JESLIB 2012; 1(2): 109-114
doi:10.7191/jeslib.2012.1014

Journal of eScience Librarianship
putting the pieces together: theory and practice

A Collaborative Framework for Data Management Services: The Experience of the University of California

Joan Starr,[1] Perry Willett,[1] Lisa Federer,[2] Claudia Horning,[2] Mary Linn Bergstrom[3]

[1] California Digital Library, Oakland, CA, USA;
[2] University of California, Los Angeles, Los Angeles, CA, USA;
[3] University of California, San Diego, San Diego, CA, USA

Abstract

The National Science Foundation and other funding agencies now require researchers to include data management plans with new grant proposals. Faced with this requirement, researchers are looking to libraries for help with various aspects of research data management and curation, from creating data management plans to archiving and providing access to their research data. The University of California Libraries deliver a growing range of services and tools such as the DMPTool, EZID, Merritt, Web Archiving Service, and campus-based data management programs. This article discusses these initiatives, tools, and methods for campus engagement and faculty outreach, plus opportunities and challenges in developing library data services.

In January 2011, the National Science Foundation (NSF) introduced a requirement for all new grant proposals: Each proposal must be accompanied by a two-page data management plan describing how the proposal would “conform to NSF policy on the dissemination and sharing of research results” (National Science Foundation 2011). Many researchers have turned to their institutions’ libraries for assistance with meeting this mandate. For libraries, the mandate provides a unique opportunity to extend and evolve their historic charge to preserve their institutions’ scholarly assets by providing services for data stewardship.

At the University of California (UC) the framework used to support this charge is both collaborative in nature and grounded in recognizing that research data management must address the entire data life cycle. Though research is necessarily iterative, this life cycle generally includes five types of activities: Plan, Collect, Manage, Share, and Publish. In the *Planning* phase, the researcher writes proposals and plans projects. In the *Collection* phase, activities include field work and data gathering, as well as studying prior work in the field. In the *Management* phase, the researcher prepares for the long-term curation of data by creating identifiers for data objects and depositing data in a trustworthy repository. The *Sharing* phase includes informal means of making data publically available, such as face-to-face exchanges and socially-mediated interchanges. By contrast, the *Publishing* phase is characterized by more formal means of making data public, thus allowing for citation and attribution, leading

Correspondence to Joan Starr: joan.starr@ucop.edu
Keywords: data management, library services, data curation, collaboration

Table 1: CDL Data Management Life Cycle Services

Life Cycle Stage		Service	Functions
PLAN	Grant application	Data Management Planning (DMP) Tool	Create, edit, share, and save data management plans
COLLECT	Data collection	DataUp	Open source add-in and web application for Microsoft Excel as a data collection tool
		Web Archiving Service	Collect & manage ephemeral web-published content
MANAGE and SHARE	Tracking & management, citation	EZID	Create and manage persistent identifiers
	Storage, management, sharing	Merritt	Curation repository: store, manage, and share research data
PUBLISH	Scholarly publication	eScholarship	Open access scholarly publishing services: papers, journals, books, seminars, and more
	Data publication	Data Publication platform	An infrastructure to publish and get credit for sharing research data

to impact tracking, providing opportunities for getting credit, and other benefits of publication.

Although most researchers have some first-hand experience with the research data life cycle, they generally lack the institutional support that could facilitate one or more of the stages. In response to this gap, the California Digital Library (CDL) has developed a suite of services on behalf of the UC campus libraries. Because CDL is positioned independent of any campus, it can provide common infrastructure and support, as well as enable collaborative solution-seeking. A group within CDL, the UC Curation Center (UC3)[1] has taken the research data life cycle as a veritable service map to create an array of services from which campus libraries may pick and choose. Table 1 lays out the options.

For each service, a CDL Service Manager is responsible for product management, outreach, and marketing. This model has the advantage of providing campus librarians with a service-specific contact person for outreach materials, webinars, training, user sign-ups, and troubleshooting and support.[2] In this way, CDL Service Managers and campus librarians collaborate in providing services for UC researchers.

These research data services exemplify CDL's overall approach to new service provision. In 2010, Former Director of Systemwide Library Planning, Gary Lawrence, explained the CDL value proposition (based on 2008 figures):

> "…with a core budget of about $14 million, the CDL attracts an additional $18.5 million annually in voluntary co-investments from campus libraries and uses the resulting $32.5 million pool of funds to deliver about $52M in direct benefits to campuses, supports an additional $46M in measurable indirect benefits, and provides a technical platform and a leadership capability which fosters development of a host of service innovations that could not readily be supported by our ten campus libraries operating independently" (Lawrence 2010).[3]

[1] For more information about the UC3 staff, see http://www.cdlib.org/services/uc3/about/staff.html.

[2] In practice, CDL has implemented neutral group email addresses in order to provide redundant coverage.

[3] Note that recent service developments and reconfigurations prevent a repetition of this analysis but the ratio of returns is expected to be of similar magnitude.

Campus co-investment has played a significant role in new services, particularly during a fiscally challenging climate. In some cases, the service development process itself provides an opportunity for collaboration. Such distributed development allows the design team to draw on expertise from among many partner institutions and move much more quickly than if development had occurred at a single institution.

Plan: The DMPTool

The DMPTool aids researchers in creating data management plans to meet requirements of the NSF and other funding agencies. While the DMPTool is available for public use, participating institutions gain the ability to customize the DMPTool for their researchers. Once researchers log on using their regular institutional credentials, they can view links to available resources and services, get campus-specific boilerplate text for their DMPs, and find out who can provide more help at their institution. The DMPTool is a simple tool that provides a relatively straightforward service, but it also serves as a gateway to other services that support the rest of the research data life cycle.

The DMPTool development process provides a good example of the service development process as an opportunity for collaboration. Development was entirely self-funded by participating institutions, with developers and other staff at CDL working closely with developers at UCLA and the University of Illinois, sharing the work of design, development, and integration. The code sharing platform Bitbucket allowed developers at various institutions to work on different parts of the software independently and to easily integrate their work, with weekly (and sometimes daily) meetings via web conference for quick updates, testing, and bug reports.[4]

In its first year of operation, the DMPTool has registered over 2,300 users from more than 450 institutions, and users have created over 1,800 data management plans. More than 60 institutions have customized the DMPTool for use by their communities. The original planning team continues to meet regularly and plans to build on feedback from this growing community of adopters to keep moving the project forward. Work also continues at individual institutions to provide further customization to meet the needs of their researchers; for example, a planning committee convened at UCLA to discuss ways to provide UCLA-specific branding and other customization for researchers who log on with their UCLA credentials.

The DMPTool has had significant use among the UC campuses, particularly because UC librarians have been active in conducting outreach to researchers and promoting the tool. At UCLA, Associate University Librarians Sharon Farb and Todd Grappone, and librarians Lisa Federer and Monica Garcia gave a presentation on the DMPTool to the campus research administrators group in early February 2012. Subsequently, new user enrollments increased by 100% from the previous month; more new users registered in February 2012 than the previous four months combined. Promotion of the DMPTool has also created new avenues for the UCLA Library to connect with researchers and increase awareness of the Library's role in supporting the entire research data life cycle.

Collect: DataUp

Partnership with two external grant funders[5] has facilitated the development of the DataUp opensource solution for the data collection phase of the research life cycle. The work addresses the problem of curation and preservation of spreadsheets created using Microsoft Excel, a tool highly used among

[4] The DMPTool code is available free and open source at http://bitbucket.org/dmptool/main/.
[5] The Gordon and Betty Moore Foundation (http://www.moore.org/) and Microsoft Research (http://research.microsoft.com/en-us/).

researchers,[6] but not ideally suited to rigorous research data management. CDL developed requirements for the tool, including features such as standardized column headers, versioning, auto-archiving, and long-term identifier assignment. DataUp will be freely available in September 2012[7] to the general public, regardless of institutional affiliation or existing repository relationships; the only requirement will be a valid email address. After the initial release, the developers aim to engage the community to elicit suggestions for improvements, such as connecting additional repositories, adding metadata schemas, and expanding the set of features to facilitate good data stewardship practices.

Though the tool has not yet been released, outreach at the campus level has already begun, with CDL project manager Carly Strasser visiting UC campuses to gather information about researchers' data practices to inform the design of the tool, as well as to promote the tool as a way for researchers to better manage their data. At UCLA, the tool will be featured as part of an ongoing series of data management workshops offered to researchers beginning in the fall 2012 quarter.

Collect/Manage/Share/Publish: Web Archiving Service

The Web Archiving Service (WAS) provides versatile and powerful tools for collecting, archiving, and publishing ephemeral content from websites. WAS allows curators to collect and manage web-published content to help scholars use the content for private research, as well as facilitating publication of the content for general public access. The archives contain a rich variety of materials, including eScience content, government documents, event captures, and archives for specific research communities, such as unique data sets, collections of sites not otherwise connected, and sites resulting from grant activity.

WAS was built with open source tools, including the Heritrix web crawler and Open Wayback, along with locally developed solutions to address the challenges of capturing web materials. WAS provides tools for analyzing site change over time and allows keyword searching for archived sites. Making archived sites public is optional; as of this writing, over half of the 93 currently active archives are publically available. At the UC campus level, subject specialists and selectors have been encouraged to submit ideas for sites for archiving using WAS, including content of local interest to campuses and materials of importance to the general public.

Manage/Share: EZID

Stable, persistent identifiers are one of the keys to good data management, as they provide consistent tracking for research components. Identifiers also provide the mechanism behind data citation, which in turn powers data sharing and re-use, as well as allowing other researchers to provide credit and attribution for the original researcher (Piwowar et al. 2007). EZID, a tool developed by CDL and available to UC campuses and external institutions, enables easy creation and maintenance of long-term identifiers. The campus co-investment for EZID involves cost-sharing, operating on a cost-recovery basis using annual subscription fees. UC campuses' fees are subsidized.

The UC San Diego (UCSD) Library's Research Data Curation Program uses EZID to provide digital object identifiers (DOIs) for UCSD researchers. In September 2011, the UCSD Library began sponsoring EZID on behalf of the campus, with Mary Linn Bergstrom as the EZID representative. As UCSD's Faculty Liaison, she promotes

[6] For data showing this trend, see Carly Strasser's blog post, "Quantitative Results from the ESA Conference" http://dcxl.cdlib.org/?p=84, accessed July 11, 2012.
[7] For project updates, see the DataUp project blog, http://dcxl.cdlib.org/.

EZID, registers interested researchers, and serves as a contact to connect researchers with CDL. Bergstrom, along with Sue McGuinness, promotes EZID in monthly Data Management Plan workshops for both faculty and librarians. Thus far, UCSD Library's workflow for registering EZID users is straightforward, with registrants tracked in a simple Excel spreadsheet, and users have benefitted from prompt service and a generally positive experience.

Manage/Share: Merritt Repository

The Merritt Repository facilitates preservation and curation of digital assets for the UC community. In its capacity as a data management tool, Merritt can function in several ways: it can be a "dark" or inaccessible archive for important digital assets, serve as a "bright" archive with direct discovery and access, provide a preservation back-end for discovery and content management systems, or integrate with distributed data grids. One example of a project using Merritt is the Datashare project, a collaboration of CDL staff and librarians, developers, and researchers at UC San Francisco (UCSF). Datashare, as the name suggests, encourages researchers to share their data by providing tools that help reduce the barriers to data sharing, including the Merritt-driven repository. In other collaborations, Merritt serves as a member node in the DataONE research grid and provides preservation services for articles published in CDL's eScholarship, UC's open access publishing platform; the Online Archive of California, which provides access to over 20,000 online collection guides; and Calisphere, a collection of digitized primary sources related to California history and culture. Merritt's flexible architecture provides multiple methods for ingest and access to satisfy a wide range of workflows and requirements.

Publish: eScholarship

The publishing component of the research data life cycle is both well-understood and very experimental, a paradox resulting from the emerging nature of data publication. UC3 sponsors a blog on data publication, DataPub, which serves as a medium for exploring developing issues within data publication. CDL's eScholarship open access publishing service provides support for the publication component of the research data life cycle. In collaboration with the UC3 team, eScholarship developers are exploring new initiatives, including support for data papers and data publications that put a familiar "wrapper" (a citable paper) around an unfamiliar object (a dataset), providing researchers with more ways to publish their data and receive credit for sharing it.

CDL participates in other collaborations exploring the possibilities of formal publication of research data, including the recently announced Peer Review for Publication and Accreditation of Research Data in the Earth Sciences (PREPARDE) project, led by researchers at the University of Leicester in collaboration with participation from other universities, libraries, government research centers, and publishers. The project aims to develop procedures and policies for publication of earth sciences-related research data, including designing a workflow and procedures for publishing with CDL.

Future Outlook

Given that researchers continue to expect institutional support to meet funder mandates and standards, data management services will continue to develop and grow in the UC Libraries, utilizing the tools and services offered by CDL as well as homegrown, external vendor, and open-source solutions. Conducting research now requires dealing with big data, or at least big amounts of small data. Even outside of the traditional scientific disciplines, researchers in many fields are beginning to rely on computational research that generates large amounts of data.

The further development of research data life cycle management services at CDL will continue to be a two-way collaboration with the UC campuses. As campuses explore the needs of their research communities, this valuable information can help inform development of new services at CDL. For example, at UCLA, a pilot project led by librarian Lisa Federer partnered public services librarians with cataloging and metadata librarians to explore how the Library could provide support for individual researchers. Claudia Horning and Chamya Kincy of UCLA's Cataloging and Metadata Center responded to a request from a researcher by providing advice on file-naming standards and recommendations for organizing the research team's web pages. Such pilot projects help the UC Libraries explore how to better meet the needs of their research communities and provide feedback to CDL on the utility of existing services and needs that remain unmet.

In the present fiscal climate, the admonishment to take care before launching new services is one that the UC Libraries must adhere to now more than ever. To ensure success, services and projects must be scalable and sustainable; combining the efforts of CDL and multiple UC Libraries gathers the diverse expertise of staff from many different backgrounds and creates economies of scale, and thus achieves a greater likelihood of a service's success. The UC's and CDL's commitment to partnerships and collaboration helps to bring together the right stakeholders in deciding which services go forward, while working to meet the diverse needs of UC researchers.

References

Lawrence, Gary. "The California Digital Library." In *Business Planning for Digital Libraries: International Approaches*, edited by Mel Collier, 219-228. Leuven, Belgium: Leuven University Press, 2010.

"NSF Data Management Plan Requirements," National Science Foundation, accessed May 7, 2012, http://www.nsf.gov/eng/general/dmp.jsp.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing Research Data is Associated With Increased Citation Rate." *PLoS ONE* 2, no. 3 (2007): e308, doi:10.1371/journal.pone.0000308.

Disclosure: The authors report no conflicts of interest.

ISSN 2161-3974