Services make the repository

Robert Chavez
Tufts University
Digital Collections and Archives
Tufts University
Medford, MA 02155
robert.chavez@tufts.edu

Gregory Crane
Tufts University
Perseus Project
Medford, MA 02155
gregory.crane@tufts.edu

Anne Sauer
Digital Collections and Archives
Tufts University
Medford, MA 02155
anne.sauer@tufts.edu

ABSTRACT

This paper provides an overview of the collaboration between the Perseus Project and the Digital Collection and Archives at Tufts University in moving the collections of the Perseus Project into the DCA's Fedora based repository as well as a listing of potential services necessary to support a successful institutional repository.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval—*digital libraries*

General Terms

Design, Experimentation

Keywords

Digital repositories, Fedora, digital library services

1. INTRODUCTION

Many assume, for better or for worse, that libraries and archives will be able to maintain bits over time, migrating them from one system to another, converting one image format into another, and executing the suite of complex processes necessary to maintain digital objects in well-established formats. While such an assumption does not do justice to the challenges of the task, the preservation of digital objects is a necessary but by no means sufficient condition for successful curation. At the same time, as institutional repositories (IR) have multiplied at academic institutions around the world, those engaged in building and managing the IRs have lamented the poor levels of contributor participation that most repositories have seen.

Launching an IR is not a small endeavor. Addressing the issues of trustworthiness, authenticity, and preservation take a tremendous investment of resources both in terms of money and staff. How can an institution justify such an investment in something that is little used? A successful IR is one that is relevant. It must meet the needs of the community it serves and be flexible enough to change as the community's needs change while preserving the digital assets deposited in it. Preservation and trustworthiness are undoubtedly the foundation of any repository, but a successful repository must do more to make itself, its collections, and its services relevant. Tufts adopted Fedora in 2001 as the software with which to build its institutional repository because Fedora was designed from the start to curate not only data but the services by which users experience that data. The Perseus Digital Library[1] initially recommended that Tufts Library work with Fedora rather than D-Space because of Fedora's service oriented architecture. The Tufts Digital Repository[2] thus represented from the start a partnership between the library infrastructure and academic projects such as Perseus.

2. THE TUFTS DIGITAL REPOSITORY

The Tufts Fedora based repository is developed and administered by the Digital Collections and Archives (DCA) in partnership with the University's Academic Technology group. We adopted Fedora as our core digital depository because it is designed to support a service-oriented architecture, that is, the Fedora digital repository does not stand as a monolithic repository.[1] In our system architecture, Fedora is the core component of a collection of services (both consumer services and provider services) that communicate with each other to achieve any number of coordinated functions or activities. In a service-oriented architecture, a service is simply a function that is well-defined, self-contained, and does not depend on the context or state of other services in the overall architecture. In other words, in the Tufts University repository system, Fedora provides the repository service to other services as well as a standard means of communication to interact or coordinate with other services, repositories, or digital libraries both inside and outside the Tufts repository system.

As we envisioned our IR, we began from the premise that a repository would need to be open, flexible, and able to deal with the wide variety of digital assets created at Tufts, be they electronic records, primary source research materials, faculty publications, or health science datasets. Openness and flexibility was an absolute must because Tufts already had a robust set of active-use digital libraries. We wanted

[1]http://www.perseus.tufts.edu/hopper/
[2]http://dca.tufts.edu/tdr/index.html

our repository to complement these projects by providing a solid foundation for the preservation of digital assets over time and an expanding set of delivery tools to promote access. The partnership between DCA and the Perseus Digital Library is a realization of this vision.

3. THE PERSEUS DIGITAL LIBRARY

The Perseus Digital Library Project has curated a growing set of digital objects over twenty years and through multiple systems. For us, a successful repository would not only maintain well designed digital objects but the behaviors that make these objects not only accessible but useful. Archives and libraries can define their role in the coming years and decades by assuming responsibility for increasingly powerful services built on top of their collections.

Since 1993, Tufts has been the home of the Perseus Digital Library, a growing collection with particular strengths in Greco-Roman antiquity and other cultural heritage areas. The PDL currently serves approximately 15,000,000 pages per month, each page representing, in effect, a dynamically generated report in response to a general query. For example, when we call up line 44 of book 23 in Homer's *Iliad*, we automatically aggregate and customize a view of all the information that we have about this canonical chunk, including multiple translations, editions, commentaries, etc. While we need to preserve the individual digital objects of the collections, dynamic behaviors that automatically tie digital objects together are what best characterize this digital library. Preserving Perseus thus implies not only preserving digital objects but also the services that go with them.

In the spring of 2006, the PDL has begun to move core digital objects, including original photography of museum objects and a library of TEI-compliant, interoperable, open access and open source primary texts into Fedora. Simply migrating these objects into Fedora is a major step forward: scholars have spent 2,000 years preserving the literature of the Greco-Roman world. The 13 million words of carefully transcribed and marked up Greek and Latin source texts are now permanent components of the Tufts Digital Repository. Unrestricted by third party rights agreements and standard in form, these texts will be able to circulate freely and to provide the basis for new editions in the future. Substantial support is at hand to augment this collection and we expect to have all major Latin texts from the classical period available by summer 2007.

4. BUILDING SERVICES FOR IRS

Although there is a current gap between the behaviors in the Fedora Digital Repository System as currently instantiated at Tufts University and those now standard in the PDL, the DCA and the PDL are actively collaborating to ensure that the necessary services became core components of the Tufts Fedora Digital Repository. For Perseus and for many of its users, these services are foundational. Digital repositories that do not incorporate such services may successfully provide long-term, secondary archives but will never, we believe, become front-line service providers. Put another way, institutional repositories that do not provide the more complex services on which users depend but only provide routine data migration services may succeed in their defined goals but have relatively little impact. If digital repositories do not address the demanding services, then they and university libraries with them may find their functions progressively shifting to a collaboration of large scale entities (e.g., the Open Content Alliance/Google Library) and various disciplinary based organizations.

Table 1: Potential Fedora Disseminators for Repositories

Annotation services
Automatic document alignment
Automatic map and timeline generation
Citation linking
Gazetteer lookup
Named entity identification
Text chunking and alignment
Vocabulary lookup

In order for an IR to successfully meet the needs of its various user communities, it should provide not only preservation and access to digital objects but a range of services that make these objects useful. Table 1 provides a list of potential services. Successful repositories will need to feature advanced browsing, searching, and visualization services, such as allowing the browsing of named entities, supporting searching for different types of entities (e.g. personal names, place names, gene names, chemical compounds) and providing visualization services such as the automatic generation of maps and timelines. For example, repositories could provide a gazetteer lookup service. Such a service could allow repository users to determine if a given document contains readily identifiable terms that lend themselves to rapid gazetteer lookup, such as technical terms, multi-word organizational names or place names.

As repositories grow in size, it will also be important for automated systems to help users determine what glossaries, encyclopedias or other resources in the collection will best serve a user reading a particular document. Successful repositories should be able to associate new documents with the most useful resources. For example, a document on local politics in Boston in 1847 should be associated with contemporary directories of the city, biographies of prominent individuals alive at the time, and historical gazetteers from the mid-nineteenth century covering Massachusetts, the United States and the world.

Repositories should also support advanced text analysis features such as annotation services that track and identify passages that comment specifically upon arbitrary subsets of an object. As many digital objects that will eventually be placed in repositories can have complicated structures, repositories should also support the display of sophisticated textual markup such as multiple text chunking schemes that capture overlapping hierarchies. Support should also be provided for scholars who wish to compare different editions of texts, with services that could include version analysis (e.g. if different editions are available the system should allow a user to analyze how much they differ or visualize differences between editions dynamically). Advanced indexing services should also be provided that can more effectively mine the content of repositories. This should include both manually generated indices and the results of automated processes such as named entity analysis. Named entity analysis services should include the ability of a system to generate a list of named entities along with estimates of the confidence

in these identifications, allow third parties to correct named entity identification errors and feed manual corrections back into the system and use these to improve subsequent automatic analysis.

5. CONCLUSION

RLG/NARA's *Checklist for the Certification of Trusted Digital Repositories* identifies many attributes necessary to establish trust, but trust is only one factor in determining a repository's success, albeit an important one. Focusing on infrastructure, policy, and procedure, the checklist does turn to usability in Section C in its discussion of designated communities and particularly C3, Use and Usability. The intention of these criteria is to establish that one component of trust in a repository is that the assets on deposit will be usable upon retrieval, though that usability is defined very broadly. C4 addresses understandability, or the repository's commitment to ensure that content information is understandable by designated communities even when a significant period of time has elapsed and the skills and tools of that community may have changed, through transformation and documentation.

All of these issues are undoubtedly integral to the long term success of any digital repository and repository managers who ignore them do so at their peril. Meeting the criteria will be a significant challenge to most institutions, and it could seem sensible to focus on them exclusively, ignoring other areas of development that could be seen as distractions. However, at Tufts we have recognized that our viability as a repository program depends on forging strong partnerships with our stakeholders and that without their support we will not be able to continue to build our program to the point when we can, in fact, fulfill all of the checklist's requirements.

Partnerships are the foundation of our model. The DCA brings expertise in preservation, policies, and user services. Perseus brings extensive experience in developing innovative tools and pushing the boundaries of digital library technology for humanities collections. Other digital libraries at Tufts provide services supporting teaching in art history, integrated course management and curricular tools for health sciences, and more. These partnerships entail allowances, such as asset level APIs for management and retrieval of assets, open transport protocols, unified content models and asset typing, within the digital repository to facilitate use, reuse, and interoperability with external applications and systems. While we could each forge ahead alone, we can strengthen all of our efforts by partnering together.

From our perspective, creating a successful repository will mean building one that is more than trusted; it must be relevant, flexible, and able to meet our diverse stakeholders' needs. For the digital repository we are building at Tufts, we believe that this marriage of preservation and services, data and tools, permanence and flexibility will make our repository not only viable, but vibrant, long into the future.

6. ADDITIONAL AUTHORS

Additional authors: Alison Jones, Adrian Packel, and Gabriel Weaver (The Perseus Project).

7. REFERENCES

[1] C. Lagoze, S. Payette, E. Shin, and C. Wilper. Fedora: an architecture for complex objects and their relationships. *Int. J. Digit. Libr.*, 6(2):124–138, 2006.