Winter 12-9-2013

UC Santa Barbara's Alexandria Digital Research Library Repository Project: Building a Better Repository by Catering to User Needs

Eric Mulhaupt
San Jose State University, emulhaupt@msn.com

UC Santa Barbara's Alexandria Digital Research Library Repository Project:

Building a Better Repository by Catering to User Needs

Eric Mulhaupt

Mary Bolin

LIBR 281 Section 11

Assignment 3

December 9, 2013

Introduction

The kickoff meeting for UC Santa Barbara's new Alexandria Digital Research Library (ADRL) Repository Project was October 16, 2013. Key players in the project have met a handful of times to discuss design and implementation plans. The project is in its early stages. IT staff, metadata librarians, and catalogers are working together to conceptualize and explore options for important procedure, tools, and technologies that will impact the effectiveness of the repository. This process is summarized, documented, and updated on the Library's staff wiki. The Library plans to develop its ADRL collection in multiple phases.

Currently, the Library uses ProQuest, a third party vendor, for ETD ingestion and viewing. The UCSB Graduate Division requires its students to submit PDF copies of their theses and dissertations to ProQuest. Proquest then processes the material and creates non-standard XML files to the Library with descriptive and administrative metadata. Library catalogers transform the XML files into MARC, enhance the record to meet their own standards, then update it to our ILS. The record displays on the OPAC with a direct link to ProQuest's database version of the document. ADRL will allow the Library to circumvent ProQuest's involvement in this process by enabling graduate students to submit their ETDs directly to a UCSB Library-run database.

In the project's first phase of development, the Library intends to retroactively populate ADRL with previously submitted ETDs currently stored with ProQuest. ADRL will ingest these ETDs after receiving a well-formed MARC record, a PDF of the document, and the Proquest XML file. It will then convert the MARC record and some metadata from the XML file into MODS to improve browsability, federated searches, and

harvest-ability of the ETDs. Once ADRL is populated with the ETDs, the Library will remove their ProQuest database links from the OPAC and become UCSB's only source for ETDs. The Library is currently discussing the process and tools needed for ingesting new ETDs directly into ADRL with the campus Graduate Division.

ADRL's metadata schema is closely based on the University of Texas's Texas Digital Library (TDL) ETDs Application Profile. It includes the same set of thirteen mandatory elements and will require encoders to follow the same practices detailed in TDL's Application Profile for each element. However, since ADRL will contain more than just ETDs, the Library's metadata librarian, Chrissy Rissmeyer is currently working on tailoring its element set to encompass all potential document types. Nevertheless, ADRL will maintain better interoperability by following TDL's standard.

Once the first phase is complete, ADRL will ingest enormous collections of digitized items from the Library's Special Collections, Map and Imagery Lab, and Digitization Project departments. Each of these departments has its own set of metadata requirements for ADRL. Much of the Special Collection material must be ingested with incomplete metadata due to types of digitized material they have already digitized. The Map and Imagery Lab intends to include about 200,000 digitized aerial photographs, GIS spatial data, and a collection of 10,000 x 15,000 pixel scanned maps. The original Alexandria Digital Library contents must also be migrated to the new repository. The Library has already digitized a large amount of additional material types it intends to include in ADRL. ADRL will also need to be adaptable to other standards not yet known. Thus, while still a work in progress, ADRL's element set already has 58 unique elements.

ADRL will utilize Hydra as its technical framework. The Hydra Project's website describes the product as an open source "ecosystem of components that lets institutions deploy robust and durable digital repositories (the body) supporting multiple 'heads': fully-featured digital asset management applications and tailored workflows." Hydra caters to ADRL's demands by providing a toolset to construct a repository that is multifaceted, adaptable, and can be specified to fit UCSB Library's unique collection and special needs. These tools include APIs and plugins to create customizable search and discovery interfaces as well as submission interfaces. It is built around open-source Fedora Repository software. The project facilitates "models that can be used for various forms of content" and supports both simple and complex digital objects. The Hydra Project allows multiple institutions to develop their repositories in a collaborative environment, thus contributing to better interoperability and standardization. Hydra is the fundamental core upon which ADRL will function.

While Hydra provides a toolset for ADRL to develop an extensive and user-friendly system, it is up to the Library to gain support among users and content contributors. The future of ADRL relies on a well thought-out plan for continuous community involvement. ADRL must look beyond their plans for the pre-loaded ETDs and digital assets described above and come up with solutions that ensure students, faculty, and other researcher appreciate and understand the services it provides.

Problem

Academic institutional repositories projects thrive on community submissions from faculty, graduate students, and other researchers. However, participation from these groups is not always guaranteed as many do not recognize the benefits of the

technology. While UCSB's Graduate Division already requires its students to submit their theses and dissertation to the Library for publication, there is currently no policy in place that encourages the submission of other research documents. ADRL will fail to reach its potential if it does not maintain appeal and cooperation among these crucial research document generators.

Literature Review

The following literature review seeks to explore the process and tools librarians across the world use to create high-functioning institutional repositories. The articles were selected based on their applicability to the goals, scope, schemes, and constituency of ADRL's educational environment. I wanted to attain a better understanding of the context, concepts, and systems surrounding the project.

Calderón and Ruiz support UCSB Library's efforts to develop ADRL in *The Participation and Web Visibility of University Digital Repositories in the European Context*. The authors argue that universities use "their repositories to make themselves better known by offering open access to a wide variety of the teaching and/or research output of their academic staff" (Calderón and Ruiz, 2013). UCSB will improve its visibility and reputation within the wider academic environment if ADRL proves successful. Therefore, ensuring its success is a worthwhile endeavor.

Zavalina's *Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval* studied collection-level subject metadata in three "aggregations" of cultural heritage digital collection. ADRA's metadata scheme focuses on collection level metadata in addition to digital object metadata. Zavalina argues that users prefered to view

"complete structured collection-level metadata records" because "it provides an added layer of granularity" and a "more consistent application of controlled-vocabulary collection-level subject metadata elements would improve collection search retrieval results and support browse functionality" (Zavalina, 2011). Collection level metadata gives records context within a wider subject area and allows records to be identified through their connection to their collection even when a specific item within the collection does not have a full record. ADRL will improve its usability by instituting a system where collection level metadata is a central priority.

Bashir writes on Grid Computing as one way to improve content-based information retrieval in *Content-based Information Retrieval Techniques Based on Grid Computing: A Review*. He argues that digital content discovery relies on complex codes and powerful computing systems to meet the "just-in-time" demands of today's users. Bashir states that "the rapid growing size of digital collections produces several challenges in the field of IR such as collection's discovery, standardization of interfaces, collection's management, cost optimization, and privacy issues" (Bashir, 2013). It is important for ADRL developers to be mindful of these challenges and properly assess the scope of their repository. ADRL's metadata scheme should balance descriptive, administrative, and rights information that best applies to UCSB's faculty and researchers with a need for interoperability beyond its campus.

Park and Tosaka examine a survey taken by catalogers and metadata professionals on current practices in digital repository metadata creation in *Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability*. The survey results revealed that MARC was the most widely used

metadata schema in 2010, followed by Dublin Core and MODS. Park and Tosaka assert that "the leading criteria in selecting metadata and controlled vocabulary schemata are derived from collection-specific considerations of the type of resources, the nature of the collection, and the needs of primary users and communities" (Park & Tosaka, 2010). While the ADRL metadata scheme will include elements that satisfy local community needs, the goal is to ensure it maintains a high level of interoperability with outside institutions. The authors warn that "while locally created metadata elements accommodate local needs, they may also hinder metadata interoperability across digital repositories and collections when shareable mechanisms are not in place" (Park & Tosaka, 2010). The ADRL committee's decision to select MODS elements as the repository's primary descriptive scheme is forward-thinking. MODS provides a modern, forward-looking solution that is sufficiently granular and also compatible with the UCSB Library's multitude of MARC records.

Bowens promotes the eXtensible Catalog Project at the University of Rochester in *Metadata to Support Next-Generation Library Resource Discovery: Lessons from the eXtensible Catalog, Phase 1*. The project was based on open source applications in order to "facilitate the use of MARC metadata outside an Integrated Library System, to combine MARC metadata with metadata from other sources in a single discovery environment, and to facilitate new functionality" (Bowens, 2012). Similarly, ADRL's metadata scheme seeks to incorporate MARC metadata with metadata from other sources through its use of MODS. In addition, it plans to incorporate its own institution-specific metadata elements that function within the MODS scheme and enable records to maintain interoperability.

Miller's *Metadata for Digital Collections: A how-to-do-it manual* is an all-encompassing guide to developing a high-functioning metadata system. It has practical application and numerous examples that readers may follow to develop efficient and effective metadata for their digital collections. As Miller states, his book "provides a practice-oriented approach to learning about and applying metadata based on the author's many years of practical experience and of teaching both students and working professionals" (Miller, 2011). The main focus of the book is on descriptive metadata created by libraries, archives, historical societies, and museums. It covers the application Dublin Core, XML, MODS, and VRA Core in detail. ADRL developers would benefit hugely by a detailed read of this text. It is possible that they already have because they are closely following the step-by-step process outlined in Miller's book.

More specifically, Miller's book applies to ADRL's metadata endeavors because it gives a detailed overview of the MODS element set. The MODS chapter covers each of the top twenty most commonly used elements and sub-elements with examples. It also details how to map from XML, MARC, and Dublin Core to MODS. Miller makes a notable point in his MODS chapter summary that "the complexity of MODS...makes it generally less interoperable than a simpler, flat scheme like Dublin Core" (Miller, 2011). Our metadata librarian has already come up with a metadata model for the repository with 61 elements and a method for determining which are required and which are optional. Hopefully, ADRL will benefit from MODS granularity and will simultaneously maintain a high level of interoperability.

Sathyanarayana explores the importance and challenges of developing electronic content for institutional repositories in *Collection Development in the E-*

content World: Challenges of Procurement, Access, and Preservation. The current landscape of digital asset procurement is filled with choices and is constantly changing. Collection development for a digital repository can be a daunting task. There are a multitude of content providers that deliver different amounts of material at a range of prices. Some are limited to specific types of material based on subjects while others encompass a variety of material types and subjects. Some put stringent controls on how the usage of their material is regulated and some are more open with their content. Laws dictate the pricing, ownership, and usage rights of digital material. Sathyanarayana asserts that "collection development may have to be just-in-time access to the relevant and required content accessible by the users from anywhere" (Sathyanarayana, 2013). It is important for the UCSB Library to invest in a repository because it facilitates an information environment that caters to the on-demand needs of our university's researchers, faculty, and other stakeholders. Maintaining its own repository will cut down on the costs of licensing agreements with outside electronic content providers.

Wang determines that instant gratification is a central value to modern library patrons and proposes that digital libraries use methods of "co-curation" in *Co-Curation: New Strategies, Roles, Services, and Opportunities for Libraries in the Post-Web Era and the Digital Media Context*. This strategy alleviates significant workload from catalogers and encourages collaboration between the library and its patrons. This model enables users to upload and process their own material through automated and simplified systems. Wang suggests *Readux*, a crowd-based free web application, as a solution for promoting co-creation. One of the goals of Readux is to "expedite delivery

of library collections to users" by addressing the fact that "libraries have more collections than their workforce can process in a timely manner" (Wang, 2013). ADRL plans to use Fedora to facilitate UCSB faculty self-publishing and expedite content processing. Project committee members have already determined procedures for automating the process of uploading, cataloging, and storing electronic theses and dissertations.

Berman and Kesterson-Townes offer some advice to digital content providers on how to market themselves to modern users in *Connecting with the digital customer of the future*. They examine how user value chains can indicate what type of digital services users want and suggest that libraries should make digital content "more social" (Berman & Kesterson-Townes, 2012). This article further supports ADRL's intent to integrate a system that allows UCSB researchers and faculty to upload and share their works themselves.

Gunning's *Metadata Creation at Institutional Repositories* discusses the general trends concerning institutional repositories within the context of today's academic environment. As publishers continue to increase prices on electronic journal subscriptions, academic libraries are forced to explore open access alternatives when developing their collections. Gunning explains that "Institutional repositories play an extremely important role within the open access movement as they are a primary conduit for providing open access to their scholarly content" (Gunning, 2011). ADRL should include as many material types as possible in order to be more versatile and increase its use-value. UCSB will "highlight the quality of [its] intellectual capital, and develop new forms of scholarly communication" (Gunning, 2011). These benefits are

achieved when faculty and researchers actively and frequently participate in the library's repository initiative. Faculty and researchers also benefit from the exposure of their material via digital repositories. Unfortunately, the idea of submitting their material to such repositories is unpopular. Many ADRL committee members fear this same phenomenon will occur at UCSB as well.

According to Gunning, submitting material can sometimes be confusing to people who are unfamiliar with digital repository technology. Some strategies that encourage participation include creating a "liaison system by which a librarian or repository staff member works with an academic department to collect published material and deposit it into the repository" (Gunning, 2011) and creating mandates that force faculty to deposit their publications. Whatever the technique, it is imperative that the repository guide them through the process to ensure that proper metadata is attached to each submitted item. Gunning recommends MODS as an excellent scheme for institutional repositories given its flexible, malleable, and hierarchical properties. Gunning posits that "by breaking down elements into sub-elements and creating a hierarchical record structure, [MODS] takes some of the ambiguity out of broad element categories" (Gunning, 2011). The MODS Guidelines has detailed descriptions and examples of elements and attributes that can simplify record creation. ADRL should look to create more mechanisms for faculty and researchers to feel comfortable uploading and creating usable metadata for their material. UCSB librarians should scan for automatic metadata tools as these technologies develop to further improve user-repository interaction.

Lagoze, Payette, Shin, and Wilper describe a widely-used digital repository framework called Fedora in *Fedora: an architecture for complex objects and their*

relationships. Following the example of other large university repository projects like Stanford's, ADRL plans to incorporate Fedora as an open-source framework to manage and deliver its digital content. Fedora will provide ADRL with an automated system that will "ingest and export digital objects that are encoded in such XML transmission formats" (Lagoze, Payette, Shin, & Wilper, 2006). Fedora will help ADRL store user-generated content and convert it into a more interoperable and future-proof standard format. Fedora will improve ADRL's user interface between faculty and researchers and the complex cataloging record management that goes on behind the scenes. This will contribute to more user-generated content because it makes the process less intimidating.

Research Question

How does a leading institutional repository deal with the ingestion and display of ETDs and other scholarly works so that its collections is frequently used and constantly expanded? How does it encourage robust metadata contributions from uploaders and improved searching for its patrons?

Methodology

I examine the user interface website for Hydra-based University of Virginia's (UVA) Libra repository. This repository has ETD and Open Access document submission components. I focus on the web interfaces by which their users search for, find, and deposit ETDs. I connect these interfaces with possible options for ADRL to improve submission and exploration for ETDs and other types of research material. I intend to uncover techniques and tools Libra uses to enhance its user experiences

through a qualitative analysis. I was unable to access certain parts of the website when conducting my research so I used screencasts provided by the Hydra Project for a portion of my exploration.

Results: Libra's Document Searching and Ingestion

Libra's homepage (http://libra.virginia.edu) is accessible from UVA's OPAC. It clearly displays a search bar and button for users seeking to add their work. There are links to discover more about Open Access, ETDs, and other new features added to the repository. Along the left side, users can select to search by type of work or department. The general format of the page is aesthetically designed and easy to use. The style persists across all Libra pages. ADRL would benefit from a similarly clear and simple style for its web interface.

The search function offers ways to limit choices by types and campus department. Users may also choose to search by title, author, or other keyword types. Upon searching, users are presented with a generated list of related material. The list is sortable by title, author, publication year, and deposit year. It is expandable to display more items per page. Each item on the list displays a title, author, work type, and year. Once an item has been selected, Libra displays a record based on metadata collected by the document's uploader. The records range in the amount of information they contain. This amount is determined by what the uploader contributes to various optional and required fields when submitting their document. The pertaining document is accessible from a link to its uploaded PDF form in the top left of the record. ADRL should look to also facilitate similarly flexible records, lists, and menus.

Users affiliated with UVA can log in the campus to search documents that are only accessible to them based on privileges settings specified by the author. Logging in is also necessary in order to upload documents. It allows uploaders authenticate their identity and auto-fills certain fields of their submission forms related to their UVA campus directory or Student Information System profiles. The campus-wide NetBadge login service provides linked data about the user to Libra. ADRL needs to include the same type of campus-linked login submission feature as it would make the uploader's experience easier and more straightforward.

There are three types of submission forms. One is for graduate students to upload their ETDs. The second allows researchers to upload Open Access works including articles, books, preprints, conference papers, and books. The third is for campus researchers to upload datasets. The form clearly distinguishes required and optional fields. There are different requirements and options for each type of form. Uploaders who fill more optional fields out contribute to fuller records, better searching, and better data harvesting. The form directly correlates to the repository's defined descriptive and administrative metadata elements. Every field on the form has a rollover button to provide tips on how to fill it out. ADRL should design their submission forms after Libra's to optimize user experiences.

Discussion

Libra has an exemplary modern repository interface. It is easy to use, helpful, and somewhat automated. It improves document submission and discovery quality by encouraging better metadata recording and database navigation. In turn, user satisfaction improves because searches yield better results. Libra's collaboration with

the Student Information System and campus enables linked data to automatically populate necessary and helpful metadata. These features are made possible by Libra's use of the Hydra framework. ADRL will likewise benefit from its own Hydra-based implementation. However, the success of a repository project within a campus community is not entirely based on the quality of its user interface.

Libra's relationship with its community is progressive and should be emulated by ADRL. UVA's Faculty Senate passed a resolution encouraging its members to make their works openly available in the repository. Similarly, ADRL should lobby its own researcher groups at UCSB beyond the Graduate Division to gather support for continuous content submissions. ADRL should focus a portion of its attention on campus outreach programs that advertise the usefulness of a local Open Access repository and the utility and recognition it can provide them. For instance, UCSB offers a service for its faculty to post personal profiles summarizing and listing their academic work. ADRL could link those profiles directly to uploaded digital works associated with each professor and could update these lists in real time.

Conclusion

ADRL has enormous potential to benefit our campus and the wider University of California community by connecting newly generated research and currently active researchers with a supportive digital library infrastructure under CDL. UCSB Library staff are developing plans, processes, and programs that will lead to an incredibly robust and modern database. It is important that they consider Libra's success as they move forward. Libra's implementation of Hydra tools is a model for ADRL to follow. With more user-friendly tools comes better user-generate metadata.

References

Bashir, M. (2013). Content‐ based Information Retrieval Techniques Based on Grid Computing: A Review. *IETE Technical Review*.

Berman, S. & Kesterson-Townes L. (2012). Connecting with the digital customer of the future. *Strategy & Leadership*, 40(6), 29-35.

Bowen, J. (2008). Metadata to Support Next-Generation Library Resource Discovery: Lessons from the eXtensible Catalog, Phase 1. *Information Technology & Libraries*, *27*(2), 5-19.

Calderón, A., & Ruiz, E. (2013). The Participation and Web Visibility of University Digital Repositories in the European Context. *Comunicar*, *20*(40), 193-200.

Cervone, F. (2013). Learning, adaptation, and digital libraries. *OCLC Systems & Services*, 29(4), 200-203.

DeGeorge, D. (2012). Metadata for Digital Collections Book Review. *College & Research Libraries*, 73(1).

Gunning, T. (2011). Metadata Creation at Institutional Repositories. *PNLA Quarterly*.

Hui, Y. (2012). What is a Digital Object?. *Metaphilosophy*, *43*(4), 380-395.

Hydra: Get ahead on your repository. Website: http://projecthydra.org.

Libra: Online Archive of university of Virginia Scholarship.
Website: http://libra.virginia.edu

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. *International Journal On Digital Libraries*, *6*(2), 124-138.

Libra screencast: Accessing content. Website: http://screencast.com/t/23dphaR4pYb.

Libra screencast: Uploading ETDs. Website: http://screencast.com/t/Aw1ZaNsa.

Libra screencast: Uploading datasets. Website: http://screencast.com/t/zibhn20Tll7b.

Mapulanga, P. (2013). Digitising library resources and building digital repositories in the University of Malawi Libraries. *The Electronic Library*, 31(5), 635-647.

Miller, S. (2011). Metadata for Digital Collections: A how-to-do-it manual.

Park, J. & Tosaka, Y. (2010). Metadata Creation Practices in Digital Repositories and Collections: Schemata, Selection Criteria, and Interoperability. *Information Technology & Libraries*.

Rissmeyer, C. In-person Interview.

Rushing, A. (2008). Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations.

Sathyanarayana, N. V. (2013). Collection Development in the E-content World: Challenges of Procurement, Access and Preservation. *DESIDOC Journal of Library & Information Technology*, 33(2), 109-113.

Shreeves, S. (2009). Digital Library Federation / Aquifer Implementation Guidelines for Shareable MODS Record.

Surratt, B., Little, A., Mitchell, A., Thomale, J., and Flannery, M. (2006). Texas Digital Library Application Profile for ETDs.

Wan, G. & Liu, Z. (2008). Content-Based Information Retrieval and Digital Libraries. *Information Technology & Libraries*.

Wang, Z. (2013). Co-Curation: New Strategies, Roles, Services, and Opportunities for Libraries in the Post-Web Era and the Digital Media Context. *International Journal of Libraries & Information Services*.

Zavalina, O. L. (2011). Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval. *Journal Of Library Metadata*, *11*(3/4), 104-128

Zhu, B. (2012). Digital repository: preservation environment and policy implementation. *International Journal on Digital Libraries*.