ALFRED P. SLOAN FOUNDATION
GRANT SUMMARY

Grant Number: G-2014-13603

Approved on:	October 14, 2014
Program Director:	Joshua Greenberg
Grantee:	REGENTS OF THE UNIVERSITY OF CALIFORNIA
Tax Status:	501(c)(3)
Principal Investigator / Primary Contact:	Patricia Cruse
Purpose:	To promote research data sharing by enhancing the usability (design, functionality, and user experience) of existing community repositories
Duration:	12 months | 11/1/2014 - 10/31/2015
Amount:	$266,958.00
Payment Schedule:	$266,958.00 **Date:** 11/15/2014
Check Payable to:	REGENTS OF THE UNIVERSITY OF CALIFORNIA Mail to: Lourdes DeMattos Research Policy Analysis and Coordination University of California, Office of the President 405 Hilgard Avenue Box 951432 Los Angeles, CA 90095
Grant Agreement and Report Due Dates:	Grant Agreement **Date:** 10/31/2014 Substantive Final **Date:** 12/1/2015 Financial Final **Date:** 12/1/2015

Please email reports to Joshua Greenberg at greenberg@sloan.org and cc: grantsadmin@sloan.org

GENERAL CONDITIONS

1. **Purpose:** This grant is provided only for the purpose described on the grant summary and the grantee agrees to use grant funds only for this purpose. The grant may be terminated if the Foundation, in its sole discretion, determines that the grantee failed to comply with the purpose or the terms and conditions of the grant.

2. **Accounting and Audit:** The grantee will record the receipt of this grant, together with any expenditures that relate to it, in such a manner (e.g. in a separate ledger account) as to enable the Foundation to verify that the funds received have been expended for the purpose for which the grant was made. All expenditures must be in accordance with the established policies of the grantee.

3. **Records:** Records pertaining to this grant, along with copies of reports submitted to the Foundation, must be retained in the grantee's files for four years after the termination of the grant. The Foundation reserves the right to audit these records during and after the term of the grant or to have an audit made by independent auditors.

4. **Budget:** If this grant is made on the basis of an approved budget, it is understood that no substantial variations will be made in the budget without the Foundation's prior approval.

5. **Unused Funds:** Upon termination of the grant, funds in excess of U.S. $100 not expended for the purposes of this grant shall be returned to the Foundation in U.S. dollars. Unused funds of U.S. $100 or less may be retained and used for general purposes.

6. **Tax Exemption:** The grantee will notify the Foundation immediately of any change or any expected change in its status as a not-for-profit tax-exempt organization or as an organization not classified as a private foundation as defined in the Internal Revenue Code.

7. **Reporting:** Within 30 days after the end of the grant period, or the completion of the use of the grant funds, whichever comes sooner, the grantee agrees to give the Foundation a final written report (in English) summarizing the work done and appraising the results. The grantee also agrees to give the Foundation a final financial report (in U.S. dollars) containing a statement of payments received, and expenditures made or incurred. If for any reason the period of the grant exceeds one year, the grantee agrees to give the Foundation interim narrative reports of progress to date and interim financial reports. The grant summary notes the dates by which the Foundation should receive these reports.

8. **Lobbying:** The Grantee agrees that no portion of these funds will be used for any attempt to influence legislation, to influence the outcome of any specific election or to carry on directly or indirectly any voter registration drive. Should the results of the project be used for technical assistance, the Grantee agrees that it shall be at the written request of such body or duly constituted committee thereof, and the results will be made available to the entire body.

9. **CounterTerrorism:** The Grantee agrees that it will use the grant in compliance with all applicable antiterrorist financing and asset control laws and regulations.

10. **Key Personnel:** If any of the key personnel identified in the grant application leave the institution(s) to which the grant was made, the Alfred P. Sloan Foundation must be notified within 30 days of that person's departure.

11. **Acknowledgment of Support:** The grantee will acknowledge the Foundation and its contribution to the project supported by the grant (the *Project*) in any materials (in any media) produced in connection with the Project (*Materials*) and in any public statements made regarding the Project, including but not limited to interviews, speeches, press releases, Web site announcements and publications (*Media Releases*). The grantee agrees to establish the Foundation’s acknowledgment rule prior to interviews or other promotional efforts. The grantee also agrees to provide the Foundation the form of its acknowledgment in all Materials and Media Releases prior to distribution of any Materials or Media Release. Acknowledgments in all Media Releases and Materials will refer to the Foundation by its full name *Alfred P. Sloan Foundation.*

Dash: Improving Community Repositories for Better Data Sharing

Contents

1. Problem Statement

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 [1]. 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 [2]. There are many reasons for this lack of data stewardship, and the most commonly cited are as follows:

1. A lack of education about good data management practices [3],
2. Poor incentives for researchers to describe and share their datasets [4], and
3. A dearth of easy-to-use tools for data curation.

The incentives problem is being addressed by increasing mandates for more proactive data management. Furthermore, it is increasingly no longer optional to provide access to data: sharing is becoming a matter of institutional policy and disciplinary best practice, and a precondition for grant funding and publication (e.g., recent directives from the US Office of Science and Technology Policy [5]). Although this means researchers have more incentives to participate in data stewardship, there is still a lack of easy-to-use tools, resulting in practices that may impede future access to datasets. As evidence, many researchers that do choose to "archive" are doing so in one of three ways, each potentially problematic:

- *Commercially owned systems (e.g., fig***share***, Dropbox, Amazon S3)*. Potential problem: these solutions are owned by groups who may not fully share the academic value of openness, and who may not have a primary goal of long-term data preservation.

- *Supplemental materials alongside the main journal article*. Potential problem: These materials are not always preserved and accessible for the long term [6].
- *Personal website.* Potential problem: personal websites are often poorly maintained and eventually abandoned. Both research and anecdotal evidence indicate the average lifespan of a website is between 44 and 100 days [7].

A better option for data archiving is **community repositories**, which are owned and operated by trusted organizations (i.e., institutional or disciplinary repositories). Although disciplinary repositories are often known and used by researchers in the relevant field, institutional repositories are less well known as a place to archive and share data.

Why aren't researchers using institutional repositories? **First**, the repositories are often not set up for self-service operation by individual researchers who wish to deposit a single dataset without assistance. **Second**, many (or perhaps most) institutional repositories were created with publications in mind [8], rather than datasets, which may in part account for their less-than-ideal functionality. **Third**, user interfaces for the repositories are often poorly designed and do not take into account the user's experience (or inexperience) and expectations. Because more of our activities are conducted on the Internet, we are exposed to many high-quality, commercial-grade user interfaces in the course of a workday. Correspondingly, researchers have expectations for clean, simple interfaces that can be learned quickly, with minimal need for contacting repository administrators.

Our Solution

We propose to address the three issues above with Dash, a well-designed, user-friendly data curation platform that can be layered on top of existing community repositories. Rather than creating a new repository or rebuilding community repositories from the ground up, Dash will provide a way for organizations to allow self-service deposit of datasets via a simple, intuitive interface that is designed with individual researchers in mind. Researchers will be able to document, preserve, and publicly share their own data with minimal support required from repository staff, as well as be able to find, retrieve, and reuse data made available by others.

2. Related Work

The Dash platform will be based on the existing DataShare service[1], collaboratively developed by the University of California Curation Center (UC3) at the California Digital Library (CDL); the University of California, San Francisco (UCSF) Library and Center for Knowledge Management; and the UCSF Clinical and Translational Science Institute (CTSI). DataShare was intended as a working prototype: it provides the basic functionality required to meet the needs of a simple archiving platform [9]. While it solves the first and second problems mentioned above – it is a self-service platform for curating datasets – it still has the look and feel of a prototype. Second, the DataShare interface was built specifically for UCSF researchers, who primarily generate biomedical data. Finally, DataShare needs substantial technical work to be more generalizable to repositories external to UCSF.

[1] http://datashare.ucsf.edu/

Other work in this area of easy, self-service research data curation for sharing includes fig**share**[2] (an open repository owned by Macmillan Publishers), and Zenodo[3], a platform from CERN (European Organization for Nuclear Research) that preserves both code and datasets. Librarians are hesitant to point researchers towards fig**share** since a commercial publisher owns the service, although their clean user interface and simple sign-up have resulted in many researchers discovering fig**share** themselves. Zenodo is a relatively new repository option, and we plan to contact the project leads to determine how we can learn from one another as we move forward with Dash.

3. Proposers' Qualifications

The success of Dash is contingent on assembling a team that can effectively respond to the key elements of the proposed project: understanding of researchers' needs, knowledge of the data sharing landscape, digital preservation expertise, ability to engage with the community, and robust institutional support. The assembled staff is comprised of internationally recognized experts in all areas touched on by Dash: data management and sharing, data publication, digital curation, preservation, scholarly research, and scholarly communication. Through participation and engagement in a range of organizations and initiatives (DataCite, Research Data Alliance, DataONE, NDIIPP/NDSA, Digital Preservation Network (DPN), HathiTrust, PREPARDE, 4C project) the UC3 team has collectively been at the forefront of many initiatives that are changing the scholarly research landscape.

[2] http://figshare.com/
[3] http://zenodo.org/

A project of this nature, which is pushing the envelope in how researchers work, demands that project staff understand the challenges today's researchers face. Strasser is an oceanographer by training who represents researchers in discussions about data sharing and open science, and in software requirements development. She is an active member of the DataONE community, primarily as a working group lead for community engagement. She is an advocate for open science and data sharing, and has an active blog[4] that covers these topics. Much of her work involves building partnerships with librarians, publishers, researchers, project managers in both industry and academia, and others to ensure that UC3's work stays relevant and valuable.

Data management, sharing, curation, and preservation are the technical underpinnings of the proposed project. For over a decade Cruse and Abrams have played a leadership role in these areas and are recognized experts in their fields. Cruse founded of the University of California digital preservation program in 2002 and has brought to fruition UC3's successful user facing services that are actively used by the UC community and beyond. Abrams is recognized as a digital preservation expert and has brought his technical expertise to many of UC3's services. UC3 has successfully deployed several groundbreaking services that directly support the rapidly changing scholarly environment: Merritt[5], EZID[6], DataUp[7], the Web Archiving Service[8], and the Data Management Plan Tool (DMPTool)[9]. All of these services were conceived, designed, and implemented by the UC3 staff and its partners. A core value of

[4] http://datapub.cdlib.org/
[5] http://merritt.cdlib.org/
[6] http://ezid.cdlib.org/
[7] http://dataup.cdlib.org/
[8] http://was.cdlib.org/
[9] http://dmptool.org/

developing these services was engaging with the community of stakeholders throughout all stages of a project, from design to development and to production and use.

The CDL exists to support the University of California's scholarship mission in an increasingly digital world. CDL's diverse and talented staff has assembled one of the world's largest digital research libraries, and has changed the ways that faculty, students, and researchers discover, access, use, and preserve information. UC3 is a programmatic unit of the CDL and is a creative partnership bringing together the expertise and resources of the CDL, the ten UC campuses, and the broader international curation community. Harnessing the collective energy and innovation of its partners, UC3 provides solutions that are out of reach of any individual partner.

4. Approach

We propose to enhance the existing DataShare application to address problems associated with community repositories. The new system, to be called "Dash" (DAtaSHare), is a data curation platform that sits on top of existing repositories. We will focus on ensuring that the Dash interface is well-designed, visually appealing, and user-friendly, in an effort to maximize its uptake by researchers in search of data curation solutions.

At its core, Dash has two main functions: self-service data deposit and data discovery. **For data deposit**, researchers can: (1) upload data using drag-and-drop or file browse modalities; (2) create metadata using an intuitive input form with minimal required fields; (3) associate the data with a persistent identifier (e.g., DOI); and (4) deposit the data to a repository without having to worry about details of submission packaging or protocols. **For data discovery**, Dash has faceted search-and-browse capabilities and is optimized for search engines. In addition, all Dash metadata will be indexed by leading abstract and indexing services (e.g., Thomson Reuter's Data Citation Index).

Other Dash features include the ability to version datasets, provision of standard data use agreements to provide controlled terms of use, and privacy and embargo options. Note that since Dash is an added-value layer on top of existing repositories, all of the underlying repository's native functions will remain intact.

In addition to deploying Dash as a website, we will extend the possibilities for incorporation of Dash functionality into other online contexts by implementing the submission and discovery functions as embeddable widgets. By exposing its functionality as embeddable widgets, Dash can be directly integrated with existing online scientific tools and workflows, providing enhanced opportunities for data sharing. We envision that in the future, the majority of Dash use may come from being embedded into contexts already in use by researchers as part of their workflows.

Three Phases of Dash Development

Phase 1: Requirements Gathering

Much of the work needed for Dash is related to the user experience. Input and iterative feedback from the community is therefore critical to the success of this project. Before the design process begins, we will build requirements for researchers via interviews and surveys. Questions will focus on understanding the challenges they face when depositing data, and assessing the usability of existing options like fig**share**. For more on the incorporation of user input throughout Dash development, see the *Feedback* section and Appendix 1 below.

Sample Interview Questions for Researchers

1. Where do you store, archive, and/or share your datasets?
2. How do you decide where to put datasets?
3. How much metadata do you typically create when sharing datasets?
4. What types of data would you like to share?
5. How easy (or difficult) are existing user interfaces for data deposit?

Phase 2: Design Work

Based on surveys and interviews with researchers (Phase 1), we will develop requirements for a researcher-focused user interface that is visually appealing and easy to use. Over the course of implementing the design, we will work closely and iteratively with the Dash user community to ensure the resulting platform meets their needs. The design process will involve collecting examples of quality websites, interviewing researchers and other stakeholders about their preferences, and surveying the community at large about design components that we should incorporate into the Dash interface.

We will identify potential design firms that have successfully worked in the area of application development, and communicate the project needs based on community conversations. After selecting a design firm, we will then work closely with them to develop webpage mockups, obtain feedback from potential users, and build wireframes for more formal user assessment. This assessment will involve one-on-one testing with researchers, and surveys via the CDL blog that are advertised widely using social media.

Once the beta version of Dash is created, we will conduct more testing and surveys to fine-tune the application for maximum appeal and usability, and to ensure the best user experience possible.

Phase 2 – Iterative Community Design Process

1. Collect examples of quality websites (based on researcher-focused criteria)
2. Identify potential design firms; obtain proposals
3. Hire design firm to develop wireframes/mockups
4. Gather feedback from user community on wireframes/mockups
5. Build new user interface; create beta version of Dash
6. Gather feedback on beta version
7. Finalize user interface based on feedback

Phase 3: Technical Work

Dash will be an added-value data sharing platform that integrates with any repository that supports two key protocols: (1) SWORD (Simple Web-service Offering Repository Deposit)[10] for data deposit; and (2) Atom[11] for metadata harvesting. These protocols are widely supported by leading open source repository systems such as DSpace, EPrints, and Fedora.

The two main functions of Dash, data submission and data discovery, will communicate through backend application programming interfaces (APIs) to subsystems implementing the necessary submission and discovery function (Figure 1). Submission will be implemented as a Ruby on Rails application, and discovery will be based on Solr[12]. Each subsystem will have an abstraction layer between its business logic and the repository; this will facilitate the future extension of Dash to support additional submission and harvesting protocols.

[10] http://swordapp.org/
[11] http://uq-eresearch-spec.github.io/atom-pmh/
[12] http://lucene.apache.org/solr/

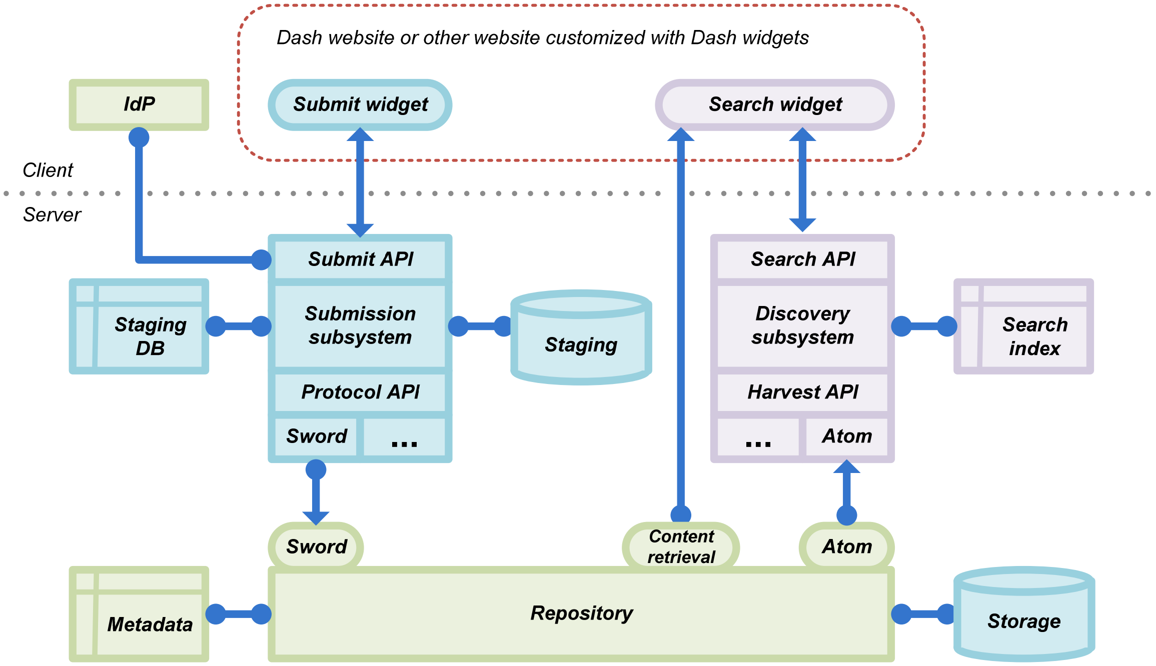

Figure 1: Provisional Dash architecture

While data discovery is open to the public, data submission will require prior authentication. Dash will support InCommon/Shibboleth[13] authentication to allow login using local institutional credentials. Authentication support will be designed and implemented in a manner that will permit the easy future extension for alternative authentication schemes such as OAuth[14].

Phase 3 – High-level Technical Implementation Tasks	
1. Stabilize the existing UCSF DataShare codebase	a) Fork the code and establish a GitHub repository. b) Deploy to local development environment.
2. Refine workflow and UI design	a) Focus group evaluation and design.
3. Enhance features	a) Add support for authentication from any InCommon/Shibboleth IdP. b) Add support for ORCID identifiers of data contributors. c) Generalize the application for applicability to any repository supporting SWORD and Atom.

[13] http://www.incommonfederation.org/
[14] http://oauth.net/

	d) Add support for SWORD deposit to the Merritt repository. e) Add support for multi-institutional configuration of style sheets and URLs.
4. Create documentation	a) Installation and operation guide. b) Customization guide.

An important feature of Dash will be its capability for institutional or organizational branding. This will be made possible via customized style sheets and some basic configuration tasks. A single instance of Dash can support customized user interfaces for multiple institutions; this is an actual use case for the CDL, which will be supporting branded Dash "views" for all ten UC campuses, including the use of campus-specific URLs.

The CDL Merritt repository will be used as the testing prototype during the project. In addition, the project team will recruit high-visibility, high-impact partners for integration testing and deployment. Our selection of these partners will be based on several criteria. We will take into account reports on researchers' preferred repositories obtained from surveys and interviews in Phase 1, plus community knowledge about reliable data repositories, e.g., those recommended by *Nature Scientific Data*[15]. We will also liaise with communities around leading open source repository platforms, including Dataverse, DSpace, and Fedora, at conferences and meetings; these may include the Open Repositories conference and Coalition of Networked Information Member Meetings. The partners will be chosen to balance technological and disciplinary heterogeneity.

[15] http://www.nature.com/sdata/data-policies/repositories

Feedback Strategy

The uniqueness and importance of Dash lies in its user interface. Because of this, the most critical component of this proposal is obtaining user feedback to ensure that the application is user-friendly and inviting. Towards that end, we envision an iterative process for designing the user interface. We have experience at UC3 in this process – team members successfully created the DataUp application, which was informed wholly by the user community via interviews, surveys, and social media [10].

The user community at our disposal for obtaining feedback is diverse and large. We have connections to researchers at the 10 UC campuses, including UC Berkeley where the recent Berkeley Institute for Data Science initiative ensures interest among researchers for developing tools for data sharing. In addition, Strasser's former work as a researcher allows her to connect with the target community in effective ways, especially via one-on-one interviews on campuses, at professional meetings, and via the DataONE community.

Metrics to Gauge Success

Dash will be a new data curation service for the UC researcher community, as well as a new data curation platform for existing repositories. Given our inability to establish clear "before" numbers, we will base our metrics on Croll and Yoskivitz's *Lean Analytics* framework [11] and track three questions applicable to new startup services:

1. Are we solving a need (empathy)?
2. Does Dash meet the need (stickiness)?
3. Are we acquiring and retaining users (virality)?

We plan address these questions in two ways: **quantitatively**, via both traditional site analytics and “alternative” metrics (e.g., mentions on Twitter and Facebook or in blogs), and **qualitatively**, via user interviews with researchers.

The most obvious **quantitative** metrics we will use for a project on improving institutional repositories include number of system users, number of return users, number of datasets uploaded, and number of dataset downloads. Ideally, we would have before-and-after statistics for each of these metrics, which would allow us to demonstrate why a well-designed user interface was critical for adoption of a data sharing culture. However, these metrics would not accurately reflect the utility of Dash for two reasons. First, there is a significant shift towards mandatory data sharing by publishers and funders. This means that upticks in data deposit or system use might be a reflection of these mandates, rather than the system’s adoption. Second, there is no good “before” instance of Dash. Although the UCSF DataShare system exists, it is set up for a unique academic situation: UCSF has no undergraduates and is primarily a biomedical research facility. These researchers often have existing disciplinary repositories for their datasets, plus complex regulations around sharing data due to their use of human subjects and/or patents associated with drug development.

For **qualitative** assessment, we will incorporate our interviews into Phase 1 above, obtaining researcher feedback on Dash as it develops. Based on our interview questions, we will be able to assess whether researchers would use Dash in the future and/or recommend it to other researchers.

Throughout the project we will capture metrics as indicators of Dash adoption and community uptake. We will particularly monitor metrics with regard to project priorities:

(1) use of Dash for data deposition and access; (2) adoption of Dash platform by community repositories. These data will provide an indication of success and a strong foundation for post facto assessment of the Dash's utility.

Category	***Relevant Question***	***Quantitative***	***Qualitative***
Empathy	Are we solving a need?		Interviews, desk-based research
Stickiness	Does Dash application meet that need?	Surveys of user experiences, number of users, return users, number of datasets uploaded	Interviews
Virality	Can Dash acquire and retain users?	Surveys, website analytics	Interviews

See Appendix 1 for our recruitment strategy and sample questions.

Dissemination of work

The main outcome of this project will be the Dash software itself. We plan to make this software and all technical documentation publicly available on GitHub after the project is completed. Based on conversations with repository administrators, there is significant interest from organizations in deploying Dash on top of their own repositories; ideally these organizations will become part of an active Dash community hosted on GitHub. Upon completion of the Dash platform, we plan to promote its widespread implementation and use by community repositories via multiple channels. We will present the software at repository platform development communities, use targeted communication to well-known repositories for potential Dash implementation, and rely on our DataONE connections to inform other community repositories that may benefit from Dash.

Project roles, responsibilities and coordination

The Dash project will be lead by Patricia Cruse, Carly Strasser, and Stephen Abrams at the UC3 (CDL). Cruse, UC3 Director, will be responsible for all aspects of the development and deployment of Dash, build relationships with key project stakeholders (UC researchers and the DataONE community), set project priorities, guide and review results, and ensure successful completion of the project. Carly Strasser, UC3 Research Data Specialist, will manage project activities, coordinate community interaction, requirements gathering, and design work, user testing, and will lead all the interaction with the selected design firm on the development of wireframes, mockups, and the final product. Strasser will also work collaboratively with the team to disseminate results. Stephen Abrams, UC3 Associate Director, will be responsible for the technical design, development, and deployment of Dash. Abrams will bring his rich technical knowledge to the project and work to integrate all user feedback and user-interface components into Dash's technical requirements.

A mid-range application developer will be hired to implement the technical design based on the gathered user cases and requirements. The developer will also interact with Strasser in the ongoing iteration of Dash. We will work closely with a professional designer on the development of Dash's information architecture, visual design, web development, and logo design. We will interact with the design firm throughout all of these phases.

5. Outputs

Output 1: Fully functioning version of the Dash software, with an improved user interface and visual appeal. The new version of the Dash with its expanded functionality will be freely accessible via the web. Any user will be able to discover data via the search interface. Initially, the UC campuses will be able to deposit and describe their data via the submission workflow.

Success Statement 1: Researchers that use Dash would be willing to recommend it to their colleagues. This will be measured via a user survey.

Success Statement 2: Researchers from several disciplines, with different types of data, successfully use the Dash repository to provide public access to their data.

Output 2: Dash code base. This will be freely available on GitHub after the project is completed (i.e., after the successful deployment of the Dash system at the CDL). All software will be licensed in accordance with UCOP policy. Related outputs will be the technical documentation around the Dash system, which will also be made available on GitHub.

Success Statement: A GitHub code repository is set up, with extensive documentation, and is able to be forked and used/understood by at least one organization other than the CDL (i.e., partner organizations).

Output 3: A thriving open-source community around the Dash code base. We anticipate interest in the Dash platform from a wide range of data curation and preservation experts and institutions, and we will foster this new community around Dash via the GitHub repository interface (issues, wiki, etc.).

Success Statement: There are several individuals or organizations that follow the Dash repository on GitHub, and several interactions on the wiki or issue tracker that involve individuals outside of the CDL.

Output 4: Survey results on the current archiving practices of researchers, and their needs for data archiving systems. The novel information we will collect pertains to the influence of user design on archiving practices, which we will focus on during our interactions with the researcher community.

Success Statement: Submission of an open-access peer-reviewed publication with the results of the survey.

6. BUDGET JUSTIFICATION

The UC Curation Center at the CDL requests $ 266,958 for the Dash project. The budget is allocated to staff salaries for project oversight, project management and community interaction, technical design and oversight, software development, user-interface and visual design, and travel. The effort funded by budget is allocated to the following activities:

- Project oversight: salary and benefits (at a rate of 41%) are requested for 5% Patricia Cruse at an annual base rate of (2014-2015) on a 12 month calendar year. Cruse will provide general oversight and direction and build relationships with build relationships with key project stakeholders (UC and DataONE) communities, set project priorities, guide and review results, and is ultimately responsible for completion of the project.
- Project management and community interaction: for salary and benefits (at a rate of 40%) are requested for 25% Carly Strasser at an annual base rate of (2014-2015) on a 12 month calendar year. Strasser will manage project activities, coordinate community interaction, requirements gathering and design work, user testing, and will lead all the interaction with the selected design firm on

the development of wireframes, mockups, and the final product. Strasser will also work collaboratively with the team to disseminate results.

- Technical design and oversight: for salary and benefits (at a rate of 44%) are requested for 10% Stephen Abrams at an annual base rate of (2014-2015) on a 12 month calendar year. Abrams will be responsible for the overall technical design of the new application. Abrams will provide technical oversight throughout the entire project. Abrams will bring his rich technical knowledge to the project and work to integrate all user feedback and user-interface components into technical Dash's requirements.
- Application development: Funds are requested for one Application Developer working at a 100% effort with an annual rate estimated of $119,540 on a 12 month calendar with benefits (at a rate of 39%). The estimated rate is based on the University of California's pay grade for a midrange technical programmer.
- User-interface and visual design: One time funds, $48,000 are requested to contract with a to-be-named professional design firm to work with the project team on the development of Dash's information architecture (based on community feedback), visual design, web development, and logo design. This cost is based on initial interaction with several design firms and based on previous experience with design firms.
- $3000 are requested in funds to support travel to meet with stakeholder community Projected travel will include attendance at well-known conferences (e.g., ESA, AGU, IDCC, iPRES, Open Repositories) and will take place early in the project schedule for face-to-face feedback with key stakeholders, and later in the schedule to apprise

the community of progress and promote the adoption of use of the final deliverables. Note that funds will not support all of these conferences and therefore will be prioritized based on timing and utility.

7. Existing Relevant Support

Currently Dash receives no funding or support outside of the California Digital Library.

8. Other Sloan Foundation Funding

The CDL, alongside University of Virginia and University of Illinois Urbana-Champaign, received funding in 2013 from Sloan for “DMPTool 2: Responding to the community.” This project is in its final phase, with all major goals met.

APPENDIX 1: REQUIREMENTS & FEEDBACK STRATEGY

Communities of interest

In the course of gathering requirements for and evaluating Dash, we will target researchers from all disciplines, especially those that produce “small” datasets (e.g., ecologists, geologists, social scientists). These individuals represent the long tail of research data[16] and are therefore most likely to benefit from a platform like Dash. In addition, these individuals are less likely to have support for data archiving since they are not producing large datasets. We will interview all levels of professional scientists, from graduate students to principal investigators, and we will include scientists from a spectrum of institutions (academia, museums, non-profit organizations).

Modes for recruiting researchers for general input

1. Campus visits: We plan to use our connection to the UC community to meet with researchers at select UC campuses to discuss their data archiving needs (before development) and to obtain feedback on the Dash platform (after development).
2. Online: We will use Twitter to send out links to short surveys and polls, and to keep the community informed on the Dash project’s progress and outcomes. We will also will provide updates on the project and ask for feedback on its progress and development via the Data Pub blog.
3. Events: We will target several disciplinary meetings and use the venues for one-on-one conversations with researchers and to administer quick polls. Potential venues include the Ecological Society of America meeting, the American Geophysical Union meeting, and the Earth Science Information Partners meeting.
4. Emails: We will use researcher-focused listservs (e.g., ECOLOG), to promote surveys and request feedback via polls.

[16] Heidorn, P. Bryan (2008). Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2) Fall 2008.

5. Targeted outreach: We will use our existing contacts within the research community (via DataONE, previous campus interactions, etc.) to identify individuals well-suited for providing useful input and feedback for Dash.

Lines of questioning for developing Dash requirements

- Have you archived data in the past? If yes,
    - What repository did you use? Why did you choose it?
    - How easy or hard was the process? Describe.
    - What did you like and dislike? Would you archive there again?
- Do you know of any repositories where you could put your data? Which ones?
- What types of data do you produce? (size, quantity of files, file type, metadata type)
- How much metadata are you willing to enter? Would you archive formal or informal metadata if it wasn't mandatory?
- How likely is it that someone could use your dataset given the minimal metadata many repositories request?
- Where do you go to find datasets?
- What are the major barriers to providing public access to your data?

Lines of questioning for obtaining feedback on Dash

- What was easy about using Dash? What was hard?
- How likely are you to suggest Dash to your institutional colleagues? To those outside of your institution?
- Would you be willing to provide more metadata than was required?
- Was the upload system easy to use? Were the problems with uploading?
- How was the user interface? Appealing? Annoying? What would you change?
- Was it clear what your next steps were in the process? Was there a point at which you were "lost"?
- Were you able to find help text when you needed it?
- How long did it take to complete the round-trip process of depositing data into Dash?

APPENDIX 2: BIBLIOGRAPHY

[1] Hey, T, S Tansley, and K Tolle (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. Available at http://fourthparadigm.org/

[2] Tenopir, C, S Allard, K Douglass, A Aydinoglu, L Wu, E Read, M Manoff, and M Frame (2011), “Data Sharing by Scientists: Practices and Perceptions”. *PLoS ONE* 6: e21101+. http://dx.doi.org/10.1371/journal.pone.0021101

[3] Strasser, C and SE Hampton (2012), “The Fractured Lab Notebook: Undergraduates and Ecological Data Management Training in the United States”. *Ecopshere* 3:art116. doi:10.1890/ES12-00139.1

[4] Borgman, C (2012), "The conundrum of sharing research data," *Journal of the American Society for Information Science* 63(6): 1059-1078.

[5] Holdren, JP (2013), “Memorandum for the Heads of the Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research.” February 22, 2013 Memo from the White House Office of Science and Technology Policy. Available at http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[6] Evangelou, E, T Trikalinos, and J Ioannidis (2005), “Unavailability of online supplementary scientific information from articles published in major journals.” *FASEB Journal* 19(14): 1943-1944.

[7] Taylor, N (2011), "The average lifespan of a webpage," *The Signal Digital Preservation* Blog, available at http://blogs.loc.gov/digitalpreservation/2011/11/the-average-lifespan-of-a-webpage/

[8] Crow, R (2006), “The Case for Institutional Repositories: A SPARC Position Paper”. Available at http://scholarship.utm.edu/20/

[9] Abrams, S, P Cruse, C Strasser, P Willett, G Boushey, J Kochi, M Laurance, and A Rizk-Jackson (2014), "DataShare: empowering researcher data curation," *Ninth International Digital Curation Conference*, San Francisco. Available at http://www.dcc.ac.uk/sites/default/files/documents/IDCC14/Parallels/StephenAbrams_PrallelA4_26Feb.pdf.

[10] Strasser C, Kunze J, Abrams S and Cruse P (2014), “DataUp: A tool to help researchers describe and share tabular data”. *F1000Research*, 3:6.

[11] Croll, A, and B Yoskovitz (2013). *Lean analytics: Use data to build a better startup faster*. Sebastopol, CA: O'Reilly Media.

APPENDIX 4: KEY PROJECT STAFF CURRICULUM VITAE

1. Patricia Cruse, Director of UC3
2. Stephen Abrams, Associate Director of UC3
3. Carly Strasser, Data Curation Specialist at UC3

Stephen Abrams

Education

Boston University, BA in Mathematics

Harvard University, ALM in the History of Art and Architecture

Employment History

California Digital Library

- Associate Director, UC Curation Center, 2009-
- Senior Manager for Digital Preservation Technology, 2008-2009

Harvard University Library

- Digital Library Program Manager, 2003-2008
- Development Team Leader, 2002
- Senior Programmer, 1999-2001

Important Publications

Abrams, S, P Cruse, C Strasser, P Willett, G Boushey, J Kochi, M Laurance, and A Rizk-Jackson, "DataShare: Empowering researcher data curation," *International Journal of Digital Curation* 9:2 (2014). (In press)

Abrams, S, "How does the changing digital landscape affect preservation services for publishers and libraries," *Ithaka Sustainable Scholarship*, New York, October 21-22, 2013.

Abrams, S, A Rizk-Jackson, J Kochi, and N Wittman, "Sharing data-rich research through repository layering," *OR 2013, The 9th International Conference on Open Repositories*, Prince Edward Island, July 8-12, 2013.

Abrams, S, "Libraries and research data curation: Barriers and incentives for preservation, sharing, and reuse," *Future of Scientific Publishing: Open Access to Manuscripts and Big Data*, Stanford University, June 27, 2013.

Abrams, S, P Cruse, and J Kunze, “Augmenting repositories to showcase research data,” *OR 2012, The 7th*

International Conference on Open Repositories, Edinburgh, July 9-13, 2012.

Abrams, S, "Unified digital format registry (UDFR): A community resource for effective preservation," *Beyond Borders: 76th Annual Meeting of the Society of American Archivists*, San Diego, August 6-11, 2012.

Abrams, S, P Cruse, and J Kunze, "Total cost of preservation: Cost modeling for sustainable services," *CNI Spring 2012 Membership Meeting*, Baltimore, April 1-3, 2012.

Abrams, S, P Cruse, J Kunze, and D Minor, "Curation micro-services: A pipeline metaphor for repositories," *Journal of Digital Information* 12:2 (2011).

Abrams, S, J Kunze, and D Loy, "An emergent micro-services approach to digital curation infrastructure," *International Journal of Digital Curation* 5:1 (2010): 172-186.

Abrams, S, "File formats," *Digital Curation Manual* (Edinburgh: Digital Curation Center, 2007).

Abrams, S, ed., ISO 19005-1:2005, Document management – Electronic document file format for long-term preservation – Part 1: use of PDF 1.4 (PDF/A-1).

Patricia Cruse

Education

University of California, Berkeley
Slavic Languages and Literature, Bachelor of Arts, 1984
School of Library & Information Studies, MLIS, 1990

Employment History

Director, University of California Curation Center, California Digital Library, University of California, 2009-

Director, UC Libraries Digital Preservation Program, California Digital Library, 2003-2009.

Program Manager, Social Science Data Initiatives, California Digital Library, University of California, 1999-2003.

Manager, Government Information and Social Science Data Initiatives, Geisel Library, University of California, San Diego,1994-1999.

Important Publications

Abrams, S, P Cruse, C Strasser, P Willett, G Boushey, J Kochi, M Laurance, and A Rizk-Jackson, "DataShare: Empowering researcher data curation," *International Journal of Digital Curation* 9:2 (2014). (In press)

Strasser, C, J Kunze, S Abrams, and P Cruse. 2014. DataUp: A tool to help researchers describe and share tabular data. *F1000 Research* 2014, 3:6. Awaiting peer review. doi: 10.12688/f1000research.3-6.v1

Abrams, S, P Cruse, and J Kunze, “Augmenting repositories to showcase research data,” *OR 2012, The 7th International Conference on Open Repositories*, Edinburgh, July 9-13, 2012.

Abrams, S, P Cruse, and J Kunze, “Total cost of preservation: Cost modeling for sustainable services,” *CNI Spring 2012 Membership Meeting*, Baltimore, April 1-3, 2012.

Michener, W., P Cruse, J Kunze, et al. "DataONE: Data Observation Network for Earth – Preserving Data and Enabling Innovation in the Biological and Environmental Sciences." *D-Lib Magazine*, 17:1/2 (Jan/Feb 2011).

Abrams, S, P Cruse, J Kunze, and D Minor, “Curation micro-services: A pipeline metaphor for repositories,” *Journal of Digital Information* 12:2 (2011).

Kunze, J, P Cruse, et al., Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines, Report for the Gordon and Betty Moore Foundation, Nov 2010, http://escholarship.org/uc/item/9jw4964t

Abrams, S, J Kunze, et al., An Emergent Micro-Services Approach to Digital Curation Infrastructure, iPRES 2009: the Sixth International Conference on Preservation of Digital Objects. Proceedings, 4-11.

Abrams, S, P Cruse and J Kunze, "Preservation is Not a Place." *International Journal of Digital Curation*, 4:1 (2009): 8-21.

Cruse, P, B Sandore. "The Library of Congress National Digital Information Infrastructure and Preservation Program." *Library Trends*, 57:3 (2009): 301-314. Guest editor. DOI: 10.1353/lib.0.0055

Carly Strasser

Education

PhD, Biological Oceanography. MIT-WHOI Joint Program. 2008.

BA, Marine Science. University of San Diego. 2001.

Professional experience

Data Curation Specialist, University of California Curation Center, California Digital Library, University of California, 2011-

Postdoctoral Investigator, DataONE. 2010-2011.

Postdoctoral Investigator, University of Alberta and Dalhousie University. 2009-2010.

Postdoctoral Investigator, Woods Hole Oceanographic Institution. 2009.

Important Publications

Kratz , J and C Strasser. 2014. Data publication consensus and controversies. *F1000Research* 2014, 3:94. Awaiting peer review. f1000research.com/articles/3-94/v1

Strasser, C, J Kunze, S Abrams, and P Cruse. 2014. DataUp: A tool to help researchers describe and share tabular data. *F1000 Research* 2014, 3:6. Awaiting peer review. doi: 10.12688/f1000research.3-6.v1

Krier , L and C Strasser. 2014. Data Management for Libraries: A LITA Guide. 112 pages. Available from alastore.ala.org

Hartter, J, SJ Ryan, CA MacKenzie, JN Parker, and CA Strasser. 2013. Spatially Explicit Data: Stewardship and Ethical Challenges in Science. *PLoS Biology* 11(9): e1001634.

doi:10.1371/journal.pbio.1001634

Hampton, S, C Strasser, J Tewksbury, W Gram, A Budden, A Batcheller, C Duke, and J Porter. 2013. Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11(3): 156-162. doi: 10.1890/120103.

Hampton, S, C Strasser, and J Tewksbury. 2013. Growing Pains

for Ecology in the 21st Century. *BioScience*. 63(2): 69-71. doi:10.1525/bio.2013.63.2.2

Strasser, C, and SE Hampton. 2012. The Fractured Lab Notebook: Undergraduates and Ecological Data Management Training in the United States. *Ecopshere* 3:art116. doi:10.1890/ES12-00139.1

Hampton, S, J Tewksbury and CA Strasser. 2012. Ecological data in the Information Age. Guest Editorial in *Frontiers in Ecology and the Environment* 10:59. doi:10.1890/1540-9295-10.2.59

APPENDIX 5: CONFLICTS OF INTEREST / SOURCES OF BIAS

We cannot identify any conflicts of interest or sources of bias on the part of the primary investigator, all key project staff, and the University of California.

APPENDIX 6: ATTENTION TO DIVERSITY

UC is committed to achieving diversity in the classroom, research lab and workplace. It strives to establish a climate that welcomes, celebrates and promotes respect for the contributions of all students, staff and faculty. In September 2007, the University of California Board of Regents adopted a Diversity Statement (see http://ucnet.universityofcalifornia.edu/working-at-uc/our-values/diversity.html) calling on the University to “seek to achieve diversity among its student bodies and among its employees” and acknowledging “the acute need to remove barriers to the recruitment, retention, and advancement of talented students, faculty, and staff from historically excluded populations who are currently underrepresented.” As part of the University of California, the California Digital Library upholds and promotes the actions of the Diversity Statement. The project will use the tools supplied by the University, particularly the Guidance promoted by the Office of General Counsel, to engage with diverse communities throughout all activities of the proposed project. Specifically, we will engage with diverse audiences as we develop functional requirements to ensure what we developed can be effectively used by diverse communities. In addition we will pay careful attention that the user interface responds to a range of disciplines and user types.

APPENDIX 7: ORGANIZATIONAL PROFILE

The CDL was founded by the University of California in 1997 to take advantage of emerging technologies that were transforming the way digital information was being published and accessed. CDL's diverse and talented staff have assembled one of the world's largest digital research libraries and have changed the ways that faculty, students, and researchers discover and access information. Our vision is to elevate the digital library for UC so that it becomes "expansively global and deeply local".

The CDL's stated values are innovation, collaboration, openness, sharing, privacy, and learning. As the "11th University Library," the CDL provides services to all ten University of California campuses. The CDL's constituency extends from the UC community of over 170,000 scholars and staff, to the 37 million citizens of California, and beyond. The CDL's web-based services are available to internet users worldwide.

As part of the CDL, the UC Curation Center (UC3) has relationships with a wide range of researchers across the University, including natural history museums, research groups, individual scientists and faculty members. UC3 has focused on developing tools and services that serve the UC community throughout the research and data life cycles. The DMPTool aids researchers in creating data management plans to meet requirements of the National Science Foundation and other funding agencies. DataUp is an open source tool that addresses the problem of curation and preservation of spreadsheets created using Excel. In partnership with the international DataCite organization, UC3 is able to provide identifiers for digital objects via the EZID service. Long-term identifiers provide the mechanism behind data citation, which in turn powers data sharing and re-use, as well as allowing others to provide credit and attribution for the original researcher. The Merritt repository facilitates preservation and curation of digital assets for the broader community. One example of a project using Merritt is the DataShare project, which is a collaboration with UC San Francisco, which encourages researchers to share their data. Merritt also serves as a member node in the DataONE research grid and provides preservation services for articles published in CDL’s eScholarship, the Online Archive of California, and Calisphere.

APPENDIX 8: LETTERS OF SUPPORT

1. Mercè Crosas, Director of Data Science and Dataverse Project Lead, IQSS, Harvard University
2. Brian Hole, Founder and CEO, Ubiquity Press
3. Michele Kimpton, CEO DuraSpace
4. Elizabeth Marincola, Chief Executive Officer, Public Library of Science - PLOS
5. William Michener, Professor and Director of e---Science, University Libraries, University of New Mexico, Principal Investigator, DataONE (Observation Network for Earth)
6. Todd Vision, Associate Professor, Department of Biology, Associate Director for Informatics, National Evolutionary Synthesis Center