Integrating ETD Services into Campus Institutional Repository Infrastructures Using Fedora

Martin Halbert

Digital Programs and Systems, Woodruff Library, Emory University, Atlanta, Georgia, USA

Abstract

Research Libraries now face a challenging set of integration tasks when establishing institutional repository system architectures which provide for the range of contemporary digital library services needed on campuses today. This paper will describe the comprehensive implementation of institutional repository services now underway at Emory University, bringing together the campus ETD program with all other digital library services by means of the Fedora repository software and web services. The user-centered process for developing value-added services for graduate research, and intellectual asset policies on top of this infrastructure will also be described. Special attention will be devoted to the aim of accommodating institutional priorities and practices in this endeavor.

1. Introduction

This paper considers the question: *"How can an ETD repository infrastructure provide a foundation for a comprehenisive and extensible campus institutional repository framework?"* Answers to this question are useful for research libraries that are in the process of formalizing their ETD and other repository services. Electronic Thesis and Disseration (ETD) services of various kinds have now been in operation at universities for more than a decade, a very long period in comparison with the range of digital services now offered in research institutions. [1] As a relatively mature concept in academic information technology services, ETD depositories are frequently one of the core components of the umbrella notion of the *institutional repository (IR)* that has emerged in the early twenty-first century. The embedded relationships of campus ETD programs and institutional repositories are still solidifying, and are worth careful analysis and unpacking. This paper reports on work in this regard being undertaken in this regard at Emory University, a large U.S. private research university located in Atlanta, Georgia.

1.1 Evolving Context of Digital Content in Libraries

Virtually all information used in research activities is becoming increasingly digital rather than print-based. This is true for observational data recorded in the field, in-progress analysis, or published scholarly content. Even for Humanities fields still dominated by examination of printed or other analog objects, digital images and word processed monographs have become *de rigueur*. If they are to remain relevant to the academic agenda, research libraries in the twenty-first century are faced with the daunting challenge of creating an entirely new infrastructure for their traditional task of administering intellectual content.

Creating this infrastructure entails the design of systems to deposit, manage, and retrieve an enormous variety of digital content acquired through varied digital ingestion workflows. The range of digital content that libraries may wish to administer is enormous, potentially including: eprints, archived post-prints, locally cached copies of online journals, digitized images of archival materials, locally indexed datasets and databases, electronically encoded etext surrogates of traditional materials, digitized audiovisual recordings, and, of course, electronic versions of theses and dissertations.

1.2 The Need for Coordinated Institutional Repository System Architectures

As individual libraries are confronted by opportunities or demands for them to manage different materials within this range of potentential possibilities, a great deal of idiosyncratic variation has arisen between institutions as to exactly which types of intellectual assets will be administered in this infrastructure. Some institutions may focus on eprints, some on images digitized in archives, etc. There is also variation in what sorts of ingestion, administration, and dissemination services may be provided. Silos of different format-centric infrastructures created by different organizational departments inevitably arise. As research libraries continue to rapidly create new digital library programs and functions to address demands, these siloed systems for administering particular formats have proliferated. The untenable prospect of maintaining a large number of separate digital library infrastructures simultaneously has gradually become obvious to research libraries everywhere.

The requirement for a sustainable and integrated array of systems and process for ingestion, management, and dissemination of digital content is now commonly referred to as the *institutional repository*. This relatively new term is usually understood to refer to a coordinated service program whereby the research library and/or other institutional support operations (such as the campus information technology department) maintain one or more coordinated file servers and a set of associated digital services that enable effective ongoing access to and management of intellectual assets of the university. [2] A list of the most common institutional repository software systems is provided at the end of this article.

1.3 The Law of the Hammer

The most common attempt to create such an integrated infrastructure entails selecting a single software system such as DSpace or CONTENTdm to be maintained by a single campus department and then migrate all previously established content infrastructures into this unified tool. [3] This strategy may be successful ultimately, but is frequently marked by difficulties in the associated migration efforts. It also may be a somewhat inflexible approach as it amounts in some ways to the so-called “Law of the Hammer”: If the only tool you have is a hammer, then everything looks like a nail.

Different kinds of content may not be most effectively managed by means of a single tool. Forcing all of the potential varieties of content that libraries are creating or acquiring into a single monolithic infrastructure may demand Procrustean decisions that impose illogical constraints or unnatural conventions on content types not suited to the tool.

1.4 Web 2.0 Concepts

Since 2004 there has been an increasing shift from centralized “monolithic” web sites toward a re-emphasis on decentered interactions between complementary web sites. This trend has been popularly termed the “Web 2.0” philosophy by Tim O’Reilly. [4] Although some commentators have criticized the Web 2.0 label as nebulous, the phrase is generally understood to emphasize decoupled, flexible, standardized interactions between separate web sites in which end-users are intimately involved in creating content and layers of value-added content re-use services.

The concepts and strategies highlighted by the Web 2.0 movement have been embraced by a growing number of librarians developing digital services. [5] Herbert Van de Sompel and Carl Lagoze have proposed the Open Archive Initiative Object Reuse and Exchange (OAI-ORE) standard as a way of formalizing the scenarios for interaction between content repositories. [6] Most recently, a report by the National Library of Australia recommended a

wholesale shift to a service based architecture. [7] The fundamental transition in thinking about digital library infrastructures represented by these developments suggests that a web services approach may now be the most useful way to approach the creation of an institutional repository, and it was this approach that informed our work.

2. Emory University Institutional Repositories and ETD Program

Emory University is typical of many large research libraries in that it sees the promise of the new digital age, is actively creating many new digital library programs and services, and is seeking effective strategies to best manage these new activities. Like many other university libraries during the last decade, separately managed library units in various Emory libraries created a gallimaufry of digital collections and services, with the earnest intention of improving user access in a rapidly changing context of technological possibilities. Like others, we at Emory are now faced with the arduous task of managing and consolidating an infrastructure that evolved in unpredictable ways over the years. Like others, we have now already had to migrate some of these systems through several generations of operating systems, hardware platforms, and failed, defunct, or otherwise vanished vendor solutions. We have particularly been frustrated by the latter experiences, in which a promised "silver bullet" comprehensive vendor solution either did not work as promised, or was abandoned as a supported product by the vendor. In both situations we had to devote significant effort to either "rescuing" content from failed vendor infrastructures, or creating "bridge" solutions to let two incompatible proprietary systems talk to one another.

While we considered the "Law of the Hammer" approach to consolidation, we were also cautious, noting our difficult previous experiences with monolithic closed solutions. But we also did not want a sprawling chaos of separate silo systems that offer no opportunities for consolidation of systems administration expertise and practices. By 2006, we had reached a juncture in which we felt that we needed to try a new line of attack to this set of problems, as we still lacked a coordinated institutional repository program and urgently believed we needed such an infrastructure. The advent of the Emory ETD program gave us an opportunity to re-think our approach.

2.1 Emory ETD Program

Electronic theses and dissertations repository services is one arena of digital library activity in which Emory has been behind the curve of adoption in other university libraries. Previous campus administrations at Emory had been somewhat technologically neophobic and reluctant to endorse an ETD deposit program. Top-to-bottom changes in university leadership from 2003-2006 (including the president, provost, dean of graduate studies, and university librarian) led to a new receptiveness to change and recognition that an ETD program was long overdue. An internal campus award of strategic funding in 2006 allowed the library to begin implementing an ETD program. The library's Digital Programs and Systems (DP&S) division was charged with the implementation of the program, and (even before the funding award was made) began immediate planning activities with the Graduate School and other campus departments.

While most of the planning entailed discussion of bureaucratic processes and policies for submission of electronic theses and dissertations, the question of technical infrastructure keenly interested the DP&S division. We felt that building the ETD service infrastructure also presented us with a golden opportunity to reconsider the prospects for an institutiuonal repository architecture that would represent a balance between over-centralization and the extreme fragmentation that we were experiencing.

Some initial consideration was given to the prefatory question of whether or not the ETD infrastructure was the right foundation for the institutional repository. Recent research has documented widespread inclusion of ETD services in academic institutional repositories, although the evidence also indicates that this is certainly not a universal practice and that there is great variation in the fundamental understanding of what constitutes an academic institutional repository. [8] Although emerging guides to implementing institutional mention ETD services as one possible function of an institutional repository, there does not seem to be a consensus that institutional repositories either should or must include ETD repositing functions. [9] There is also widespread debate over what software to use for ETD depository infrastructures. [10] But this uncertainty in the published literature was contrary to what we were hearing from many colleagues at other universities, who saw a very close relationship between ETD and IR infrastructures. Indeed, we received the explicit advice from several peer institutions that we should plan these systems hand-in-glove. This also matched up well with planning efforts that had recently taken place at Emory.

The library had conducted an internal planning effort in February of 2006 which included an analysis of strategic questions and functional requirements for posited institutional repository. The recommendations of this internal study included an explicit conclusion that any system developed must be capable of providing ETD services. This internal report informed our decision to use the ETD implementation project as a means of laying the foundation for a larger IR infrastructure.

The questions nevertheless remained about tradeoffs of centralization versus decentralization, standardization versus flexibility, etc. As we entered the latter half of 2006 we had to make an immediate implementation decision on the ETD infrastructure, as the campus had now mandated that we implement a working pilot program in time for Spring graduation in 2007.

2.2 Ambitions for Emory Institutional Repository Infrastructure

As mentioned, our hope was to avoid forcing ourselves into a single monolithic architecture, while simultaneously realizing benefits from a standardized infrastructure and set of system administration practices. As we analyzed the strategic infrastructure questions in early 2006, we were also wrapping up participation in a multi-institutional NSDL project titled OCKHAM. [11] The findings from this project led us to believe that standardization could be achieved in repository architectures without compromising flexibility, primarily by means of Web 2.0 approaches and specifically by focusing on the Web Services Architecture framework of the W3C. [12] The full range of reasons for our belief in this approach are not necessary to recount here, especially as we feel there is a growing consensus in the digital library field on this point and the need for flexible exchange of data via XML. The main point here is that we decided to re-architect our infrastructure *not around any particular software **product**, but rather around an **approach to interoperability**.*

Having said that, we still needed a software tool to implement our ETD service program. Hopefully, whatever tool we selected would allow us to begin moving our institutional repository infrastructure toward the web services approach to standardized interoperability that we were seeking.

2.3 Selection and Benefits of Fedora

To select an IR software package we first consulted the comparative literature. [13] A recent ARL survey reports that the most commonly deployed IR software package is the open source software DSpace, with the DigitalCommons hosted service offered by Proquest the most popular commercial system. [14] We considered the DigitalCommons solution briefly,

but felt that a hosted solution would not give us the flexibility that we were seeking in an institutional repository infrastructure. Similarly, the option of implementing one of the commercial software packages that exist, such as CONTENTdm or Digitool, felt like creating another monolithic silo. We have had the best results previously with popular open source software packages, which provide access to the source code for customization but also access to a large community of other implementers. After considering the available open source options we narrowed our evaluation to three: DSpace, Fedora, and Eprints. DSpace, as mentioned, seems to currently be the most popular selection in ARL libraries. Eprints was arguably the first major IR software package, and is probably the most popular option in European libraries. Fedora is a quite different sort of software, developed by Cornell computer scientists and UVA librarians. [15] Unlike DSpace or Eprints, Fedora provides virtually no front end user interfaces, but primarily offers a sophisticated set of tools for repositing content and an associated model for how to conceive of interacting digital library services. In many ways we felt that the Fedora model was the best match to our infrastructural philosophy, but we did not immediately settle on it before talking to others.

To gain the practical perspectives of others, we interviewed staff at other instituions that had implemented institutional repositories of various kinds, or which had special insights into issues surrounding both the IR and ETD topics. Virginia Tech, as the ETD field leader, was able to give us valuable perspectives on both implementation of an ETD program as well as the context of the NDLTD program. We obtained useful information from our colleagues at Georgia Tech, who had first implemented the ETD-db for their ETD service and then later implemented the DSpace software as a comprehensive solution for ETD and other repository services. The NSDL Core Integration team at Cornell had many insights into Fedora. Finally, the University of Edinburgh had evaluated our top three software choices (DSpace, Eprints, and Fedora), and was able to give us comparative advice on all three products.

After considerable internal discussion, we finally decided on Fedora, much to our own surprise. We had known that Fedora represented the most flexible option, but it had not been a frontrunner because we recognized that it was also the most ambitious option in terms of its technical development requirements. But after carefully considering the development efforts that *any* institutional repository solution would mandate over the long run in terms of integration efforts, we realized that Fedora would be the most effective infrastructure for a coordinated and comprehensive solution to IR needs. Fedora is inherently not so much a single tool as a method of approaching digital library services. Implementing Fedora is not like implementing a monolithic repository system like DSpace in which various functions are built into the software. Rather, it means implementing an abstracted approach to modular digital library components that accomplish functions separately but work together through documented protocols. [16]

An example is the ETD service itself. Fedora provides no functions for managing the ETD submission/ingestion process. A separate software package must be installed to provide that apparatus. We selected the Fez software developed by the University of Queensland Library to adapt for our ETD program. The software provides an entire workflow enabling graduating students to log in and submit their dissertation or thesis, for oversight functions by faculty or graduate school officials to take place, search and browse of submitted items, and other user interface functions. Because Fez, like Fedora, is open source, we were able to easily adapt and troubleshoot the installation for our purposes.

Items reposited in Fedora may be accessed through any number of modular user interface systems. We are thereby able to address currently known needs while leaving the possibilities for other interfaces totally open. We are already anticipating the development of additional interfaces for digital content in the repository. The abstracted architecture enables a much higher degree of flexibility than would otherwise be available.

2.4. Web Services Architecture

The interoperability abstraction that we focused on was the W3C web services framework. By requiring that all systems communicate using XML via web services protocols, we are able to *separate the user interface from the repository proper*. This enables us to further modularize the virtual institutional repository in another way: there may be multiple instances of repository software systems at work. As long as we are able to impose a consistent web services interface on a repository, we can include it seamlessly in the interior of the following architectural schematic, which is generally organized acoording to the OAIS reference model concepts. [17]

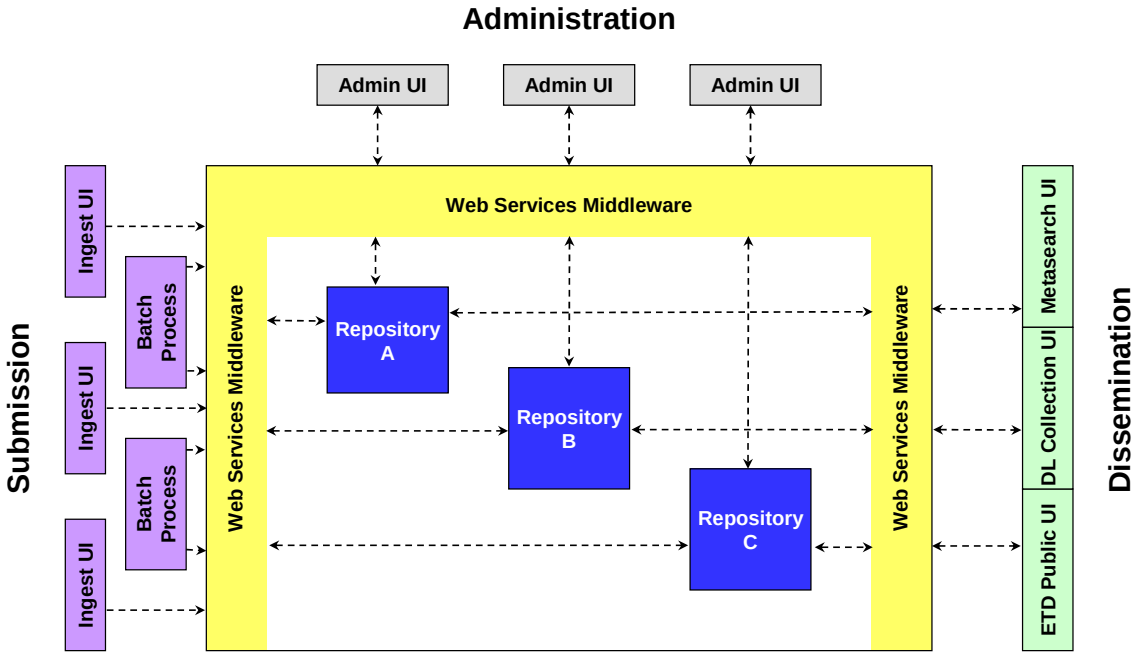

Fig.1 Virtual Repository Architecture based on Web Services Interfaces

Submission may conceptually occur in many ways, through particular ingestion user interfaces or a variety of batch processes. These ingestion processes communication with underlying repository systems through a web services layer. These user interfaces and batch processes may in turn be driven by other standard protocols, such as the OAI-PMH.

Similarly, *dissemination* and *administration* functions occur in separate modules that also communicate with underlying repositories via web services.

It may often be the case that related modules for submission, dissemination, and administration are clustered together in a single clustered software system. Fez is an example of such a system, in which the relevant ETD submission, dissemination, and administration modules are collected together. What keeps this from being a monolithic solution is that any of these functions can be supplemented or replaced by another system that meets the same web services standards.

This components-based approach lends itself to reuse of software. We have realized that we will be able to reuse many of the scripts and routines from the ETD interfaces in other digital collection interfaces. We plan on continuing to accumulate a reusable toolkit of such software in coming months.

2.5 Fedora/Fez Implementation

The process of implementing the new repository architecture and concommittant ETD program began in the summer of 2006 and proceeded rapidly through the winter of 2006-2007. The development team was limited, consisting of 1 FTE programmer with oversight provided by both the team leader for systems development and the ETD implementation project manager. The combined Fedora/Fez system was in operation by April 2007, with several spring graduation dissertations and theses reposited during the pilot project.

The greatest significance of this implementation was not that we were able to accomplish it in minimal time with modest effort. Rather, we felt that we had not simply implemented another system that would have to be replaced at some point, but had successfully laid the foundation for a range of subsequent systems that could leverage many of the same tools and models.

3. User-Centered Process of Service Development Prioritization

Obviously, this approach to systems development is open ended, and could quickly lead to overwhelming our development staff without effective prioritization. The Emory University Libraries are scaled similarly to other university libraries, and have a very modest programming staff available to deploy on ad hoc projects. We have spent a significant amount of time considering processes for prioritizing service development requests based on reported user needs and strategic alignment of systems expansion efforts.

We are consequently careful to document systems development requests, including the source of the request, factors of urgency, scale of effort required, alignment with institutional strategic priorities, technical requirements, and other relevant issues. Requests from various vectors are submitted into a development queue and reviewed on a quarterly basis for approval or deferral. Usability studies of key systems are conducted after implementation, with requested feature changes or redesign work channeled through this same process.

The ETD program implementation was closely guided by feedback from the campus implementation committee. Key stakeholders were (respectively) students for submission interfaces, faculty and librarians for administrative interfaces, students and faculty for dissemination interfaces, and systems staff for the repository functions. By recording and prioritizing all requests, we were able to limit scope creep to a reasonable number of requests.

4. Findings and Implications

Our main findings are that a modular, standards based approach guided by direct user feedback is the best option for systematic development of an institutional repository framework. This strategy provides the following benefits:

- Enables incremental advances to an infrastructure
- Limits vulnerability to being locked into a particular tool
- Maximizes flexibility and capacity for rapid adaptation to changing requirements

We believe that implementing our ETD program using these principles as a conceptual and functional foundation for our institutional repository infrastructure has been a successful strategy; a strategy that we will continue with in coming years.

Common Institutional Repository and ETD Software Systems

CONTENTdm	This commercial software package is frequently applied to digital archives operations, and less frequently to institutional repositories or ETD systems. URL: http://www.dimema.com/
DigitalCommons	This hosted solution by Proquest/UMI is a powerful tool, but requires that an institution depend totally on an external vendor. URL: http://umi.com/products_umi/digitalcommons/
Digitool	This commercial software package (one component of the offerings by Ex Libris for library operations) is frequently applied to digital archives operations, and less frequently to institutional repositories or ETD systems. URL: http://www.exlibrisgroup.com/digitool.htm
DSpace	This open source software is currently the most popular option for institutional repositories in the USA. It has an active developer community in the DSpace Federation. URL: http://www.dspace.org/
Eprints	This was the first open source institutional repository software, and is widely deployed in European institutions. URL: http://www.eprints.org/
ETD-db	The earliest ETD repository software, it is typically not used as a full institutional repository. URL: http://scholar.lib.vt.edu/ETD-db/
Fedora	Arguably the most sophisticated repository software, Fedora is a set of back-end tools for creating web service based institutional repositories. URL: http://www.fedora.info/
Fez	Not itself an IR softeware, Fez is an ETD module for Fedora instances. Official site: http://www.library.uq.edu.au/escholarship/ Development Site: http://sourceforge.net/projects/fez/

References

1. Fox, Edward A.; Eaton, John L.; McMillan, Gail; Kipp, Neill A.; Weiss, Laura; Arce, Emilio; and Guyer, Scott. "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources." D-Lib Magazine, September 1996. http://www.dlib.org/dlib/september96/theses/09fox.html

2. Johnson, Richard K. "Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication." D-Lib Magazine, Vol. 8, No. 11 (November 2002). http://www.dlib.org/dlib/november02/johnson/11johnson.html

3. Lynch, Clifford. A. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." Coalition for Networked Information, ARL Bimonthly Report 226 (February 2003). http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

4. O'Reilly, Tim. "What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software." O'Reilly Network, 09/30/2005. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

5. Cohen, Laura. "Library 2.0: An Academic's Perspective." Blog site. http://liblogs.albany.edu/library20/

6. Open Archives Initiative. *Object Reuse and Exchange Web Site.* http://www.openarchives.org/ore/

7. National Library of Australia. *IT Architecture Project Report.* March 2007. http://www.nla.gov.au/dsp/documents/itag.pdf

8. Van Westrienen, Gerard, and Clifford Lynch. "Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005", D-Lib Magazine, Vol. 11, No. 9 (September 2005). http://www.dlib.org/dlib/september05/westrienen/09westrienen.html

9. Barton, Mary R. and Waters, Margaret M. Creating an Institutional Repository: LEADIRS Workbook. MIT Libraries, 2004. http://www.dspace.org/implement/leadirs.pdf

10. Jones, Richard D. "DSpace vs. ETD-db: Choosing software to manage electronic theses and dissertations." Ariadne Issues 38 (January 2004). http://www.ariadne.ac.uk/issue38/jones/intro.html

11. The OCKHAM Initiative. Website. http://www.ockham.org/

12. Web Services Architecture. W3C Working Group Note 11 February 2004. http://www.w3.org/TR/ws-arch/

13. Open Society Institute, A Guide to Institutional Repository Software, 3rd ed., August 2004. http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_v3.pdf

14. Bailey, Charles. "SPEC Kit 292: Institutional Repositories." ARL SPEC Kit Series, July 2006. Survey Results available at: http://www.arl.org/bm~doc/spec292web.pdf

15. Staples, Thornton; Wayland, Ross; and Payette, Sandra. "The Fedora Project: An Open-source Digital Object Repository Management System." D-Lib Magazine, Vol. 9 No. 4 (April 2003). http://www.dlib.org/dlib/april03/staples/04staples.html

16. Lagoze, Carl; Krafft, Dean B.; Payette, Sandy; and Jesuroga, Susan. "What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL." D-Lib Magazine, Vol. 11, No. 11 (November 2005). http://www.dlib.org/dlib/november05/lagoze/11lagoze.html

17. Consultative Committee for Space Data Systems (CCSDS). *Reference Model for an Open Archival Information System (OAIS).* CCSDS 650.0-B-1, Blue Book, January 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf