Holt	Intention Is Commitment with Expectation	18 Jan 2008
Pampell, Stephen M.	Needs Correction	Amy Motquin	I like titles	18 Jan 2008
Liiii, Yixuan	Published	Philip Mattingly	Assembly of the Mot protein complex into the Escherichia coli flagellar mo...	18 Jan 2008
Greenly, Cody B.	Under Review	Unassigned	Underwater left-handed basket weaving	18 Jan 2008
Mikeal, Adam, III	Withdrawn	Unassigned	Viewing Dexter with open eyes	18 Jan 2008
Tarpley, Jeremy	Submitted	Diane Holt	DETERMINING THE VALIDITY AND RELIABILITY OF THE CULTURAL AWARENESS AND BEL...	18 Jan 2008
Maslov, Alexey	Pending Publication	Unassigned	On the breeding habits of porcupines	18 Jan 2008
Messinger, Christina	Submitted	Unassigned	how stuff works	18 Jan 2008
Mikeal, Adam	Needs Correction	Laura Hammons	Fear and Trembling: A Fresh Look	18 Jan 2008
Tarpley, Jeremy	Submitted	Diane Holt	How to dig holes of various sizes.	18 Jan 2008
Liles, Benjamin F., IV	Needs Correction	Amy Motquin	Implementing Persistent Identifiers	18 Jan 2008
Ponsford, Bennett	Submitted	Unassigned	A PROCESS FOR CREATING CELTIC KNOT WORK	22 Jan 2008
Ponsford, Bennett	Submitted	Unassigned	Beatrix and Autti of Huntingdon and Their Daughter Theodora: Parent/Child ...	22 Jan 2008
Hammons, Laura	Under Review	Amy Motquin	My Dissertation	30 Jan 2008
Mikeal, Adam	Submitted	Unassigned	How We May Think	01 Feb 2008
Peschel, Joshua M.	Submitted	Unassigned	TRADE, PIRACY, AND NAVAL WARFARE IN THE CENTRAL MEDITERRANEAN: THE MARITIM...	29 Feb 2008
Krueger, Cara B.	Submitted	Unassigned	Cara's Dissertation	29 Feb 2008
Catanach, Therese A.	Submitted	Unassigned	t	29 Feb 2008
Adair, Irene N.	Submitted	Unassigned	Interactive Storytelling Engines	29 Feb 2008
Wyckoff, Robert T., II	Submitted	Unassigned	Laying Claim to the Domestic	29 Feb 2008
Ponsford, Bennett	On Hold	Unassigned	sdkjf;la	29 Feb 2008
Ji, Yue	Submitted	Unassigned	aaaa	29 Feb 2008
Anuar, Faiz I.	Published	Unassigned	INTERACTIVE STORYTELLING ENGINES	29 Feb 2008
Karadkar, Unmilsdfg P.	Needs Correction	Diane Holt	MIDAS	29 Feb 2008
Pang, Chengzong	Submitted	Unassigned	PhD Disseration	29 Feb 2008
Sosa, Gabriela	Submitted	Unassigned	Trade, Piracy, and Naval Warfare in the Central Mediterranean	29 Feb 2008
Ponsford, Bennett	Submitted	Unassigned	adj;alfskjd	29 Feb 2008
Ponsford, Bennett C.	Submitted	Unassigned	"THAT TRUTH THAT LIVES UNCHANGEABLY": THE ROLE OF ONTOLOGY IN THE JUST WAR...	03 Mar 2008
Holt, Diane R.	Waiting On Requirements	Adam	Help Is on the Way	06 Mar 2008
Motquin, Amy C.	Needs Correction	Unassigned	Fat Boys Eat and Dogs Run	06 Mar 2008
Johnson, Carole A.	Needs Correction	Unassigned	Bowdlerisms in the Bible	06 Mar 2008
Mikeal, Adam	Submitted	Unassigned	My Cool Title	02 May 2008

Figure 3: Screenshot of Vireo's staff admin interface



Document Information

In this step you describe your thesis or dissertation. Please make sure that the information entered below matches the information contained in your document.

Document Information

* Title:

* Degree Date: ...choose ...choose

* Document Type: Thesis

* Abstract:

* Keywords: Keyword 1; Key Phrase One; Keyword 2; Key Phrase Two

For the degree date, enter the semester in which your degree will be conferred (typically your graduation semester).

You can use Unicode characters (like accents and diacritics) in these fields.

You can enter multiple keywords. Use semi-colons (;) to separate the entries.

Figure 2: Screenshot of Vireo's student submittal interface

Vireo supports multiple workflow models and a limited degree of customization for each school. It can be described as a single application with two interfaces: a submission interface for students (see Figure 3), and an administrative interface for thesis clerks and other staff (see Figure 2). The student interface is built on the familiar "wizard" paradigm, and walks the student through a simple 5-step submission process. It is assumed that most users of this interface will only see the application once, so attention was given to ensure that any documentation and help was available on-screen and embedded directly within the interface.

The administrative staff interface uses a "job queue" paradigm, and the basic screen shows an unfiltered list of all submissions, sorted by submission date (see Figure 3). A set of *filters* are available to the left of the list, creating a faceted browsing experience, and allowing the staff to create customized queries to show precisely the set of submissions applicable to his or her work. Staff are granted significant control over the application environment, and can customize the filters and columns that appear on the screen.

Selecting an individual submission from the list allows the staff member to edit the record, manage the metadata of the item, change its state in the workflow, and interact with the student through the web interface and email. All actions taken on the submission are recorded in a permanent log, and notes and other files may be attached to the record as necessary.

2.5 Repository Platform

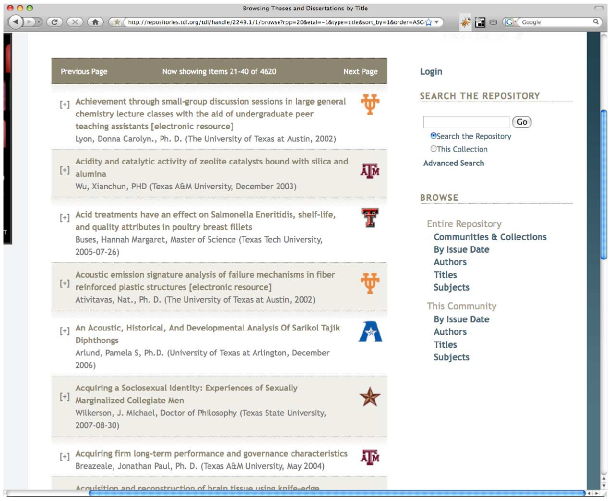
The end result of a statewide ETD system is the ability to deliver the scholarly output of your state's universities to the world. To achieve this a repository platform with the following capabilities is necessary:

- **Publication**. Basic tools and services for browsing, searching, retrieval, ingestion and workflow management should be provided, as well as policy enforcement for management of content administered by the content owners.
- **Branding**. When creating a federated collection that combines content from multiple schools, institutional branding becomes a significant factor. The repository platform should provide the ability to alter the visualization and display of the content at multiple levels, to either embed

the repository fully within an established web presence, or merely brand individual items as they are displayed.

- **Infrastructure**. The capability to build additional tools (such as customized workflows as mentioned above) within the repository platform provides a basic set of services without reinventing the wheel, such as session management, caching, database connectivity, user privilege management, and a storage access layer.

Robust digital repository software platforms provide all these services to some degree. Depending on the platform, there will be a greater or lesser emphasis placed on the different properties of the repository.



[+] Achievement through small-group discussion sessions in large general chemistry lecture classes with the aid of undergraduate peer teaching assistants [electronic resource]
Lyon, Donna Carolyn., Ph. D. (The University of Texas at Austin, 2002)

[+] Acidity and catalytic activity of zeolite catalysts bound with silica and alumina
Wu, Xianchun, PHD (Texas A&M University, December 2003)

[+] Acid treatments have an effect on Salmonella Eneritidis, shelf-life, and quality attributes in poultry breast fillets
Buses, Hannah Margaret, Master of Science (Texas Tech University, 2005-07-26)

[+] Acoustic emission signature analysis of failure mechanisms in fiber reinforced plastic structures [electronic resource]
Ativitavas, Nat., Ph. D. (The University of Texas at Austin, 2002)

[+] An Acoustic, Historical, And Developmental Analysis Of Sarikol Tajik Diphthongs
Arlund, Pamela S, Ph.D. (University of Texas at Arlington, December 2006)

[+] Acquiring a Sociosexual Identity: Experiences of Sexually Marginalized Collegiate Men
Wilkerson, J. Michael, Doctor of Philosophy (Texas State University, 2007-08-30)

[+] Acquiring firm long-term performance and governance characteristics
Breazeale, Jonathan Paul, Ph. D. (Texas A&M University, May 2004)

Acquisition and reconstruction of brain tissue using knife-edge

Figure 4: Screenshot of item-level branding in TDL's repository

At the time of TDL's formation, the initial members already had well-established repository projects, and several were using the DSpace platform [40]. Texas A&M University was heavily involved in the development of the new XML-based interface layer for DSpace, called Manakin [32, 33, 34].

DSpace provides an out-of-the box repository solution. Most of the basic repository infrastructure is handled by DSpace: the structure imposed on the items into communities and collections; the internal structure of the item's files and metadata, and the tools necessary to manage the repository content and users. The addition of Manakin provides the ability to atomically add and remove functionality from the repository in discrete chunks called *Aspects* [13, 34]. Manakin's *Themes* apply branding and interface-level changes at the item, collection, or repository level (see Figure 4).

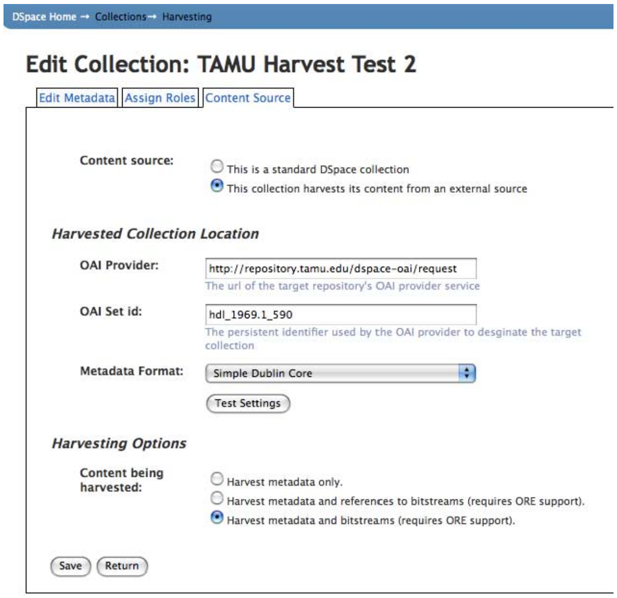
2.6 Interoperability

Interoperability between repository systems enhances preservation through the dissemination of multiple copies of the items [7] and increases availability by providing multiple points of access. Two significant challenges exist that limit the ability of repository systems to interoperate: divergent metadata standards and the immaturity of existing transmission protocols.

As discussed in section 2.2, the lack of sufficiently descriptive metadata standards for ETDs necessitates the creation of customized schemas. Unfortunately, in order to engage repositories outside your control, it is necessary to "step down" from a precise internal schema into an inexact generic standard such as Dublin Core. This challenge can be found in reverse in the situation of a federated repository with exacting metadata standards that must incorporate content from a repository unable to provide metadata with similar granularity.

The Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) is the current *de facto* standard for metadata interoperability between repository systems [8, 47]. However, the models for transmitting content between repositories are still immature. Several possibilities exist, such as embedded METS documents within the PMH harvest, the Digital Item Declaration Language (DIDL) subset of MPEG21 [2], or the newer OAI Object Reuse and Exchange (ORE) protocol [30]. However, none of these options enjoy widespread adoption as of yet, which reduces their effectiveness for repository interoperability.

The first statewide system deployed by TDL was a federated collection for ETDs that aggregated items from the individual member schools. The initial implementation utilized a variety of manual batch processes, and relied on scripts to export, transform, and import the data from one repository to another. This model suffered from significant drawbacks: it was fragile, inflexible, and unable to scale as new schools were added.



Edit Collection: TAMU Harvest Test 2

Content source:
○ This is a standard DSpace collection
● This collection harvests its content from an external source

Harvested Collection Location

OAI Provider: http://repository.tamu.edu/dspace-oai/request
The url of the target repository's OAI provider service

OAI Set id: hdl_1969.1_590
The persistent identifier used by the OAI provider to desginate the target collection

Metadata Format: Simple Dublin Core

Harvesting Options

Content being harvested:
○ Harvest metadata only.
○ Harvest metadata and references to bitstreams (requires ORE support).
● Harvest metadata and bitstreams (requires ORE support).

Figure 5: Screenshot of new collection harvesting options in DSpace

To address these deficiencies, the harvester/provider model based on OAI-PMH was adopted. As TDL's DSpace repositories already supported OAI-PMH dissemination, TDL allocated resources to add harvesting support into the DSpace platform. This enhancement includes the ability for collection administrators to configure and manage harvested collections, and for the repository software to automatically check for updates to the remote collection (see Figure 5). Next, DSpace was modified to both generate and interpret OAI-ORE resource maps. The combination of these two functional enhancements allows two repositories to automatically replicate each other's content.

The lesson learned during this process was that a robust interoperability solution must be simple and direct. A potential solution investigated was to use DSpace's METS dissemination capabilities to describe items with their structure [19]. However, the lack of well-defined and widely adopted METS profiles creates significant challenges for this solution, challenges the simpler ORE protocol avoids.

2.7 Preservation

The preservation task has always been a significant consideration for digital libraries, and much has been written comparing long-term preservation in the physical and virtual worlds [6, 7, 17, 21, 35, 36]. The challenges that affect the preservation of ETDs are the same challenges faced by any collection of digital content, especially when that content is born digital; preserving bits is not always the same as preserving access to the message [24].

TDL is in a unique position to leverage the geographic size of Texas in establishing its preservation strategy. Following the same two-tiered development strategy used in building its federated repository (see previous section), TDL has constructed a largely manual process in the short-term to ensure geographic distribution and archival storage. Meanwhile, a long-term project is underway to create a fully automatic, distributed preservation network for the state. This network will communicate directly with TDL core services—such as the repository—and will automatically handle the complex preservation metadata and migration utilities necessary for long-term viability.

3. SYSTEM ARCHITECTURE

A robust ETD system requires the technical components identified in sections 2.3 through 2.7. How these components interact in a larger ecosystem will depend on the particular technologies adopted for each component. If each piece is carefully implemented with attention to communication and interoperability, the end result will be a flexible systems architecture.

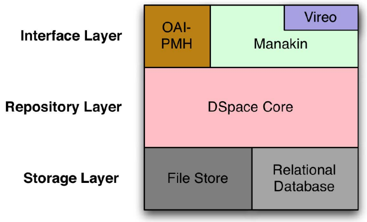


Figure 6: Central components in DSpace repository application stack

One such implementation decision is how tightly many of these components are integrated into the repository. For example, if the workflow is integrated directly into the repository application stack, then additional architecture paradigms are available that keep the submission, workflow and publication tasks together in a single service. Furthermore, several of the other critical components, such as an interoperability transmission protocol or distributed identity management, will need to be implemented only once, rather than in two separate systems.

There are various competing products that can fill the roles in the application stack as described [5, 15, 29]. The application stack chosen by TDL is DSpace with Manakin. Figure 6 shows the tight integration of the Vireo workflow module implemented as a Manakin Aspect. Vireo is thus able to reuse components from the entire stack, such as database and file storage, DSpace's management of items, the interface framework provided by Manakin, and the transmission protocol provided by OAI-PMH.

3.1 Architecture layers

For a statewide system, there are two layers of architecture to consider: one to describe the components at an individual school—a single ETD system *instance*—and a second layer to describe how independent instances interact to form a federation which can properly be described as the statewide ETD repository system.

3.2 Paradigms

To describe an instance, there are two intersecting factors to consider which create 4 possible paradigms: discrete or aggregated collections, and monolithic or distributed services.

In an instance with a discrete collection, all submitted ETDs are directed into the same collection, and the assumption is that all students may be treated the same through the workflow process (this option would be used by a school running a workflow system only for its own students). Allowing aggregated collections allows multiple schools to share the same instance while customizing the workflow for their own students.

A monolithic instance is a repository system that combines the submission, workflow and publication components into the same application instance; it is ideal for smaller schools that have neither the technical resources to run multiple servers or the submission volume to require it. A distributed instance allows a school to separate the submission and verification workflow away from the publication needs of its central institutional repository, providing better performance and scalability.

3.2.1 Simple all-in-one (discrete monolithic)

This paradigm is appropriate for a single institution that desires to host the workflow and publication components together in the same service. This model is *discrete* in that it collects all incoming submissions into a single logical unit. It is *monolithic* in that all aspects of the repository service—submission, workflow, and publication—are handled by the same repository platform (see Figure 7).

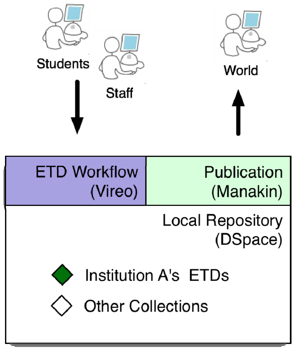


Figure 7: Simple all-in-one model

A smaller school that has the technical capacity to host a repository service is capable of using this model. At a minimum, this will require hardware support for the service and at least a part-time effort from a systems administrator. Additionally, programmer resources would be required if customization of the interface or workflow is required beyond what is available from the base installation. The advantage of this model is its simplicity; by keeping all components of the system in one service, maintenance and administration overhead are reduced.

3.2.2 Cooperative all-in-one (aggregated monolithic)

The aggregated monolithic model differs from discrete models in that it assumes that multiple schools share the instance. The workflow component directs each submission into the appropriate collection based on the student's affiliation. As with a discrete model, all aspects of the service are still served by the same repository platform (see Figure 8).

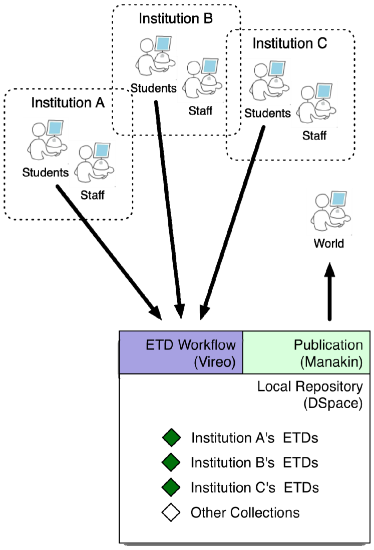


Figure 8: Cooperative all-in-one model

This model would be useful for a university system with multiple universities, a consortium of small schools or even a single school with semi-independent colleges or divisions. For smaller schools that lack the technical resources necessary to host a mission-critical application service, this model allows for resource pooling and the ability to distribute the hosting costs. In any case, in order to support users from multiple institutions, some form of distributed identity management, such as Shibboleth, will be required.

3.2.3 Separated workflow (discrete distributed)

The distributed models separate the submission and workflow components away from the publication component in the repository, resulting in two independent services. This provides added robustness and security to both the repository and the submission platform, and is better able to scale to the needs of a larger institution (see Figure 9).

A larger school with the technical resources necessary to deploy mission-critical services would benefit from this model, particularly if the volume of ETDs processed per year is significant, or if the repository service has grown to a significant size. Additionally, this service independence allows for either component to be replaced with different technologies, assuming that the transmission protocol between the two parts remains consistent.

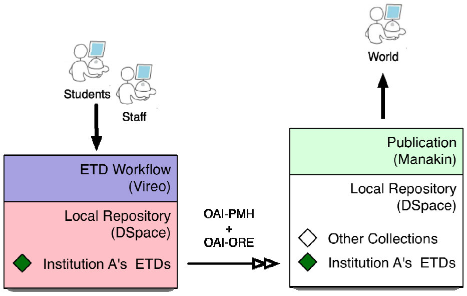


Figure 9: Separated workflow model

3.2.4 Cooperative and separated (aggregated distributed)

Finally, the aggregated distributed model allows for the application of a distributed server architecture to the cooperative model described in section 3.2.2. The workflow component would still direct each submission according to user affiliation, but the submission and workflow service itself has been separated from the final publication platform (see Figure 10).

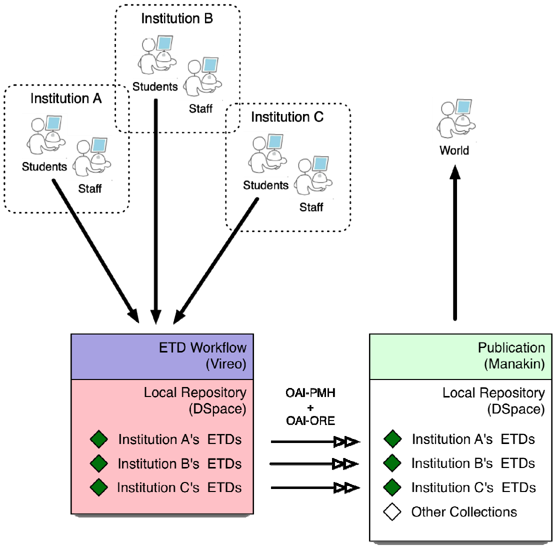


Figure 10: Cooperative and separated model

This model would serve a larger consortium that finds it volume growing to a scale that presents performance challenges to an "all-in-one" paradigm. The service separation relieves these growing pains, and allows for high availability techniques, such as load balancing and caching, to be applied to one or both services independently.

3.3 Federation layer

While the paradigms described above apply to the individual school layer, the second — or federation — layer combines the various instances within a state into a larger interconnected ecosystem. The ability to disseminate and harvest metadata and content ensures that no institution is an island of data. Regardless

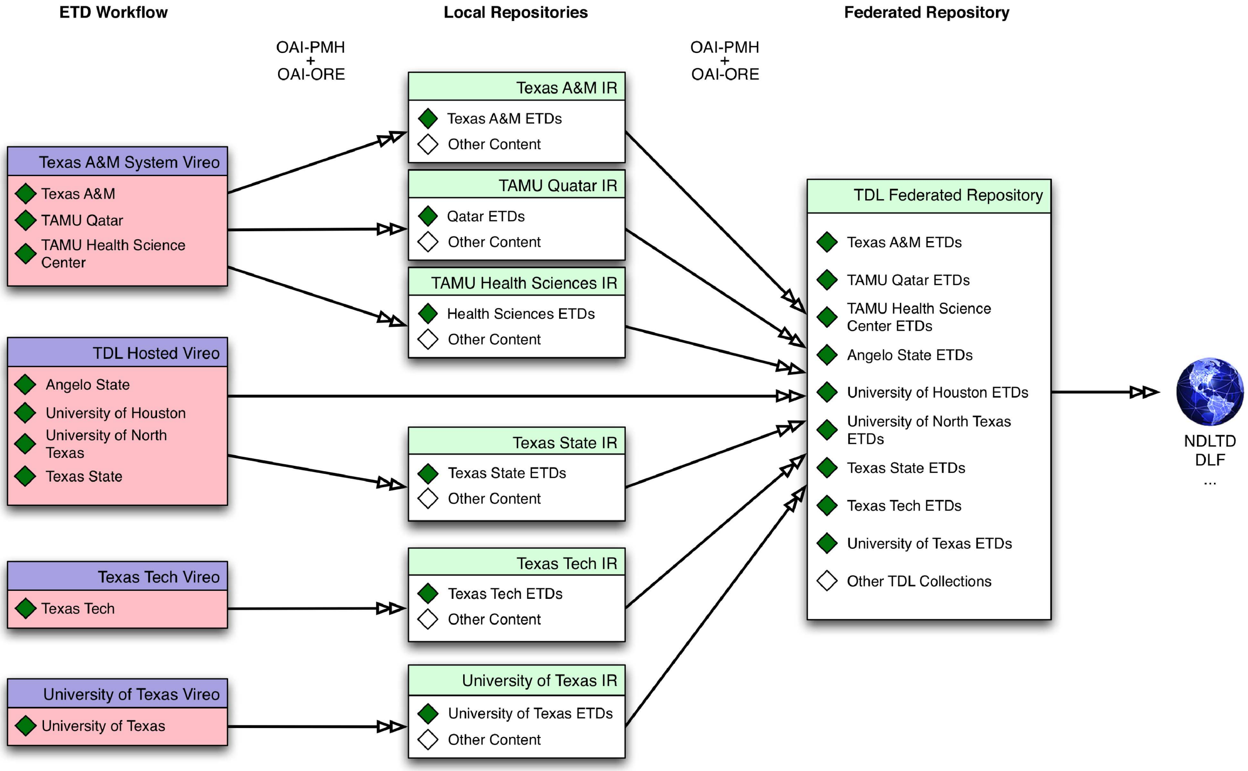


Figure 11: Representation of federation model for Texas

of the paradigm an individual institution might warrant, it can be integrated into the statewide system in a seamless fashion.

3.4 Application in Texas

The particular layers of governance and organization among competing university systems in Texas create unique challenges for architecting a statewide system. In order to respond to these needs, TDL has implemented multiple paradigms within the state. The flexibility to select alternate paradigms for different schools is vital, as the instance for an individual school can migrate from one paradigm to another as the school's needs change over time.

In Figure 11, the workflow components are located in the left-hand column. Larger schools such as The University of Texas and Texas Tech University can be seen to host their own individual Vireo instances that feed into locally hosted repositories. This arrangement corresponds to the separated workflow model discussed above. Each school maintains an individual Vireo instance which is harvested by their independent institutional repositories, and not until the final federation harvest are ETDs combined with those from other schools.

Several smaller schools can be seen grouped together in a single Vireo instance hosted by TDL. This content is fed directly into the federated repository which also harvests content from the local repositories of the larger schools mentioned above. In this example, one of those smaller schools also feeds into its own local repository prior to transmission to the federated collection.

The Texas A&M University System represents a third case, where a cooperative and shared model is being used for several schools within that system. Each of these school's ETDs are then harvested by locally-hosted repositories before final transmission into the TDL federated collection.

This arrangement offers several advantages. Similar schools can be grouped in a variety of ways, allowing organization along administrative boundaries, or arrangement by alternative criteria such as volume of ETDs. For the smaller schools, the ability to share an instance with other schools eliminates duplication of effort and allows for significant cost savings. For larger schools, the flexibility provides room for growth, allowing the service to scale along with the school's needs.

4. RELATED WORK

ETDs are not unfamiliar in the digital library. There are existing applications that have tackled the workflow challenges mentioned in section 2.4, and there are various large-scale efforts at publishing federated ETD collections within a repository context. The ETD repository created by TDL occupies a unique position in this space, based on its scope and architecture.

One of the earliest entries into the ETD workflow management field was the ETD-db system developed by Virginia Tech University in the late 1990's. Showing many of the same basic design decisions as TDL's workflow management application, ETD-db was fundamentally different in that it was not directly integrated into a repository platform. The Tapir project was an early attempt to move these management tasks into DSpace; and although the functional enhancements of Tapir have now been folded into

DSpace proper, the workflow and customization options are limited [11].

There are several examples of national or regional federated collections. NDLTD has produced a union catalog of ETD records for nearly a decade, but this only contains metadata records, and not the full items [27]. EThOS is an impressively comprehensive project in the UK, but like NDLTD, its focus is on federating metadata, and not full item replication or comprehensive workflow management. OhioLINK's ETD repository is closer to the system TDL has developed, but is a fully centralized model, without the distributed architecture that provides so much flexibility for our needs [4].

5. LESSONS LEARNED & FUTURE WORK

From the earliest stages of the statewide ETD repository, we anticipated that member needs would change over time, and we knew that a successful system must adapt to those needs. What was surprising was the rapidity of these changes; how each school's perception of their needs would change and grow as they were engaged in the planning process.

The four strategies outlined earlier in the introduction were developed as a result of the lessons we learned attempting to adapt to our member's ever-changing needs. Stakeholder participation must begin early in the process to ensure buy-in to the core principles of the system. Agreement on core issues and policies ensures that as things change, school participation does not waver. Flexibility is needed at the architectural layer to address organizational requirements that may change during the planning and implementation stages. Our integration of the ORE protocol directly into DSpace allowed us to arbitrarily combine multiple architectural paradigms at the federation layer.

Scalability is vital as the repository system grows from a demonstrator to a mission critical service for the state. Architectural flexibility is a necessary component of scalability at the federation level. Integration with existing university infrastructure allowed us to focus on repository-specific development, ultimately making the whole system lighter and more responsive. Integrating Vireo directly into the repository application stack made possible two of the four architectural paradigms that are currently in use in Texas.

Although the system is currently live at Texas A&M University, there still remains work to move this project into a mature system. Vireo, the document workflow component that was developed by TDL, will be released as an open-source project this year. Support documentation must to be written for this application, and scalability testing is currently underway. As the source code is released, effort will be spent to develop an active user and support community around the system, providing multiple avenues for users and administrators to seek help.

The OAI-PMH harvester and OAI-ORE implementation for DSpace will be given back to the DSpace community. It is expected that this functionality will be added to the DSpace platform within the year; testing and documentation are part of this process. The preservation network mentioned in section 2.7 is still in its seminal stages in Texas; within the next year we expect to see development started and significant progress toward the final outcome.

6. ACKNOWLEDGMENTS

This project is made possible by a National Leadership Grant from the United States Institute of Museum and Library Services (LG-05-07-0095-07).

The authors would like to thank Tim Donohue from the University of Illinois for his feedback on the Vireo project during its beta stages, and Thomas Dowling of OhioLINK for his assistance and advice during the planning stages of our repository.

7. REFERENCES

[1] Bailey, C.W. Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals, Association of Research Libraries, 2005.

[2] Bekaert, J., Hochstenbach, P., and Van de Sompel, H. "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library" D-Lib Magazine, Volume 9, Number 11, November 2003. http://doi:10.1045/november2003-bekaert

[3] Chan, L. "Supporting and enhancing scholarship in the digital age: the role of open access institutional repositories", Canadian Journal of Communications, Volume 29, Number 3, 2004. p 277-300.

[4] Dowling, T. Personal Interview in capacity as Asst. Director of Library Systems, OhioLINK, April 2008.

[5] Fedora Commons. http://www.fedora.info/

[6] Granger, S. "Emulation as a Digital Strategy", D-Lib Magazine, October 2000.

[7] Hedstrom, M. "Digital Preservation: A Time Bomb for Digital Libraries", Computers and Humanities, Kluwer Academic Publishers, 1998, pp. 189–202.

[8] Hochstenbach, P., Jerez, H., and Van de Sompel, H. "The OAI-PMH static repository and static repository gateway". In Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, Houston, Texas, May 27 - 31, 2003. p210-217.

[9] Institute for Museum and Library Sciences, National Leadership Grant. "The Texas ETD Repository: Promoting our Scholarship and Preserving our Legacy". PI, John Leggett; Co-PIs, Adam Mikeal and Jay Koenig (LG-05-07-0095-07).

[10] Jewell, C., Judge, L., Oldfield, W., and Tomalty-Crans, L. "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.

[11] Jones, R. "The Tapir: Adding E-Theses Functionality to DSpace". Ariadne 41. October 2004. http://www.ariadne.ac.uk/issue41/jones/

[12] Jøsang, A. and Pope, S. "User Centric Identity Management". In Proceedings of AusCERT, 2005

[13] Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C., Loingtier, J-M., and Irwin, J. "Aspect-oriented programming". In Proceedings of ECOOP'97, Lecture Notes in Computer Science, Volume 1241, June 1997. p220 - 242.

[14] Kim, J. "Finding Documents in a Digital Institutional Repository: DSpace and Eprints." In Proceedings 68th Annual Meeting of the American Society for Information Science and Technology (ASIST), Charlotte, NC, USA, October 28 - November 2, 2005.

[15] Kortekaas. C. "Making Fedora easier to implement with Fez". Open Repositories, San Antonio, Texas, January 23 - 26, 2007. http://espace.library.uq.edu.au/view.php?pid=UQ:11924.

[16] Leggett, J., McFarland, M., and Racine, D. “The Texas Digital Library: A Business Case”. Prepared for and published by the Texas Digital Library, July 2005, revised July 2006.

[17] Levy, D. “Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation”, Proceedings of the 3rd ACM Conference on Digital Libraries, ACM, Pittsburg PA, June 24–27 1998, pp. 152–160.

[18] Levy, D. and Marshall, C. “Going Digital: A Look at Assumptions Underlying Digital Libraries”, Communications of the ACM, ACM, April 1995, pp. 77–84.

[19] Library of Congress. "METS: An Overview and Tutorial". Metadata Encoding & Transmission Standard Home Page. http://www.loc.gov/standards/mets/METSOverview.v2.html.

[20] Lonestar Education and Research Network. http://www.tx-learn.org/

[21] Lorie, R. “A Methodology and System for Preserving Digital Data”, Proceedings of the Joint Conference on Digital Libraries 2002, ACM, Portland OR, July 14–18 2002, pp. 312–319.

[22] Lynch, C. and Lippincott, J. "Institutional Repository Deployment in the United States as of Early 2005" D-Lib Magazine, Volume 11, Number 9, September 2005. http://doi:10.1045/september2005-lynch

[23] McCray, A., and Gallagher, M. “Principles for Digital Library Development”, Communications of the ACM, ACM, May 2001, pp. 49–54.

[24] Mikeal, A., Green, C., Maslov, A., Phillips, S., and Leggett, J. “Preserving the Scholarly Side of the Web”. In *Proceedings of the Fourth IEEE Latin American Web Congress* (LA-WEB'06), Puebla, Mexico, pp. 162-171. October 2006.

[25] Mikeal, A., Brace, T., Phillips, S., Leggett, J. and McFarland, M. “Developing a Common Submission System for ETDs in the Texas Digital Library”. In *Proceedings of the 10th International Symposium on Electronic Theses and Dissertations* (ETD07), Uppsala, Sweden. June 13-16, 2007.

[26] Mikeal, A., Phillips, S., Koenig, J., Leggett, J., Paz, J., Brace, T., and McFarland, M. “ETD Management in the Texas Digital Library: Lessons learned from a demonstrator”. In *Proceedings of the 11th International Symposium on Electronic Theses and Dissertations* (ETD08), June 2008.

[27] Networked Digital Library of Theses and Dissertations. ETD-MS: An Interoperability Standard for Electronic Theses and Dissertations. http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html

[28] O’Leary, K., Needham, P., Kent, T., and Troman, A. “UK EThOS – Opening access to UK theses”. In *Proceedings of the 11th International Symposium on Electronic Theses and Dissertations* (ETD08), June 2008.

[29] Open Access and Institutional Repositories with EPrints. http://www.eprints.org/

[30] Open Archives Initiative. Object Reuse and Exchange. http://www.openarchives.org/ore/

[31] Pashalidis, A. and Mitchell, C. Information Security and Privacy, chapter “A Taxonomy of Single Sign-On Systems.” Lecture Notes in Computer Science. Springer, 2003.

[32] Phillips, S., Green, C., Maslov, A., Mikeal, A., and Leggett, J. "Manakin Developer's Guide". http://di.tamu.edu/projects/xmlui/resources/DevelopersGuide.pdf

[33] Phillips, S., Green, C., Maslov, A., Mikeal, A., and Leggett, J. “Introducing Manakin: Overview and Architecture”. In Proceedings of the 2nd International Conference on Open Repositories. January 23—26, 2007. San Antonio, TX, USA.

[34] Phillips, S., Green, C., Maslov, A., Mikeal, A., and Leggett, J. “Manakin: A New Face for DSpace”. D-Lib Magazine, Vol. 13 No. 11, November 2007.

[35] Rothenberg, J. Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation, Council on Library & Information Resources, Washington DC, January 1999.

[36] Rothenberg, J. “Ensuring the Longevity of Digital Documents”, Scientific American, January 1995, pp. 24–29.

[37] Rushing, A. “Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations, Version 1.0”. Prepared for and published by the Texas Digital Library, May 2008.

[38] Shneiderman, B. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley Publishing Company, 1997.

[39] Shibboleth (Internet2). In Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Shibboleth_(Internet2)&oldid=205883216 (accessed May 31, 2008).

[40] Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R., and Walker, J. "An Open Source Dynamic Digital Repository", D-Lib Magazine, Volume 9, Number 1, January 2003. http://doi:10.1045/january2003-smith

[41] Surratt, B. “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.

[42] Tansley, R., Bass, M., and Smith, M. "DSpace as an Open Archival Information System: Current Status and Future Directions" Lecture Notes in Computer Science, Volume 2769, 2004. p446-460.

[43] Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., and Smith, M. "The DSpace institutional digital repository system: current functionality". In Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, Houston, Texas, May 27 - 31, 2003, p87-97.

[44] Texas Digital Library. http://www.tdl.org

[45] Texas Digital Library: Shibboleth Federation. Prepared for and published by the Texas Digital Library. March 26, 2007. http://www.tdl.org/documents/TDLShibbolethFederationPolicies.pdf

[46] The Texas Digital Library collection of Theses and Dissertations. http://repositories.tdl.org/tdl/handle/2249.1/1

[47] Van de Sompel, H., and Lagoze, C. “The Santa Fe Convention of the Open Archives Initiative”, D-Lib Magazine, Corporation for National Research Initiatives, February 2000