Sharing Data-Rich Research through Repository Layering

Stephen Abrams
University of California Curation Center

Noah Wittman
University of California, Berkeley

Angela Rizk-Jackson
Julia Kochi
University of California, San Francisco

Abstract

Information technology and resources have become thoroughly integral and indispensable to the contemporary conduct of science, as they have in culture, commerce, entertainment, and education. These technological advances have nurtured the development of a new paradigm of data-intensive science [4]. However, far too much scientific activity still takes place in silos, to the detriment of open scientific inquiry and advancement. Data-intensive science is facilitated through the more universal adoption of good data management practices ensuring the ongoing viability and usability of all legitimate research outputs, including data, and the encouragement of data publication and sharing for reuse.

The centerpiece of such data sharing is the digital repository, which acts as the foundation for surrounding value-added services supporting and promoting effective publication, discovery, and dissemination of research data. Any general-purpose repository will not, however, be able to support the same level of specialized user experience that can be provided by disciplinary portals and discovery mechanisms. Thus, a layered model built on a stable repository core is an appropriate division of labor, taking best advantage of the relative strengths of the concerned systems.

The Merritt repository, operated by the University of California Curation Center (UC3) at the California Digital Library (CDL), functions as a foundation repository for several data sharing initiatives. Merritt has been successfully integrated into a number of external data grids and discovery services, including CDL's eScholarship open access publishing platform, the DataONE network, and the Open Context archaeological data portal. This paper will focus on two recent examples of external integration for purposes of research data sharing: DataShare, an open portal for biomedical data at UC, San Francisco; and Research Hub, an Alfresco-based content management system at UC, Berkeley. These efforts showcase the catalyzing effect that coupled integration of curation repositories and well-known public disciplinary search environments can have on research data sharing and scientific advancement.

1 Merritt

The Merritt curation repository was developed by the UC3 as a comprehensive platform for both long-term preservation of, and access to, digital resources. Merritt is based on a micro-services architecture in which repository function is decomposed into a growing granular set of small RESTful services that can be combined in strategically useful ways [1]. The flexibility inherent to this approach particularly facilitates the integration of Merritt with external systems and services in a variety of ways. Merritt supports geographic storage replication, data integrity, audit, versioning, newly implemented support for asynchronous delivery of GB-scale data objects, and "model free" content eligibility, that is, there are no prescriptive technical requirements on data or metadata.

Effective data publication is predicated on stable citation [4]. Merritt provides persistent identifiers, including DOIs, via integration with the UC3 EZID service, and through it, DataCite. One factor impeding wider data sharing is a concern held by many researchers about the potential loss of control over their data. All Merritt disseminations are mediated through curatorially-assigned access control rules. Additionally, specific terms of use can be codified into data use agreements (DUAs), click-through acceptance of which is then an enforced condition of access.

Merritt's robust storage and stable identifiers create a solid preservation base upon which systems offering higher-level data sharing features can rely. Providing the institutional infrastructure to enable the sharing of data constitutes a powerful step towards advancing the culture of open data.

2 DataShare

The goal of the DataShare project is to achieve widespread voluntary sharing of scientific data [4]. Currently, data sharing is not widespread across all disciplines; scientific advancement and society as a whole would benefit if research data were more widely shared and easily discoverable. To facilitate this goal, DataShare enables investigators at the University of California to publish all of their research outputs – tabular data, images, software, etc. – to a public portal where that data can be discovered via browsing or metadata searching by metadata fields. The DataShare project is a collaboration between UC3, UCSF's Clinical and Translational Science Institute (CTSI), and the UCSF Library. An initial pilot project is underway with digital imagery produced by the UCSF Center for Imaging of Neurodegenerative Diseases; however, DataShare will be open to all research data in the future.

The DataShare portal is based on CDL's eXtensible Text Framework (XTF), a Lucene-based open source platform for access to digital content [2]. The portal also incorporates a

schema-directed editing facility for the collection of meaningful descriptive metadata that is automatically included as part of Merritt submission packages. DataShare is synchronized with new data acquisitions in Merritt by subscribing to the collection-specific Atom feed that Merritt offers as a standard feature. Access and download requests are serviced directly from the Merritt repository, contingent on acceptance of the relevant DUA terms of use. Object-level metadata, passed directly as item-level feed properties or indirectly through feed-embedded links, drive XTF's faceted discovery environment so that data consumers can search and browse using terminology meaningful to their scientific domain. The loose coupling through Atom facilitated an easy integration path that required no modification of standard Merritt features.

3 Research Hub

The Research Hub was developed by UC, Berkeley, as a platform for active content management and collaboration [6]. Built upon the standards-based open source Alfresco CMS, the Hub offers an effective environment for organizing, enriching, and preparing research data for analysis and publication. Rather than independently developing a redundant set of services, the Berkeley Research Hub team preferred to the leverage existing UC3 expertise and preservation services. This has allowed the Hub team to focus its attention on the needs of the early stages of the research lifecycle, addressing collaboration and integration needs for in-process research projects. Berkeley researchers can now easily publish data from within Research Hub through its integration with the Merritt repository using the repository's RESTful APIs for asynchronous content submission. All relevant data and metadata are automatically packaged together in proper form and submitted to Merritt. The normal response to submission requests is handled via email notification. However, Merritt has been modified to support a polling mechanism by which external submission clients, such as the Hub, can track the progress of ingest processing.

Most researchers are not schooled in good curation practices and have difficulty in expressing meaningful descriptive metadata and packaging data in repository-compliant form. Research Hub minimizes these impediments to wider data sharing by mediating many of the problematic activities. Its UI permits a user to request the submission of any selected set of local content to the Merritt repository for permanent retention and access. Descriptive metadata (e.g., Dublin Core, DataCite, ORCID, etc.) facilitates data discovery and reuse. The Research Hub supports a wide range of metadata schemas and is highly extensible and customizable. Data submitted to Merritt are assigned persistent ARK or DOI identifiers, which can be used by researchers to cite data with confidence.

Although the integration mechanism is different from that employed with DataShare – at

the API rather than Atom-level – it again can be characterized as a loose coupling, with the advantage of easy interfacing without necessitating any internal development. Furthermore, the Research Hub infrastructure can be made available as a working content repository upon which other applications are developed for research purposes using standards-based APIs.

4 Conclusion

Merritt's micro-services-based architecture provides a number of options for easy integration with diverse external discovery services with specific disciplinary focus on scientific data sharing. By removing many of the barriers faced by researchers interested in data publication, the integrations of Merritt with DataShare and Research Hub exemplify a new service model for cooperative and distributed data sharing. The widespread adoption of such sharing is critical to open scientific inquiry and advancement.

References

[1] Abrams, Stephen, Patricia Cruse, John Kunze, and David Minor (2010), "Curation micro-services: A pipeline metaphor for repositories," *OR 2010, The 5th International Conference on Open Repositories*, Madrid, 6-9 July 2010 <http://biecoll.ub.uni-bielefeld.de/frontdoor.php?source_opus=5081&la=en>.

[2] Hastings, Kirk, Martin Haye, and Lisa Schiff (2008), "Publishing with the CDL's eXtensible Text Framework (XTF)," *ELPUB 2008, 12th International Conference on Electronic Publishing*, Toronto, 25-27 June 2008 <http://elpub.scix.net/cgi-bin/works/Show?_id=432_elpub2008&sort=DEFAULT&search=%22ELPUB%3a2008%22&hits=52>.

[3] Hey, Tony, Stewart Tansley, and Kristin Tolle, eds. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery* <http://research.microsoft.com/en-us/collaboration/fourthparadigm/>.

[4] Kunze, John (2012), "New metaphors: Data papers and data citations," *National Federation of Advanced Information Services Conference*, Philadelphia, 26-28 February 2012 <http://www.slideshare.net/jakkbl/jak-data-metaphorsfeb12-11805770>.

[5] Rizk-Jackson, Angela, Julia Kochi, and Carly Strasser (2013), "The DataShare Project: Collaboration Yields Promising Tool," *CNI Spring Membership Meeting*, San Antonio, 4-5 April 2013.

[6] Jaffe, Rick (2012), "New and improved Research Hub ready for the academic year," *UC Berkeley iNews* <http://inews.berkeley.edu/articles/Oct-Nov2012/Research-Hub>.