IJDC | *General Article*

Dash: Data Sharing Made Easy at the University of California

Stephen Abrams, John Kratz,
Stephanie Simms, Marisa Strong
and Perry Willett
California Digital Library
University of California

Abstract

Scholars at the ten campuses of the University of California system, like their academic peers elsewhere, increasingly are being asked to ensure that data resulting from their research and teaching activities are subject to effective long-term management, public discovery, and retrieval. The new academic imperative for research data management (RDM) stems from mandates from public and private funding agencies, pre-publication requirements, institutional policies, and evolving norms of scholarly discourse. In order to meet these new obligations, scholars need access to appropriate disciplinary and institutional tools, services, and guidance. When providing help in these areas, it is important that service providers recognize the disparity in scholarly familiarity with data curation concepts and practices. While the UC Curation Center (UC3) at the California Digital Library supports a growing roster of innovative curation services for University use, most were intended originally to meet the needs of institutional information professionals, such as librarians, archivists, and curators. In order to address the new curation concerns of individual scholars, UC3 realized that it needed to deploy new systems and services optimized for stakeholders with widely divergent experiences, expertise, and expectations. This led to the development of Dash, an online data publication service making campus data sharing easy. While Dash gives the appearance of being a full-fledged repository, in actuality it is only a lightweight overlay layer that sits on top of standards-compliant repositories, such as UC3's existing Merritt curation repository. The Dash service offers intuitive, easy-to-use interfaces for dataset submission, description, publication, and discovery. By imposing minimal prescriptive eligibility and submission requirements; automating and hiding the mechanical details of DOI assignment, data packaging, and repository deposit; and featuring a streamlined, self-service user experience that can be integrated easily into scholarly workflows, Dash is an important new service offering with which UC scholars can meet their RDM obligations.

Received 27 January 2016 ~ Accepted 24 February 2016

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

An earlier version of this paper was presented at the 11th 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

Information technology and resources permeate the academic enterprise and are transforming scholarly communication. The careful management of all forms of scholarly output – whether publications, data, software, or workflows – is important to provide public disclosure, promote academic integrity, avoid needless duplication of effort, and enable new, synergistic forms of collaboration and intellectual advancement. The imperative for effective research data management (RDM) is being driven by mandates from public and private funding agencies, pre-publication requirements, institutional policies, and evolving norms of scholarly discourse and practice. In order to meet these new obligations, scholars need access to appropriate disciplinary and institutional tools, services, and guidance. When providing help in these areas, it is important that service providers recognize the variable degree of familiarity that scholars have with data curation concepts and practices. The UC Curation Center (UC3) at the California Digital Library (CDL) supports a growing roster of innovative curation services for use by the University of California (UC) community. Most of these, however, were intended originally to meet the needs of institutional information professionals, such as librarians, archivists, and curators. In order to address the new curation concerns of individual scholars – whether faculty, students, or staff – UC3 realized that it needed to deploy new systems and services optimized for the needs of new stakeholders with widely divergent experiences, expertise, and expectations. This led to the development of Dash, an online data publication service that makes campus data sharing easy.

While Dash gives the appearance of being a full-fledged repository, in actuality it is only a lightweight overlay layer that sits on top of, and freely interoperates with, standards-compliant repositories supporting common protocols for submission and harvesting, such as UC3's existing Merritt curation repository. The open source Dash system provides intuitive, easy-to-use interfaces for dataset submission, description, publication, and discovery. By imposing minimal prescriptive eligibility and submission requirements; automating and hiding the mechanical details of DOI assignment, data packaging, and repository deposit; and featuring a streamlined, self-service user experience that can be integrated easily and unobtrusively into multifarious scholarly workflows, Dash is an important new UC3 service offering with which UC scholars and beyond, can effectively and efficiently meet their RDM obligations.

Research Data Management

The rise of data-intensive scholarship (Hey, Tansley, and Tolle, 2009) highlights the importance of effective RDM as a fundamental component of scholarly work. While organizations such as CODATA,[1] FORCE11,[2] and the Research Data Alliance[3] fulfil important roles regarding education, policy, and advocacy, the responsibility for appropriate management of research data ultimately rests with individual scholars. Unfortunately, recent surveys of scholarly attitudes towards data sharing (Tenopir et al., 2011; Tenopir et al., 2015) indicate that most researchers feel that their home

[1] CODATA: http://www.codata.org/
[2] FORCE11: https://www.force11.org/
[3] Research Data Alliance: https://www.rd-alliance.org/

 doi:10.2218/ijdc.v11i1.408

institutions do not provide adequate support for either short- or long-term management of research data. Thus, the importance of promoting RDM best practices and placing effective tools for effectuating those practices in the hands of scholars cannot be overstated. It is also important to recognize, however, that while RDM is vitally important to the practice and success of scholarly pursuits, many scholars still think of data management as something apart from their primary scholarly focus. While they undoubtedly want positive RDM outcomes, they want them without impinging upon long-standing work patterns and practices. It is therefore incumbent upon data management service providers to tailor their systems and services to meet the explicit and implicit needs of the research community with minimal impact on established workflows.

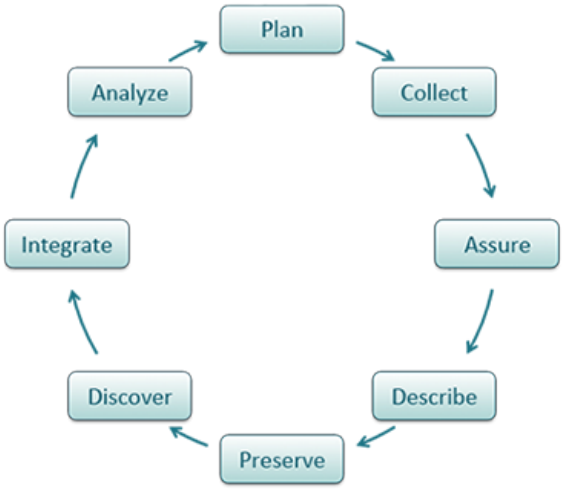

Figure 1. DataONE lifecycle model (Michener and Jones, 2012).

UC3 is a founding partner of the NSF-funded Data Observation Network for Earth (DataONE) project, which is developing a foundation for innovative environmental science through distributed and sustainable cyberinfrastructure.[4] The DataONE lifecycle model (Figure 1) is a useful instrument for structuring the analysis of and planning for RDM solutions. UC3 and its parent organization, the California Digital Library (CDL), have long provided services addressing the Collection through Discovery phases of the lifecycle to the UC community. Historically, the primary users of those services have been campus libraries, archives, and museums; similarly, the collections managed in those services have been primarily composed of cultural heritage material, for example, reformatted or born-digital texts, still images, sound, and moving images. UC3's first response to requests to support individual faculty researchers with RDM needs was to promote the use of existing systems and services. However, it became apparent that this was an inadequate response. While the overall goals of pre-existing solutions were consistent with RDM aspirations, such as ensuring the long-term viability and usability of managed digital assets, the expectations regarding system use varied tremendously between campus information professionals – librarians, archivists, and curators – and individual scholars. The needs of this new user constituency required new, targeted solutions. (UC3 also provides a solution for RDM planning, the DMPTool (Strasser,

[4] DataONE: http://www.dataone.org/

Abrams, and Cruse, 2014b),[5] but this was intended since its inception for use by individual researchers.)

The centerpiece of any effective RDM solution is a repository providing managed datasets with actionable, persistent identifiers; secure, long-term storage; active preservation oversight and intervention; and broad public exposure for discovery and retrieval (Kunze, 2012). The UC3 Merritt repository[6] supports all of these functions and more, but was designed as a general-purpose repository for use by institutional information professionals, rather than a data repository used by individual scholars. Since no general-purpose repository can provide the same level of specialized user experience offered by disciplinary portals, Merritt needed enhancement to improve its utility for RDM purposes. In terms of implementation, UC3 had two choices: directly modify the existing Merritt software to support the new function, or encapsulate this function in an independent, but interoperable service overlaying Merritt. Consistent with its long-standing promotion of micro-services-based infrastructure (Abrams et al., 2011), UC3 chose the latter approach of designing Dash as an independent overlay layer. Augmenting repository function with a surrounding constellation of added-value services minimizes development time, increases deployment flexibility, and facilitates experimentation and innovation.

Dash is an online data publication service for making data sharing easy at UC (Figure 2). It provides streamlined and intuitive self-service submission and search interfaces optimized for use by individual scholars. Since it is an independent system, it can be integrated with any standards-compliant repository, with which it is loosely coupled through standard, community-supported protocols for repository submission and harvesting. Dash is intended to promote beneficial data curation best practices and outcomes with minimal researcher effort.

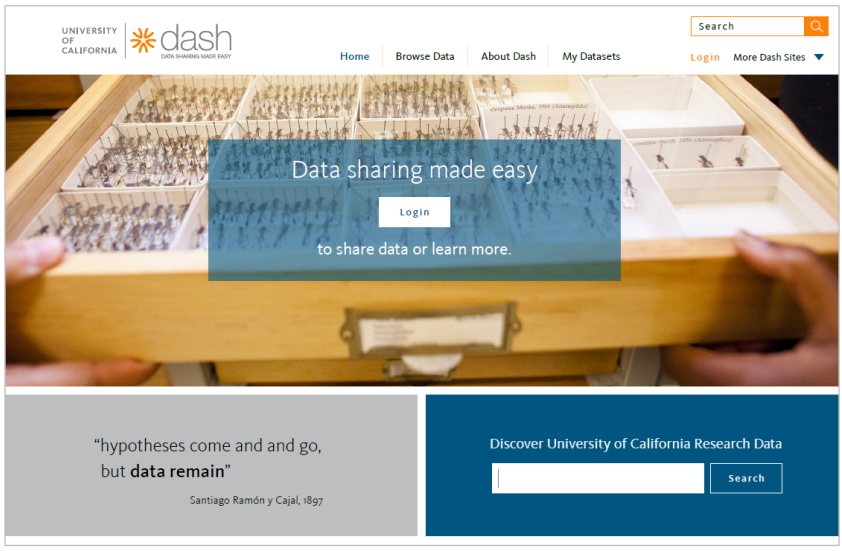

Figure 2. Dash home page.

[5] DMPTool: http://dmptool.org/
[6] UC3 Merritt Repository: http://merritt.cdlib.org/

 doi:10.2218/ijdc.v11i1.408

Dash: Data Sharing Made Easy

Dash addresses six key functions that enable individual scholars to:

1. *Prepare* datasets for curation by reviewing best practice guidance for the creation or acquisition of research data;
2. *Select* data for curation through local file browse or drag-and-drop operation;
3. *Describe* data in terms of intellectually-meaningful metadata;
4. *Identify* datasets for persistent citation, reference, and retrieval;
5. *Preserve*, *manage*, and *share* data in an appropriate data repository;
6. *Discover*, *retrieve*, and (*re*)*use* data through faceted search and browse.

By alleviating many of the barriers that have historically precluded wider adoption of open data principles, Dash empowers individual scholars to assert active managerial control over their research outputs; encourages more widespread data preservation, publication, sharing, and reuse; and promotes open scholarly inquiry and advancement.

The Dash UI features a responsive design that automatically adapts for optimized viewing experience on a range of desktop, laptop, tablet, and mobile computing devices. Drag-and-drop file selection is a familiar behavior supported by many well-known public online services. Support for various metadata schemas is extensible through a plug-in mechanism. By default, the DataCite metadata schema is used.[7] This minimizes the mandatory element set to four: author(s), title, abstract, and data type (Figure 3); plus recommended, but optional, enhanced description in terms of keywords, methods, usage notes, funder(s), related links, and geospatial location (Figure 4).

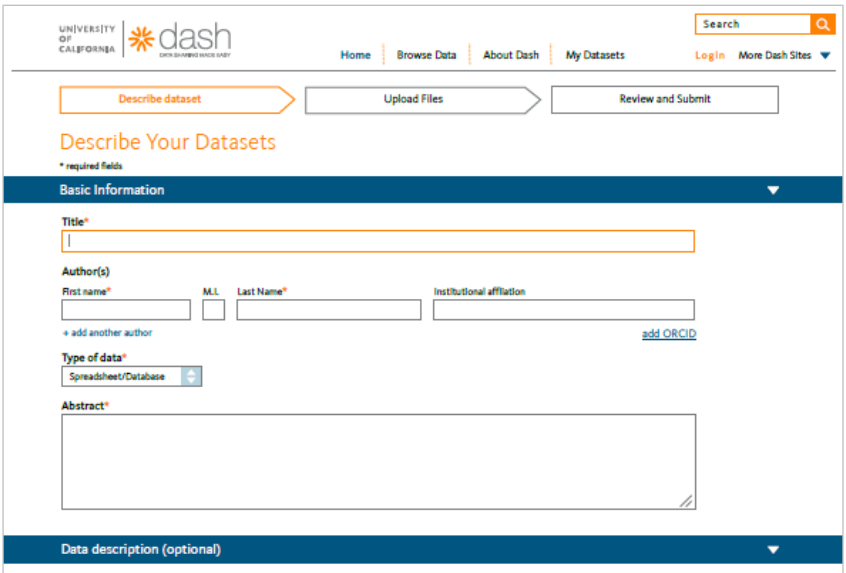

Figure 3. Metadata entry.

Reuse of data depends on stable citation and actionable persistent identifiers. All datasets in Dash are automatically assigned DOIs provided through CDL's EZID

[7] DataCite Metadata Schema: https://schema.datacite.org/

service, which in turn receives them from DataCite, an international consortium working towards easier access to research data, increased acceptance of research data as legitimate contributions to the scholarly record, and supporting data archiving, of which UC3 is a founding member.[8]

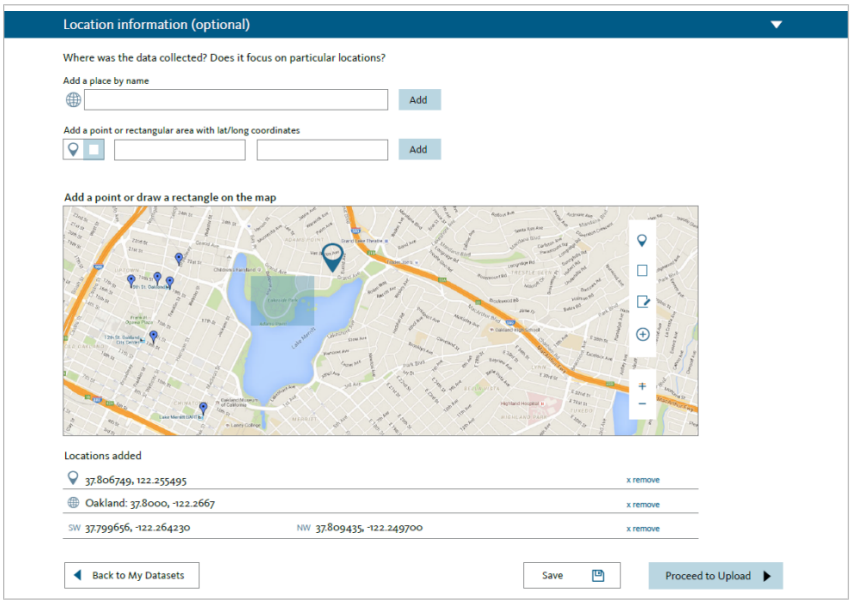

Figure 4. Geospatial location.

Dash delegates the responsibility for the data preservation function to the underlying repository with which it is integrated. Dataset discovery, on the other hand, is provided directly by Dash, which incorporates mechanisms for faceted browsing and keyword searching of metadata, as well as visual geospatial search (Figure 5).

The genesis of Dash began in 2012 with the development of DataUp, a cloud-based web service and Excel plugin for curating tabular datasets (Strasser, Abrams, and Cruse, 2014a). This effort, which garnered UC3 the 2013 Innovation Award from the National Digital Stewardship Alliance, was funded by Microsoft Research and the Gordon and Betty Moore Foundation. DataUp was integrated with ONE*Share*[9], an open data repository co-sponsored by UC3, DataONE, and the University of New Mexico Library, and intended as a repository "of last resort" for researchers without recourse to other appropriate institutional or disciplinary solutions. Although branded as a repository, in actuality ONE*Share* is a public Merritt collection, whose underlying storage is hosted at the University of New Mexico. ONE*Share* is also a member node on the larger DataONE data grid, so all datasets are discoverable through DataONE's catalog and search interface, which aggregates the contents of 29 international data repositories.[10]

At about the same time, a parallel effort named DataShare was started as a pilot project between UC3, the UC San Francisco (UCSF) Library, and UCSF Clinical and Translational Science Institute to create a portal for biomedical science (Abrams et al., 2013). While DataUp focused on a particular data type, tabular datasets, DataShare focused on a particular data genre, biomedicine. While both were of limited scope, the DataUp and DataShare projects sought to achieve the same goal: facilitating effective RDM practices. So it made sense for UC3 to merge them into a single product that has

[8] DataCite: http://www.datacite.org/
[9] OneShare: https://oneshare.cdlib.org/xtf/search
[10] DataONE Search: https://search.dataone.org

 doi:10.2218/ijdc.v11i1.408

since been enhanced into a comprehensive service for data curation and publication under the Dash name (Abrams et al., 2014). The UC3 deployment of Dash is integrated with the Merritt repository and is offered for use to the entire UC community.

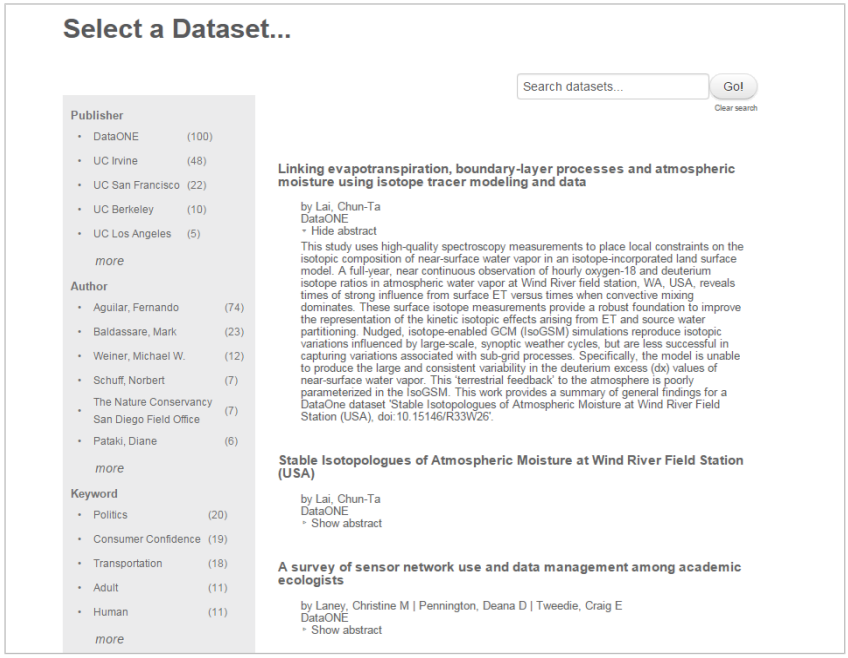

Figure 5. Search results.

UC3 is now completing a major reimplementation of the Dash system with funding from the Alfred P. Sloan Foundation. Whereas the previous version was tightly coupled to the Merritt and ONE*Share* repositories, the new release will be applicable to any standards-compliant repository supporting common protocols for data submission and harvesting. It will also feature a new UI refactored through extensive focus group evaluations to provide a simpler and more intuitive user experience. UC3 is also working to develop and promote an open source community for Dash as a component of its long-term sustainability plan.

Adoption

Dash has been adopted for use by six UC campuses, the Lawrence Berkeley National Laboratory, and the UC Office of President, the administrative home for centralized research initiatives such as the UC Natural Reserve System, a group of 38 biological and environmental field stations throughout the state. Outreach and frontline support for Dash are shared with campus library partners. Dash employs a multi-tenancy user interface (UI) providing partners with extensive opportunities for local branding and customization, use of existing campus login credentials, and, importantly, offering the Dash service under a campus-specific URL, an important consideration helping to drive adoption. UC3 expects that the forthcoming revision of Dash will facilitate use by the four remaining UC campuses, as well as other organizations wishing to provide a simple, intuitive data publication service on top of more cumbersome legacy systems. In addition to campus library collaborations, Dash, like its DataUp predecessor, is

integrated into the DataONE network via the ONE*Share* repository. The DataONE tenant in the Dash UI is available for use by scholars unaffiliated with UC. The UC campus-specific Dash instances accept data under the terms of the Creative Commons CC-BY license; the DataONE instance relies on the CC0 public domain declaration.

UC campus use of Dash incurs service fees based on storage utilization. These fees are covered by some campus units, generally the library, on behalf of the entire campus community. Public use of the DataONE instance and ONE*Share* repository is free; operational costs for ONE*Share* are subsidized by DataONE and the UNM Library. UC service fees are currently assessed and billed on an annual basis, but UC3 is pursuing an alternative option for paid-up pricing under which a one-time, up-front fee is assessed at the time of submission. This fee is calculated to be sufficient, when augmented by subsequent investment income, to fund data management services for a fixed term. Paid-up pricing is particularly important to address the issue of data that would otherwise be "orphaned" at the conclusion of funded research projects. Allowing paid-up fees to be built into project budgets enables continuity of management following the conclusion of funded activities.

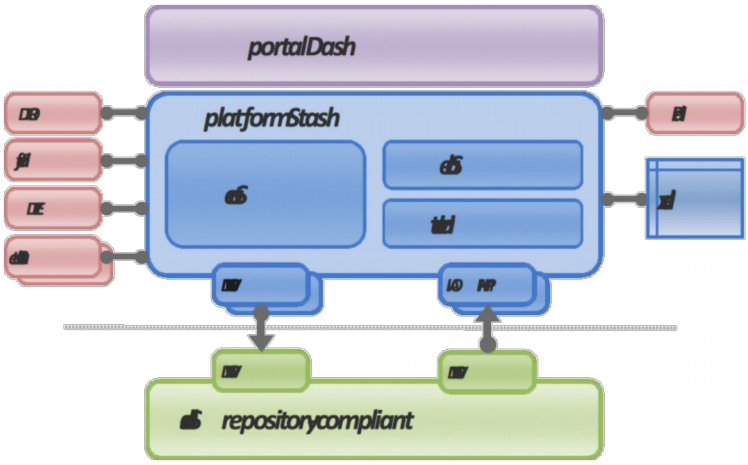

Figure 6. Dash high-level architecture.

Architecture and Implementation

Dash is based on an underlying Ruby-on-Rails data publication platform called Stash. Stash encompasses three main functional components: Store, Harvest, and Share (Figure 6). The Store component is responsible for the selection of datasets; their description in terms of configurable metadata schemas, including specification of ORCID[11] and Fundref[12] identifiers for researcher and funder disambiguation; the assignment of DOIs for stable citation and retrieval; designation of an optional limited time embargo; and packaging and submission to the integrated repository. The Harvest component is responsible for retrieval of descriptive metadata from that repository for inclusion into a Solr search index shared with the Share component, based on GeoBlacklight,[13] which is responsible for the faceted search and browse interface.

[11] ORCID: http://orcid.org/
[12] Fundref: http://www.crossref.org/fundingdata/
[13] GeoBlacklight: http://geoblacklight.org/

 doi:10.2218/ijdc.v11i1.408

Individual dataset landing pages are formatted as an online version of a data paper, presenting all appropriate descriptive and administrative metadata in a form that can be downloaded as an individual PDF file, or as part of the complete dataset download package, incorporating all data files for all versions.

To facilitate flexible configuration and future enhancement, all Stash support for the various external service providers and repository protocols are fully encapsulated into pluggable modules. Metadata modules are available for the DataCite and Dublin Core[14] metadata schemas. Protocol modules are available for the SWORD 2.0 deposit protocol[15] and the OAI-PMH[16] and ResourceSync[17] harvesting protocols. Authentication modules are available for InCommon/Shibboleth[18] and Google/OAuth[19] identity providers (IdPs). UC3 anticipates that modules for additional metadata schemas and repository protocols will become available in the future through internal and community-supported development efforts. Other anticipated enhancements include support for data-level metrics (DLM), such as those aggregated by the Making Data Count project (Kratz and Strasser, 2014a; 2014b) on which UC3 is collaborating with DataCite, DataONE, and PLOS.[20]

Conclusion

The intention of Dash is to offer maximal curation function with minimal intrusion on scholarly work practices. Researcher responsibility is minimized to self-service data uploading and providing meaningful description, while the technical details attendant to identifier assignment, metadata serialization, dataset packaging, repository deposit, and metadata harvesting and indexing are all handled automatically by Dash. Researchers can concentrate on scholarly activities while still reaping the benefits of curation best practices with little extra effort. Thus, Dash is a key component for making research data sharing easy at the University of California and beyond.

Acknowledgements

The development of the current version of Dash was generously supported by the Alfred P. Sloan Foundation, #G-2014-13603.

[14] Dublin Core: http://dublincore.org/
[15] SWORD 2.0: http://swordapp.org/sword-v2/sword-v2-specifications/
[16] OAI-PMH: https://www.openarchives.org/pmh/
[17] ResourceSync: https://www.openarchives.org/rs/toc
[18] InCommon: http://www.incommon.org/
[19] Google OAuth: https://developers.google.com/identity/protocols/OAuth2
[20] DLM: https://dlm.datacite.org/