A Technology Analysis of Repositories and Services:

A Proposal Submitted to the Mellon Foundation

Introduction

The promise and potential of repositories to house content and facilitate services such as digital preservation, e-learning and scholarly communication have been documented in various forums. Many organizations, including both universities and companies, have developed repositories that emphasize different services or objectives.   Some well-known examples include Fedora, DSpace, EPprints and ContentDM. Institutions also adopt systems for specific needs; for example, the News and Information Office at Johns Hopkins led an effort to adopt WebWare, a commercial digital asset management system, for public relations digital images. These software systems support various services such as content management, digital preservation, e-learning, or e-publishing.

These systems reflect a diversity of perspectives and functionalities that result in a fuller, richer range of development efforts. However, as institutions adopt repositories to support the delivery of services, the array of choices can become overwhelming. Providing interoperability among systems also becomes increasingly challenging. This proliferation of systems should be balanced with an objective evaluation that will illuminate relative strengths and areas of overlap.

The current landscape bears a strong resemblance to the courseware universe a few years ago. Many institutions began developing courseware systems in response to rising demand for e-learning services. Some of these courseware systems disappeared; others remained in local environments while a few were adopted on a wider scale. WebCT moved from an academic setting into a for-profit company; Blackboard represents a venture that began as a corporate entity. The University of Michigan, Indiana, Stanford and MIT developed open-source courseware systems at their local institutions, but with consideration for the others' work. Eventually, these institutions launched the Sakai Project, an effort to identify and integrate the best tools or modules from each of initiatives at these four institutions. Sakai's goal emphasizes a framework that facilitates the use of tools and modules from various systems as necessary to support users. Today, Sakai represents a multi-million dollar, multi-institution, large-scale initiative, but it began with a smaller-scale architecture and technology research phase that helped develop a roadmap.

The Digital Knowledge Center (DKC) at Johns Hopkins University, working with the University of Virginia (UVA), the Massachusetts Institute of Technology (MIT) and the Sheridan Libraries' network of international partners, proposes an architecture and technology research evaluation of repository software and services such as e-publishing, e-learning, and digital preservation. Each system will be evaluated against a series of use cases. The result will be a set of best practices and recommendations. These efforts will inform the current development of Fedora and DSpace, both of which are expected to be intensive over the next one to three years, result in a typology of repositories and repository users, and would allow us to begin, as part of this project, planning for an interface layer that would facilitate the integration of modules from various applications. Perhaps most importantly, this effort will create a greater understanding of the relative merits of these systems and provide a roadmap for enhancing interoperability among their services. While a Sakai-like effort may ultimately be worthwhile, major (and current) development efforts with both Fedora and DSpace argue for an initial analysis, evaluation, and planning phase in conjunction with both UVA and MIT.

Gap Analysis and Use cases

The Sakai core team institutions used a gap analysis of their courseware systems as a starting point for their functionality roadmap. The DKC has been conducting a preliminary gap analysis of various systems through its existing projects and activities. For example, we ingested the same content into both EPrints and DSpace to compare their relative merits; we continue to evaluate various options for publishing electronic theses and dissertations.

These initial evaluations provide sufficient context and experience to choose an appropriate set of systems for greater exploration through a series of use cases. These use cases will provide a realistic, comprehensive set of experiments using the repository software and services systems. It should be noted that while the use case scenarios are proposed cognizant of realistic policies, our evaluation would be focused on the technological aspects. Enclosed below are representative use cases that represent an initial set for consideration:

1. Self-archiving: A researcher wants a place to put working papers and similar documents, some of which would be available to the public, some of which would not. She will provide links to the content from her web site. The content will also be available via OAI harvesting.

2. Species image repository: An organization wishes to provide a service with which field biologists can upload and make searchable images of various plant and animal species. Included in the field data will be georeferencing information. Information from this repository will be harvested to support a georeferenced species finder service.

3. Library digital collections: The Sheridan Libraries at Johns Hopkins University wishes to ingest a large amount of sheet music (metadata and associated digital images) into a digital repository. The metadata uses a sheet music-specific format, which can be transformed to oai_dc for OAI harvesting, but should be directly accessible for OAI harvesting using a different metadataPrefix. It should be possible to see thumbnails and some metadata on a summary search results page and to view the complete sheet music on an 800x600 monitor.

4. Learning objects: A university runs a learning management system. From that system, the University wishes to link to or access content in the repositories in scenarios (1), (2), and (3). Additionally, the University would like to store the learning objects created in the LMS into another repository, which also contains content to which it would like to link.

5. E-learning: An instructor wants to store problems for a web based homework system in a repository. The problems themselves are expressed in a markup language that requires external applications to render. The instructor would like to be able to efficiently search for problems in order to create problem sets, and to be able to have the problems render appropriately when delivered to students' browsers.

6. E-portfolio: A university has a policy that requires that students retain meaningful control over work that they produce for courses. The student wants to grant access privileges to various entities for some material she has created. These privileges may have different expiration times for the different entities.

7. Publishing: a scenario similar to the self-archiving example from (1), except that there is a review (peer or otherwise) process introduced before the content enters the repository.

8. Publishing: Project Muse wishes to examine its workflow and functionality with a diversity of submission processes from authors, editors and journals.

9. Corporate: A company has a mixture of papers, notes, and other information stored in various folders and websites. The company wishes to better manage their content by depositing it into a repository.

10. Repository management: Operators of the repositories listed in these use cases need to be able to manage these facilities and the content. To do so, they will need to undertake activities such as integration with external services, format migrations, replication and/or backup, and inventory. For example, a repository manager may need to identify portions of an archive that might be at-risk, perhaps because a commercial entity makes an intellectual property claim regarding a file format. The repository manager then needs to develop tools to deal with a large-scale format migration. Repositories need to provide reporting facilities and interfaces that will support these activities.

11. Distributed file organization: A researcher uses a personal information manager and P2P applications to manage her individual and academic files, email messages, calendar notices, research files, etc. and wishes to consider long-term archiving of these materials through repositories.

Kulak and Guiney[1] distinguish among requirements, use cases and scenarios. Requirements, which are aspects that “a computer application must do for its users,” can be reduced in volume by removing conflicts, redundancy, and design assumptions. Use cases are a tool that should show the “what” of the interactions between the users and the computer system. About use cases, the authors state, “it is not only possible but also extremely wise to keep the number of use cases very small.” In producing a small number of use cases for functionality, the analysts and the users are forced to abstract the activities of the system until they truly represent what the system must accomplish. Scenarios are “individual instances of use cases that transverse a specific path using specific data.”

[1] Kulak, D. and Guiney, E. (2000). Use Cases: Requirements in Context. New York: ACM Press.

Carroll and Rosson[2] describe both empirical and analytic approaches to generating use cases. Carroll identifies several techniques such as ethnographic field study, participatory design, reuse of prior analyses, typologies, theories of design and human activity, technology, and transformations (i.e., brainstorming variants of other scenarios). Most importantly, these researchers stress the importance of avoiding bias in use cases and weeding out scenarios.[3] We will work closely with both UVA and MIT to refine and finalize the use cases, and move toward scenarios that will highlight strengths and areas of overlap for systems. The project plan includes a visit to both institutions early in this effort to better understand the development plans of both Fedora and DSpace, and to help build consensus regarding appropriate use cases and scenarios.

With the use cases and scenarios as the guide, the DKC will test and evaluate multiple repository software and services systems. To date, we have evaluated in a preliminary manner the following systems:

- DSpace
- Fedora
- Greenstone
- EPrints
- WebWare
- LOCKSS
- DiVA
- Virginia Tech electronic theses and dissertations software
- Sakai
- WebCT
- WebWork
- Internet Scout
- Open Source Portfolio Initiative
- uPortal

This proposed set of experiments will highlight relative strengths of various systems in the context of realistic, precise specifications. The findings from these explorations will provide invaluable feedback to developers. An intermediate outcome of the proposed analysis would be a typology of repositories and repository users, which would help improve understanding of these topics. Such a systematic understanding of repositories and types of users would facilitate the planning for an interface layer to facilitate the use of modules from various systems.

[2] Carroll, J.M., and Rosson, M.B. (1996). Getting around the task-artifact cycle: How to make claims and designs by scenario. In Rudisill, M. et al. (Eds.), Human-Computer Interface Design: Success stories, emerging methods, and real-world context (pp.229-268). San Francisco: Morgan Kaufmann.

[3] Carroll, J.M. (2000). Making Use: Scenario-Based Design of Human-Computer Interactions. Cambridge, MA: MIT Press.

Repository Interfaces Layer

Given the variety of systems listed above, it is clear that some level of integration would be desirable. As the number of systems increases, so will the complexity of creating specific application-to-application interfaces. Creating an interfaces layer—a set of common services—offers the potential to reduce this complexity substantially. This layer would not need to specify the lowest common denominator of all the services supporting a given function, but would provide standard ways of interacting with the services that are available. These services could continue to expose their native interfaces independent of this layer.

In addition to easing the development burden for those who need to interface multiple systems, these definitions would provide benchmarks to support evaluation of various products, a task that is extremely difficult in the current environment. For example, an evaluator might create a checklist that includes the interfaces supporting the required functionality and then use the checklist as a tool to weed out systems do not meet the minimum requirements.

Some initial areas of focus would be ingestion workflow, bulk ingestion, repository access, bulk export, and reporting. These areas represent important activities for managing content over its full lifecycle.

A Network of Partners

Since its inception, the DKC has emphasized a holistic approach to the development of digital libraries. It was founded with a research and development mission focused on the ingestion of and access to content to support the research, learning and dissemination activities of the academy. This research mission has highlighted the importance of considering issues of access and services at the point of ingestion. As an early example, in the mid 1990s, the DKC worked closely with the University of British Columbia as beta testers of WebCT. Through an early and innovative e-learning initiative with Wilda Anderson, Professor of Romance Languages at Johns Hopkins, the DKC developed a WebCT-supported undergraduate seminar course, explored digital workflow for e-reserves and Special Collections, and highlighted the potential of digital libraries for fostering critical thinking. This early effort eventually led to a grant from the Arthur Vining Davis Foundation to establish similar courses in other departments.

The DKC is perhaps the only organization to receive grants through the Mellon Foundation, the Institute of Museum and Library Services (IMLS) National Leadership Grant Program, and three programs within the National Science Foundation—Digital Libraries Initiative, Phase 2 (DLI-2), Information Technology Research (ITR), and the National Science Digital Library (NSDL). We bring the diverse perspectives of these programs to each of our projects.

Most recently, the DKC has initiated efforts related to digital preservation. Not surprisingly, it is also worthwhile to address issues of preservation (as well as access) at the point of ingestion. As part of the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP), the DKC is involved with the Archive Ingestion Handling Test (AIHT), a project to ingest the 9/11 Archive from George Mason, and subsequently transfer it to the AIHT partner institutions. The DKC has been identified as an objective evaluator of multiple systems in this regard.

While the DKC possesses the relevant expertise and experience to conduct a technology evaluation of repository software and services systems, it collaborates with an extensive network of international partners. Within Hopkins, the DKC acts as an adviser to Johns Hopkins University Press, particularly for Project Muse. We are working with Hopkins Press to consider both workflow and functionality for the next incarnation of Muse. The Sheridan Libraries has formal partnership agreements with the Edinburgh University Library and Uppsala University Library. These partnerships comprise a range of collaborative activities, including a consideration of Edinburgh's "Theses Alive!" Project[4], which focuses on DSpace as a platform for electronic theses and dissertations. Our work with Uppsala University includes an evaluation of DiVA[5], their electronic publishing system. As part of its evolving program on digital preservation, the DKC has submitted two proposals to the National Science Foundation[6], one of which includes the Royal Library of the Netherlands, which will provide feedback regarding the Elsevier Archive.

This network of partners will help foster a broader perspective regarding the DKC's evaluation of technology platforms. However, UVA and MIT represent the two most important partners for this particular effort. The DKC has ongoing dialogue with both organizations regarding the issues that would be explored in this project. Both Thornton Staples[7] and MacKenzie Smith[8] have influenced the development of this proposal. Staples mentioned the value of developing a set of use cases, and Smith provided feedback regarding specific use cases and the potential timing for the development of an interface layer. We will work closely with them, particularly to understand the development efforts for both Fedora and DSpace, to refine the use cases, and to plan for a repository interface layer. By doing so, we ensure that this project proceeds in a collaborative manner while retaining an objective evaluator role for the DKC.

Personnel

The project team brings a diverse, and appropriate, set of expertise and perspectives for the proposed activities. Each individual would focus on a unique aspect of the technology evaluation, but collaborate closely with the other project members, consistent with the manner in which the DKC operates. Resumes or curriculum vitae are attached for each of the personnel; the following list provides a brief description of qualifications along with the role within this project:

- Principal Investigator Sayeed Choudhury, Associate Director for Library Digital Programs, will act as administrative head. Choudhury has acted as PI for eight grant-

[4] http://www.thesesalive.ac.uk/

[5] http://www.dlib.org/dlib/november03/muller/11muller.html

[6] http://dkc.jhu.edu/proposals.php

[7] Thornton Staples is the Principal Investigator of the Fedora Project at UVA.

[8] MacKenzie Smith is the Principal Investigator of the DSpace Project at MIT.

funded projects, organized digital library conferences, published papers and provided presentations in various forums.

- Co-Principal Investigator Timothy DiLauro, Digital Library Architect, will act as technical lead and analyze the use cases to develop tests for the systems. DiLauro has acted as co-PI and Senior Personnel for DKC grants, and leads the overall effort to evaluate multiple technology systems.

- Senior Personnel Mark Patton, Programmer/System Administrator, will act as lead programmer and offer bug fixes or new features to Fedora and DSpace. Patton has developed a fast, disk-based search engine, automated metadata tools, and the NSDL service, SCALE.[9] Patton is the lead programmer for AIHT.

- Senior Personnel James Martino, Digital Pedagogy Specialist, will examine and build a reference implementation for e-learning systems. A former Math faculty member at Hopkins, Martino has explored the connections between e-learning systems and repositories, and acts as Hopkins' lead representative to Sakai[10] and the Medbiquitous Consortium.[11] He will focus specifically on connections between Sakai and digital repositories as part of the associated working group in the Sakai Educational Partners Program.

- Senior Personnel Michael Droettboom, Scholarly Communication Specialist, will examine and build a reference implementation for e-publishing systems. Droettboom has examined the utility of Greenstone and EPrints as publishing platforms, working on LADARK[12], a preprint and reprint service for Latin American scholars, which was identified as "best of the bunch" in the June 2004 issue of Library Journal.

- Senior Personnel Teal Anderson, Usability Specialist, will lead the development of use cases. Anderson provides usability expertise to the Libraries and University through a range of activities including focus groups, surveys, and task-based testing.

- Other Personnel Jacquelyn Gourley, Project Manager, will manage the project. Gourley currently manages DKC projects thorough weekly update meetings and various web-based project management tools including a wiki.

Project Plan

[9] http://nils.lib.tufts.edu/scale/
[10] http://www.sakaiproject.org/
[11] http://www.medbiq.org/
[12] http://eprints.ladark.dkc.jhu.edu/

The proposed activity will begin in January 2005 with duration of one year. The project timeline is as follows:

Phase 1: January 2005 – March 2005

- Visits to UVA and MIT to understand development goals for Fedora and DSpace and to obtain feedback regarding use cases and scenarios
- Refinement of use cases and scenarios through focus groups, contextual inquiry, interviews and/or surveys
- Setup of software and systems for evaluation

Phase 2: April 2005 – September 2005

- Experiments with software and systems consistent with use cases and scenarios
- Development of bug fixes, tools and interfaces (internally) within systems

Phase 3: September 2005 – December 2005

- Development of recommendations and best practices
- Planning effort for a repository interface layer with UVA and MIT and other key stakeholders

The project plan and budget includes provisions for two specific meetings with UVA and MIT, initially (phase 1) at their institutions and finally (phase 3) at Hopkins. Furthermore, both PI Choudhury and Co-PI DiLauro meet Staples and Smith at DLF, CNI, and other meetings regularly. We would take advantage of those meeting opportunities, and communicate via email as well.