**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 publication 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 provides 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 are 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, datashare.ucsf.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) production release last Fall 2016. User feedback was obtained on the MVP version to assess and refine the features of the tool with continuing, iterative development. The project continues to provide releases to the service in 2-4 week increments. Development and release iterations can be tracked on the Github project page.

Technology

Dash utilizes a combination of technologies, the web application itself, hosted on Amazon Web Services Cloud infrastructure (EC2 and RDS), is built on a Ruby On Rails framework. Many of the technologies used are open source. Dash utilizes both Shibboleth and Google authentication mechanisms, provides submission processing to the Merritt institutional repository via the SWORD protocol, which in turn exposes metadata for harvesting via the OAI-PMH protocol. The harvested metadata is indexed using SOLR technology with the discovery of datasets and publications provided by a GeoBlacklight portal. Persistent identifiers (DOIs) for assigned utilizing the EZID API, another service designed and implemented at CDL. All of these technologies are implemented modularly to allow for customization of campus and institutional branding, storage upload limits, and defining time periods for time-released publication of datasets.
Measuring Project Success

For qualitative assessment, our product manager has been coordinating with each campus utilizing Dash capturing feedback from both the researchers and libraries. A team of representatives from each campus have made up a Dash User Group that meets regularly to advise on future releases and necessary improvements.

Throughout the project we have captured usage metrics as indicators of Dash adoption and community uptake. Particularly we have monitored metrics with regards to the use of Dash for data publication and access.
Distance decay relationships in foliar fungal endophytes are driven by rare taxa

Oono, Ryoko, UC Santa Barbara
Lefevre, Emilie, Duke University
Rasmussen, Anna, UC Santa Barbara
Publication date: March 23, 2017
Publisher: DataONE
https://doi.org/10.15146/R3Q534

Citation
Oono, Ryoko; Lefevre, Emilie; Rasmussen, Anna (2017), Distance decay relationships in foliar fungal endophytes are driven by rare taxa, Text, https://doi.org/10.15146/R3Q534

Abstract
Foliar fungal endophytes represent a diverse and species-rich plant microbiome. Their biogeography provides essential clues to their cryptic relationship with hosts and the environment in which they disperse. We present species composition, diversity, and dispersal patterns of endophytic fungi associated with needles of Pinus taeda trees across regional scales in the absence of strong environmental gradients as well as within individual trees. An empirical designation of rare and abundant taxa enlightens us on the structure of endophyte communities. We report multiple distance-decay patterns consistent with effects of dispersal limitation, largely driven by community changes in rare taxa, those taxonomic units that made up less than
APPENDIX 2: BIBLIOGRAPHY


