IJDC | *General Article*

DataShare: Empowering Researcher Data Curation

Stephen Abrams, Patricia Cruse, Carly Strasser, Perry Willet
California Digital Library

Geoffrey Boushey, Julia Kochi, Megan Laurance, Angela Rizk-Jackson
University of California, San Francisco

Abstract

Researchers are increasingly being asked to ensure that *all* products of research activity – not just traditional publications – are preserved and made widely available for study and reuse as a precondition for publication or grant funding, or to conform to disciplinary best practices. In order to conform to these requirements, scholars need effective, easy-to-use tools and services for the long-term curation of their research data. The DataShare service, developed at the University of California, is being used by researchers to: (1) prepare for curation by reviewing best practice recommendations for the acquisition or creation of digital research data; (2) select datasets using intuitive file browsing and drag-and-drop interfaces; (3) describe their data for enhanced discoverability in terms of the DataCite metadata schema; (4) preserve their data by uploading to a public access collection in the UC3 Merritt curation repository; (5) cite their data in terms of persistent and globally-resolvable DOI identifiers; (6) expose their data through registration with well-known abstracting and indexing services and major internet search engines; (7) control the dissemination of their data through enforceable data use agreements; and (8) discover and retrieve datasets of interest through a faceted search and browse environment. Since the widespread adoption of effective data management practices is highly dependent on ease of use and integration into existing individual, institutional, and disciplinary workflows, the emphasis throughout the design and implementation of DataShare is to provide the highest level of curation service with the lowest possible technical barriers to entry by individual researchers. By enabling intuitive, self-service access to data curation functions, DataShare helps to contribute to more widespread adoption of good data curation practices that are critical to open scientific inquiry, discourse, and advancement.

Received 28 October 2013 | *Accepted* 26 February 2014

Correspondence should be addressed to Stephen Abrams, 415 20th Street, Oakland, CA 94612, US. Email: stephen.abrams@ucop.edu

An earlier version of this paper was presented at the 9th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The integration of information technology and resources into all phases of scientific activity has led to the development of a new paradigm of data-intensive science (Hey, Tansley, & Tolle, 2009). However, this paradigm can only realize its full potential in the context of a scientific culture of widespread data curation, publication, sharing and reuse. Unfortunately, the record to date is not encouraging: far too few datasets are appropriately documented, effectively managed and preserved, or made available for public discovery and retrieval (Tenopir et al., 2011). This is due in part to a lack of awareness among researchers of the manifold benefits of good data curation practices and a dearth of easy-to-use tools for data curation. The first condition is being addressed by increasing calls, recommendations, and mandates for proactive data management and sharing as a matter of institutional policy and disciplinary best practice, and as a precondition for grant funding and publication. For example, the recent directives from the US Office of Science and Technology Policy (OSTP) require all federal agencies to put into place policies and practices to ensure that 'digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze.' (Holden, 2013). By tying grant funding to compliance with beneficial data curation practices, significant bodies of data should become and remain available for sharing and reuse in the immediate future.

The second factor inhibiting data sharing and reuse is being addressed by the development of a new generation of easy-to-use tools and services that facilitate effective data curation by researchers. One such service is DataShare, collaboratively developed by the University of California Curation Center (UC3)[1] at the California Digital Library (CDL), the University of California, San Francisco (UCSF) Library and Center for Knowledge Management[2], and the UCSF Clinical and Translational Science Institute (CTSI)[3]. DataShare provides a common platform with simple, intuitive interfaces for a comprehensive suite of curation functions that helps researchers to document, preserve and publicly share their own data, and to find and retrieve – and thus reuse – data made available by others. By facilitating and encouraging data sharing, DataShare enables the reproducibility and validation of research results that lie at the heart of the scientific process. Thus, DataShare is an important vehicle for promoting open scientific inquiry, discourse and advancement.

Features

The DataShare portal offers researchers a number of important curation functions applicable across the entire data lifecycle, from initial dataset creation or acquisition to consumer retrieval and reuse. All of these functions are made available via simple, intuitive and easy-to-use interfaces.

1 University of California Curation Center: http://www.cdlib.org/uc3
2 UCSF Library and Center for Knowledge Management: http://www.library.ucsf.edu/
3 CTSI at UCSF: http://ctsi.ucsf.edu/

Preparation

Surveys of the University of California research community indicate that many researchers would like to engage in data publication and sharing but are unaware of what specific practices they should follow. A manuscript presenting the survey findings is in preparation for publication. The preparation section of the DataShare portal offers best practice recommendations for effective long-term data curation, covering topics such as intellectual property rights (IPR); versioning; privacy of personally identifiable information (PII), and anonymization of human subject and clinical data; metadata collection; and packaging conventions. This guidance was developed by UC3 curation analysts in conjunction with campus science librarians and other subject area specialists.

Currently, the best practice advice is general in nature, applicable to data across scientific domains. We expect to augment this in the future with domain-specific guidance as such best practices continue to evolve and become codified.

The preparation section is a fundamental feature of the DataShare service, as decisions made either explicitly or, as is all too often the case, implicitly early on in the data lifecycle can have significant downstream effects on long-term curation activities and outcomes. Thus, the importance of ensuring good preparatory decisions cannot be overestimated.

Selection

An important goal of the design and implementation of DataShare was to effectuate often unfamiliar data curation activities through familiar online workflows and actions in order to remove technical impediments to widespread service adoption. User expectations regarding user interface (UI) and user experience (UX) interactions with online services, including specialized scholarly and scientific services, are largely set by well-known commercial, consumer and social service offerings, such as Facebook, Dropbox and Twitter. DataShare supports modern UI/UX design principles that enable simple, intuitive operation. Datasets intended for submission to DataShare can be selected using a traditional file browsing option, or they can be designated via drag-and-drop operation.

Description

Associating a dataset with descriptive and scientific metadata is essential for effective long-term preservation and discovery. DataShare metadata support is currently limited to the DataCite metadata schema[4], although planning is underway to expand coverage to other scientifically meaningful schemas in the future. The DataCite schema provides a means to document the general properties of datasets, independent of scientific domain. This includes descriptive properties, such as title, publication year, data type, data creators, and keywords, as well as scientifically significant abstracts and methodological statements. Citations to publications making use of the data can also be specified. DataShare will ensure that all required metadata fields are supplied and suggest the addition of recommended fields. The decision to require only a minimal set of DataCite metadata elements was based on balancing the twin goals of supporting effective data discovery and minimizing the level of effort of compliance on the part of researchers.

4 DataCite metadata schema repository: http://schema.datacite.org/

Preservation

Datasets submitted to DataShare are hosted in the UC3 Merritt[5] repository for proactive management. Merritt is a general purpose curation repository supporting long-term preservation of and access to digital assets (Abrams et al., 2011). Merritt provides automated, geographically-dispersed storage replication; ongoing data integrity checking; versioning with complete dataset change history; comprehensive metadata catalog; curatorially-defined collections and access control rules for submission, update, metadata viewing, and data downloading; data use agreements (DUAs) for asserting researcher-specified terms of use; asynchronous delivery of GB-scale datasets; and active preservation analysis, planning, and (if necessary) intervention.

Merritt is content agnostic and model free; that is, there are no prescriptive requirements regarding content genre, format, structure or degree of associated metadata. This removes a significant impediment to wider data sharing found in other repository solutions: the need for researchers to comply with often unfamiliar and thus, seemingly arbitrary eligibility requirements. While UC3 encourages researchers to follow best practices to the fullest extent possible, it recognizes that this is often not practicable for various reasons. In these cases researchers can use DataShare to submit any research data in its original form and reap considerable benefits from having that data hosted in a managed preservation and access environment.

Although Merritt supports its own native submission and discovery interfaces, these are intended primarily for use by institutional librarians and archivists through RESTful APIs and customized automated workflows, rather than manual interactions by individual faculty members, research staff or students. The DataShare portal that sits on top of Merritt was designed and optimized by the UCSF Library for use by a non-specialist audience and incorporates a range of features, such as drag-and-drop dataset submission and faceted search and browse, intended to facilitate ease of use by the research community.

While the primary point of discovery and retrieval for DataShare resources is the DataShare portal, these data can also be accessed directly through the Merritt UI. All DataShare resources are currently collocated in a collection designated for anonymous public access.[6]

Citation

One of the key ideas behind the concept of data publication is to try to surround data with a similar intellectual and administrative superstructure that previously has been only applied to the traditional academic literature (Kunze, 2012). This includes providing datasets with citations equivalent to those provided for other forms of publication.

Merritt is integrated with UC3's EZID[7] service for the assignment, management and resolution of persistent identifiers. Through UC3's founding membership in the DataCite consortium[8], EZID provides Merritt-managed datasets with permanent DOIs. These DOIs can be used as the basis for permanent citations to the submitted datasets. The recommended DataCite citation format generated automatically by DataShare includes creator(s), publication date, title, institution, format, and identifier, e.g.,

5 UC3 Merritt: http://www.cdlib.org/uc3/merritt
6 Merritt collection: https://merritt.cdlib.org/m/ucsf_datashare
7 EZID: http://n2t.net/ezid
8 DataCite: http://www.datacite.org/

Weiner, Michael W. (2012): Frontotemporal Lobar Degeneration (FTLD). University of California, San Francisco. Dataset. http://dx.doi.org/doi:10.7272/q62z13fs

Exposure

DataShare DOIs and their associated metadata are automatically registered with DataCite's federated catalog, the DataCite Metadata Store (MDS), for discovery via its aggregated search interface.[9] These same metadata are further harvested from MDS for inclusion in the Ex Libris Primo[10] and Thomson Reuters Data Citation Index (DCI)[11] services for additional, high-level discovery opportunities.

Notification of the availability of new datasets in DataShare is also possible by subscribing to the Atom feed published by the underlying Merritt DataShare collection.[12] This is a standard collection-level feature of Merritt. The feed provides a summary of descriptive metadata and an actionable repository link to the dataset in the repository for each DataShare dataset.

Control

One common concern often expressed by researchers is the potential for loss of control over the dissemination and reuse of their data once it is made publicly available. In response to this concern, all datasets submitted to DataShare can be associated with data use agreements (DUAs) that explicitly state license restrictions or other terms of use.[13] Acceptance of these terms by a data consumer is required before being provided access to the underlying data. A potential consumer must supply identifying information, such as name, email address, and institutional affiliation; this information is provided to the data owner, providing him or her with an understanding of the community of interest and use for the data. This should function as a confidence boosting measure to help assuage researchers' concerns over control of their data, thus removing another significant impediment to more widespread data sharing.

Discovery

The metadata associated with all data submitted to DataShare are automatically harvested through the public Merritt Atom feed to populate the DataShare discovery portal. This portal provides a synoptic view of all UC-contributed data. Its search and browse capabilities may be refined easily through faceting of all associated metadata elements (Figure 1). Detailed information is available for each dataset (Figure 2). Requests to retrieve data are serviced directly from the underlying Merritt repository, possibly dependent on DUA acceptance.

9 DataCite metadata search: http://search.datacite.org/
10 Ex Libris Primo: http://www.exlibrisgroup.com/category/PrimoOverview
11 Thomson Reuters Data Citation Index: http://thomsonreuters.com/data-citation-index/
12 Merritt collection – UCSF DataShare Atom feed: https://merritt.cdlib.org/object/recent.atom?collection=ark:/13030/m5ng4nz1
13 For an example DataShare DUA, see: https://merritt.cdlib.org/dua?file=producer/GSE27255_iReport_stats_QC_results.pdf&object=ark:/b7272/q6td9v7j&version=5

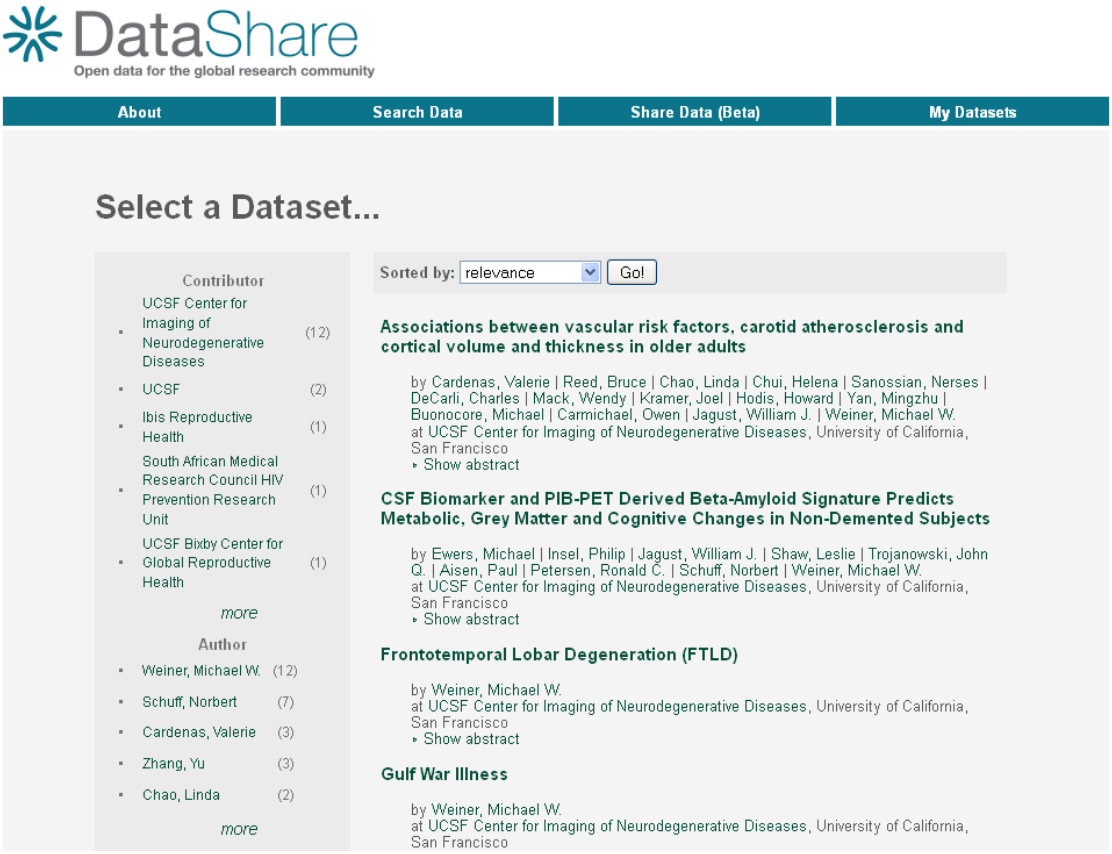

Figure 1. DataShare faceted search and browse interface.

Figure 2. Dataset landing page.

Architecture

DataShare is a public web portal encompassing two primary subsystems – data submission and data discovery – that interoperate seamlessly with UC3's Merritt repository (Figure 3). Merritt is based on a highly distributed micro-services architecture that facilitates robust high availability operation and easy upgrade paths for periodic maintenance and functional enhancement (Abrams et al., 2013).

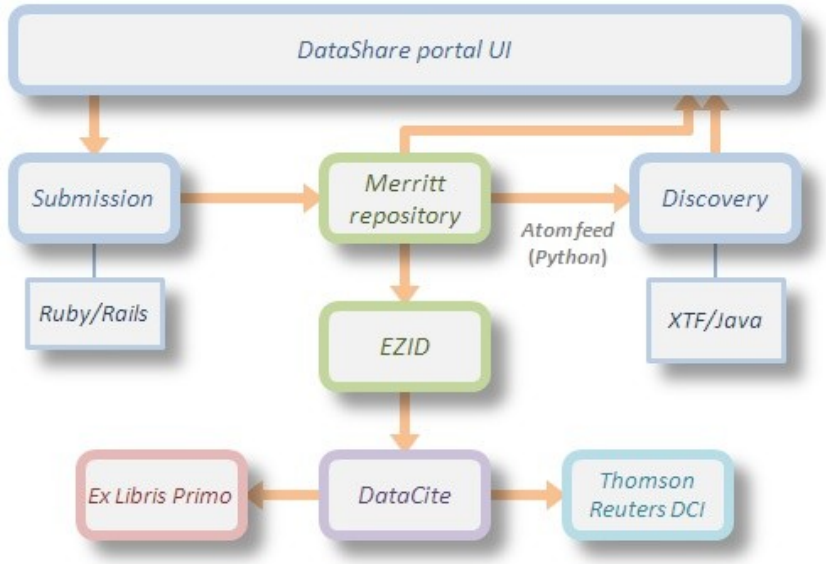

Figure 3. DataShare architecture.

The DataShare portal was designed collaboratively, drawing on the experience and expertise of UC3 in data curation, UCSF CTSI in domain science, and the UCSF Library in online information systems, and was implemented by developers at the UCSF Library. The DataShare submission system is a Ruby on Rails[14] application; the discovery system is based on CDL's XTF[15] (eXtensible Text Framework) platform. Both of these technical systems support extensive opportunities for user interface and experience configuration through customizable stylesheets. The submission system acts as an online wizard, guiding a researcher through the various steps attendant to dataset submission. The XTF search index is automatically populated by subscribing to the Atom feeds published by each Merritt collection and is updated after the successful ingest of each newly submitted dataset.

Merritt relies on CDL's EZID service for managing and resolving its persistent identifiers. EZID provides support for both ARK and DOI identifiers; support for URNs is forthcoming. DOIs are allocated to EZID by the DataCite consortium. Descriptive metadata associated with DOI-identified datasets are automatically passed from Merritt to EZID, where it is harvested by DataCite for inclusion in its aggregated catalog for high-level discovery. This metadata is also harvested by two well-known abstracting and indexing services, Primo and Data Citation Index (DCI), for similar global dataset discovery.

14 Ruby on Rails: http://rubyonrails.org/
15 XTF: http://xtf.cdlib.org/

As can be seen, DataShare operates in a highly distributed ecosystem of independent but highly interoperable components communicating via well-defined public APIs. This division of responsibility across a connected network of services offers significant benefits, facilitating the rapid provisioning of a comprehensive suite of curation functions. Relying on existing service providers and trusted partners for core curation functions obviates the need for costly and time-consuming duplication of effort.

Next Steps

The initial prototype of DataShare was limited to use on the UC San Francisco campus (Rizk-Jackson, Kochi, & Willett, 2013). The primary user community was provided by the UCSF Center for Imaging of Neurodegenerative Disease (CIND). The biomedical imaging data contributed during this period was extremely useful for refining the operation of DataShare, since these data displayed a wide variation in number and size of individual file-level components, ranging upwards to 10,000s of files and 90GB in total size.

Recent campus-wide efforts to promote the release of the latest version of DataShare to the broader UCSF research community have resulted in increased awareness of data sharing best practices, deposit of new datasets, and inclusion of DataShare in Data Management Plans for NIH grant submissions.

Hosting of the DataShare application and control over future maintenance and enhancement of the codebase is being transferred to the UC Curation Center in early 2014, at which time use will be opened to all ten University of California campuses. Although DataShare will be hosted centrally, it is important to ensure that individual campus identity is maintained. Campus branding will be supported through campus-specific URLs, such as http://datashare.berkeley.edu, and extensive opportunities for UI customization. For example, a researcher logging in to DataShare using his or her Berkeley credentials will experience the DataShare UI in a distinctly Berkeley context.

UC3 is working with an external funder to develop a multi-year project of functional enhancement for DataShare. One of the primary enhancements will be the development of a generic version of the service that will be applicable to *any* repository or content management system supporting common submission and syndication protocols, such as SWORD and Atom. This will involve inserting new abstraction layers at the point of repository submission and population of the search index. This index will be based on Solr[16] and Blacklight[17], popular open source search technologies with active development communities.

Conclusion

DataShare is intended to empower research data curation and encourage widespread data publication, sharing and reuse. While the underlying Merritt repository offers native capabilities for content submission, management and discovery, its interface and workflows are more suitable for mediated use by librarians, archivists, and information professionals familiar with repository concepts and operation. DataShare, on the other hand, is intended for use by the University of California research community. By

16 Apache Solr: http://lucene.apache.org/solr
17 Blacklight: http://projectblacklight.org/

incorporating added-value layers with simplified submission and discovery interfaces, DataShare streamlines these processes for direct individual researcher use.

Many researchers are eager to share their data with institutional colleagues, disciplinary collaborators and the wider scholarly community, but don't know how to do so effectively. Research libraries can play an important role in enabling and promoting widespread data sharing through the provisioning and deployment of innovative data curation solutions. DataShare provides a comprehensive platform for data sharing activities, with researcher-friendly interfaces for the acquisition, description, preservation, citation, sharing and reuse of the University's research outputs. By removing many of the barriers currently faced by researchers interested in data curation, the integration of the DataShare portal with the Merritt repository, and through it, EZID, DataCite, Primo and Data Citation Index, exemplifies a new service model for simplified, cooperative and distributed data preservation, publication, citation, sharing, and reuse. The widespread adoption of such practices is critical to open scientific inquiry and advancement.

References

Abrams, S., Cruse, P., Kunze, J., & Minor, D. (2011). Curation micro-services: A pipeline metaphor for repositories. *Journal of Digital Information*. http://journals.tdl.org/jodi/index.php/jodi/article/view/1605

Abrams, S., Rizk-Jackson, A., Kochi, J., & Wittman, N. (2013). Sharing data-rich research through repository layering. Paper presented at the Eighth International Conference on Open Repositories, Charlottetown, PEI, CA. Retrieved from http://or2013.net/sessions/sharing-data-rich-research-through-repository-layering

Hey, T., Tansley, S., & Tolle, K., eds. (2009). *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Holden, J.P. (2013). *Increasing access to the results of federally funded scientific research.* Retrieved from White House, Office of Science and Technology Policy website: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Kunze, J. (2012). *New metaphors: Data papers and data citations.* Paper presented at the National Federation of Advanced Information Services Conference, Philadelphia, PA. Retrieved from http://www.slideshare.net/jakkbl/jak-data-metaphorsfeb12-11805770

Rizk-Jackson, A., Kochi, J., & Willett, P. (2013). *The DataShare project: Collaboration yields promising tool.* Paper presented at the CNI Spring 2013 Membership Meeting, San Antonio, TX. Retrieved from http://www.cni.org/topics/digital-curation/the-datashare-project-collaboration-yields-promising-tool/

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*. doi:10.1371/journal.pone.0021101