Project Title: Dash Improving Community Repositories for Better Data Sharing

Submitter: Marisa Strong, Application Development Manager, UC Curation Center, California Digital Library, University of California, Office of the President 510-987-0228
marisa.strong@ucop.edu

Team consists of John Chodacki and Perry Willett, Product Managers, Stephen Abrams, Principal Investigator, Marisa Strong, Technical Development Managr, Scott Fisher, Lead Front-end developer, David Moles, Lead Backend Developer, Bhavitavya Vedula, Developer, John Kratz, UI/UX Designer, Joel Hagedorn, Web Production Developer

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

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., figshare, 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

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.

Solution

We are addressing 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.

Collaboration

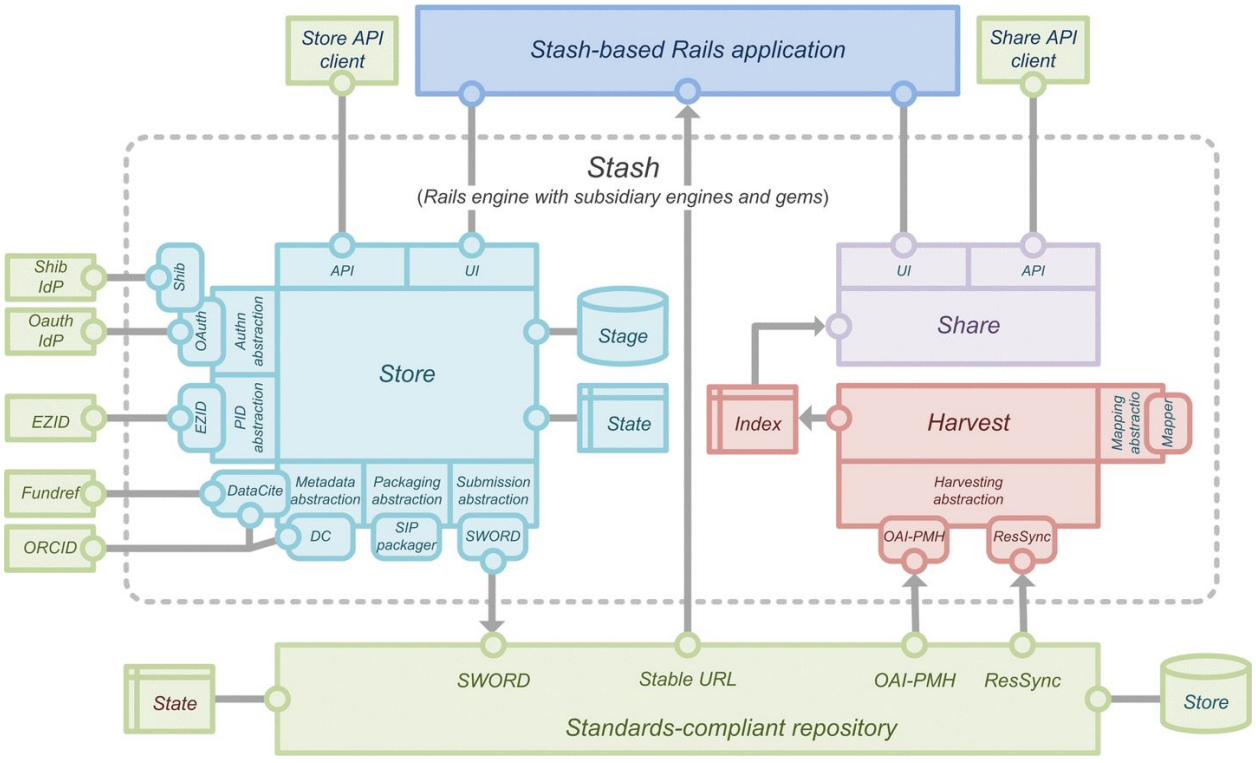
Dash is very much a service that has involved collaboration across campuses, external organizations (DataONE and Orange County Data Portal), and CDL's UI/UX department. Campuses have and will continue to provide feedback via usability testing which will influence an iterative development model. While campus has their own URL and landing page (example: dash.berkeley.edu, dash.ucop.edu, etc.) Dash is a single instance application hosted by CDL.

Deployment Timeline

After initial research into existing platforms and frameworks, Dash development began in earnest in Summer 2015. An agile development methodology was utilized to create user stories which produced the feature set of the Minimum Viable Product (MVP). User feedback is being obtained on the MVP version to assess and refine the features of the tool with iterative development continuing for a production release in Summer 2016.

Technology

Dash utilizes a combination of technologies, the web application itself, hosted on Amazon Web Services Cloud infrastructure, is built on Ruby On Rails framework. It utilizes both Shibboleth and Google authentication mechanisms, provides submission processing to an institutional repository via the SWORD protocol, harvesting is provided via an OAI-PMH protocol, and indexing is supported by SOLR. All of these technologies are implemented modularly to allow for customization and

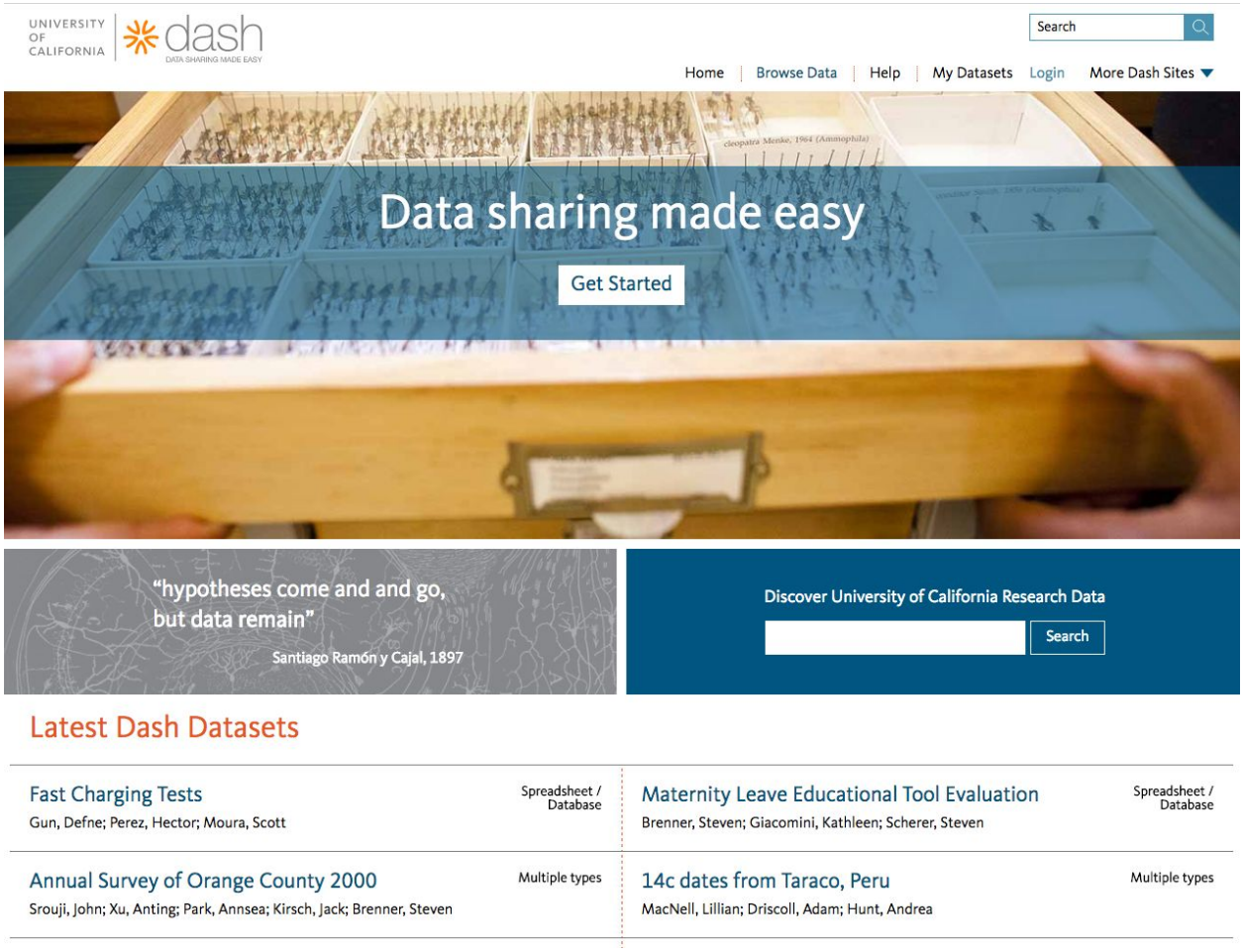


Measuring Project Success

For qualitative assessment, we will incorporate user interviews into Phase 1 above, obtaining researcher feedback on Dash as it develops. Based on 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.



Type of Data

Title of Datasets

Author Name, Affiliation, ORCID

Submission Date:
Published Date:
Publisher:
doi: xxxxx/xxxxxxxx

Cite as

Author Name and Author Name, Title_of_dataset. Type_of_data. Publisher. Version n (DD Month YYYY).
http://dx.doi.org/xxxx.xxxx

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Distinctio quod, eius aliquid, placeat quos nisi consectetur maiores iusto accusamus! Repellat nihil esse quis consequatur mollitia, optio itaque nulla veritatis, natus!

Veritatis tenetur, perspiciatis, commodi numquam repudiandae, voluptas asperiores reprehenderit dolor sint voluptatibus quam quo, itaque quidem nobis. Harum atque, officiis modi ipsum neque, odit, voluptate aspernatur, error porro vitae ad.

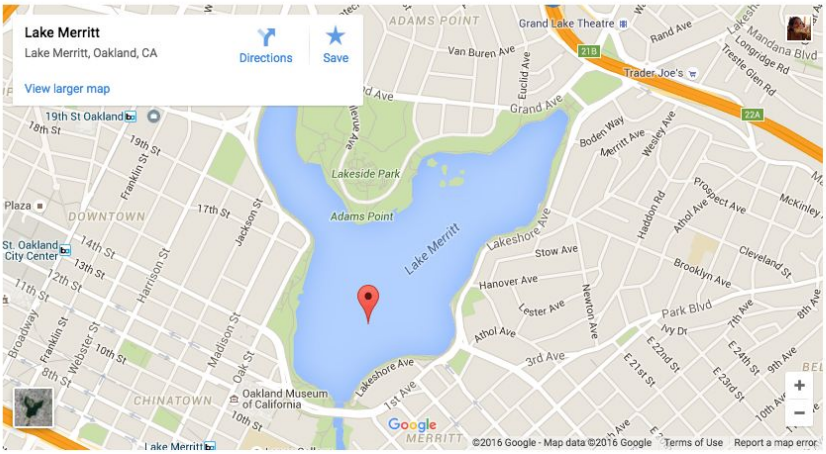
Methods

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Praesentium laborum inventore totam dolorem ipsam in at ratione, odio, atque sequi aliquid. Exercitationem nobis maiores facilis vel repellendus, neque, quas assumenda.

Usage Notes

Lorem ipsum dolor sit amet, consectetur adipisicing elit. Quas eveniet tenetur amet ipsum ullam modi dicta illo repellendus, reiciendis, labore aut minus. Quidem amet architecto odit. Quae, expedita maxime assumenda.

Location



Data Files

- V001 456 kb

Metrics

100 views

50 downloads

Keywords

Keyword
Keyword
Keyword
Keyword
Keyword

License

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/