From Harvesting to Cultivating: Transformation of a Web Collecting System into a Robust Curation Environment

Christopher A. Lee, Richard Marciano, Chien-yi Hou, Chirag Shah
School of Information and Library Science
University of North Carolina
Chapel Hill, NC 27599-3360
{callee, marciano, chienyi, chirags}@email.unc.edu

Categories and Subject Descriptors

C.2.4 [**Distributed Systems**] Client/server, Distributed applications, D.2.12 [**Interoperability**] Data mapping, Distributed objects, D.4.3 [**File Systems Management**] Distributed file systems

General Terms

Algorithms, Management, Documentation, Design, Verification

Keywords

Interoperable Repositories; Collection Lifecycle

ABSTRACT

Much has been written about the lifecycle of digital objects. This study is instead concerned with the lifecycle of collections and associated services. Online collection environments are built to fulfill specific collecting objectives and constraints. If a collection proves useful within its original hosting environment, it will often be necessary or desirable to move the collection to new environments, in order to support new forms of use and re-aggregation or extract resources from legacy data environments. Such a transformation can be extremely expensive, challenging and prone to error, especially if the collections include complex internal structures and services. When “services make the repository” [1], moving raw data from one location to another will often not be sufficient. Digital curators can pre-empt costly and problematic system migration efforts by integrating collections into environments specifically designed to support long-term preservation, scalability and interoperability [2]. We report on an integration of content and functionality of a feature-rich collecting environment (*ContextMiner*) into a robust data curation environment (iRODS).

ContextMiner is a web-based service for building collections, through the execution and management of “campaigns” (i.e. sets of associated queries and parameters to harvest content over time). As a part of the VidArch project, we have been using the *ContextMiner* framework and services for harvesting YouTube videos and associated contextual information on a variety of topics. In July 2008, we released a public beta of *ContextMiner*, allowing anyone to run similar crawls. There are now more than 100 users. The current implementation – based on a single MySQL database and associated code – has served its intended purposes very well, but it is not a scalable or sustainable basis for offering wide-scale collecting services in support of the diverse array of potential users and use cases.

iRODS (integrated Rule-Oriented Data System), is adaptive policy-driven data grid middleware, which addresses aspects of growth, evolution, openness, and closure – fundamental requirements for digital preservation [3]. iRODS currently scales to hundreds of millions of files, tens of thousands of users, and petabytes of data. It operates in a highly distributed environment with heterogeneous storage resources and allows for growth through federation. It supports evolution through the virtualization of the underlying technology and supports changing business requirements through customization of repository behaviors. It supports openness through a data type agnostic treatment of content. iRODS can be instrumented with policies that support the management of the lifecycle of digital assets and will serve as a unique platform to study repository integration. One key feature is the automation of policy enforcement across distributed data that have been organized into a shared collection. The coupling of other open repositories and iRODS can create greater efficiencies and new types of repository services.

We discuss various repository integration scenarios, their potential benefits, and implications for collection life cycles. The approaches co-locate metadata and content in varied ways and rely on efficiencies found in one repository only, or on the ability to combine policies in both spaces: (1) iRODS to *ContexMiner* data migration, (2) Policy-based data management for *ContextMiner* collections, and (3) Policy interchange between *ContextMiner* and iRODS collections.

Copyright is held by the author/owner(s).
JCDL’09, June 15–19, 2009, Austin, Texas, USA.
ACM 978-1-60558-322-8/09/06.

REFERENCES

[1] Aschenbrenner, A., et al. 2008. The Future of Repositories? Patterns for (Cross-)Repository Architectures. D-Lib Magazine 14, 11/12.

[2] Chavez, R., Crane, G., Sauer, A., Babeu, A., Packel, A., and Weaver, G. 2007. Services Make the Repository. Journal of Digital Information 8, 2.

[3] Thibodeau, K. 2008. Architectural Issues in Preservation. Sun Preservation and Archiving Special Interest Group meeting. (Baltimore, November 20, 2008).