code{4}lib JOURNAL

Issue 49, 2020-08-10

Open Source Tools for Scaling Data Curation at QDR

This paper describes the development of services and tools for scaling data curation services at the Qualitative Data Repository (QDR). Through a set of open-source tools, semi-automated workflows, and extensions to the Dataverse platform, our team has built services for curators to efficiently and effectively publish collections of qualitatively derived data. The contributions we seek to make in this paper are as follows:

1. We describe 'human-in-the-loop' curation and the tools that facilitate this model at QDR;

2. We provide an in-depth discussion of the design and implementation of these tools, including applications specific to the Dataverse software repository, as well as standalone archiving tools written in R; and

3. We highlight the role of providing a service layer for data discovery and accessibility of qualitative data.

Keywords: Data curation; open-source; qualitative data

by Nicholas Weber, Sebastian Karcher, and James Myers

Introduction

The automation of curation tasks – including portions of the ingest and transformation of submitted data to a repository – is crucial for scaling services to researchers that are managing increasingly complex and large data collections [1]. However, the diverse characteristics of qualitative data often limit the applicability of fully automated tools. Moreover, when data include privacy sensitive information, many automated processes require a curator's intervention to mediate, check quality, and evaluate compliance with responsible practices for collecting and properly handling personally identifiable information (PII) [2].

At the Qualitative Data Repository (QDR), we have started to design curatorial interventions that act as a "human-in-the-loop" to the automation and verification of different components of our data repository infrastructure. Practically, this takes the form of tools that handle mundane preparation of data at the point of ingest and provide graphic user interfaces to APIs that allow curators to publish different aspects of a data collection unique to QDR. In the following paper, we describe four different initiatives and the way that QDR uses this model of human-in-the-loop curation to efficiently publish qualitative data at scale. The three contributions we seek to make in this paper are as follows:

1. We first describe 'human-in-the-loop' curation and the tools that facilitate this model at QDR;
2. We then provide an in-depth discussion of the design and implementation of these tools, including applications specific to the Dataverse software repository, as well as standalone archiving tools written in R; and
3. Finally, in describing each tool we briefly highlight the role of providing a service layer for data discovery and accessibility of qualitative data at QDR.

Human-in-the-Loop Curation

In a repository setting, data curation often begins with the transform of a submission information package – including data, metadata, and information documentation – into an archival information package. This transformation (from SIP to AIP) may include manual or semi-automated tasks performed by a curation staff, such as data normalization, cleaning, and file migration, as well as fully automated workflows that can ingest, perform preservation actions, and prepare an archival information package for reliable long-term storage.

Reducing the number of manual tasks that need to be carried out by curators is an important and long-standing challenge for repository developers [3]. Early innovations with workflow automation, such as rule-based systems for automating preservation tasks [4], were important preemptive steps towards automation in a curation and preservation setting. Further, by building assistive technologies that can support practical curation work repositories can reduce the time, effort, and financial costs of publishing data collections. Enabling automation of manual tasks also enables curation services to scale with the complexity and size of data that are generated by researchers.

Despite the promise of large-scale automation, a number of scenarios require the purposeful intervention of a curator. These include consultation with depositors as well as the ingest and publication of privacy sensitive data that contains personally identifiable information about human participants. These scenarios require deliberate decisions that are heavily dependent upon the context and content of data.

A key emerging challenge is the development of assistive curation technologies that can take advantage of affordances offered by workflow automation, and enable curation interventions [5].

Human-in-the-loop systems attempt to do exactly this – they provide a means for human curators to avoid (or at least significantly reduce) the amount of time spent on repetitive, automatable tasks while keeping the entire curatorial process under human supervision. A crucial component of human-in-the-loop curation is its modularity: individual components of the curation process can be removed from automation should data or context call for a manual approach. In this vein, we see the development of tools that can effectively help curators to manage deposits, transform and package data, and provide reliable means to prepare data for long-term preservation as vital to the future sustainability of large scale data collections. In the following section, we review a set of open-source tools developed to assist curators at the Qualitative Data Repository. For each tool, we describe its development goal, the way in which the tool helps curators perform curation tasks, and the relation to QDR's overall archiving workflow.

Overview of Tools and Approach to facilitating Human-in-the-Loop Curation at QDR

In the following section, we review four tools that help QDR to practice human-in-the-loop curation: dvCurator, a tool to automate processing and documentation of data deposits; ArchivR, a tool that automates the transformation of hyperlinks into persistent web-identifiers; Annotation-Fetcher, an interface for the Dataverse platform that allows web-based annotations to be archived; and, dv-Previewers, a set of technologies that allow multimedia files to be viewed within a repository.

Data Ingest: dvCurator [6]

One of the rote tasks that data curators at QDR must perform is the copying of submitted data projects to local machines for manual curation tasks (e.g. checking for PII anonymization, reformatting files, changing file names to a common standard, etc) and to systematically document and track those steps within a multi-person curation team. To automate portions of this initial process, we have written a series of R scripts that will simultaneously fetch submitted data from QDR's Dataverse, create local and back-up copies in a cloud storage environment, log a set of general issues on our Github repository for tracking curation workflows, and then extract metadata for creating a compliant .tsv to re-upload to QDR's Dataverse at publication.

dvCurator's main function `start_curation()` initiates the process of curation by creating GitHub issues for standard curation tasks and associating them with a Github project for the duration of the data project. The Github issues track individual curation tasks using checkboxes to track and issue comments to document curation steps. The project board, a simple Kanban-style board with three columns, allows for quick assessment of curation status and access to individual issues (See figure 1).

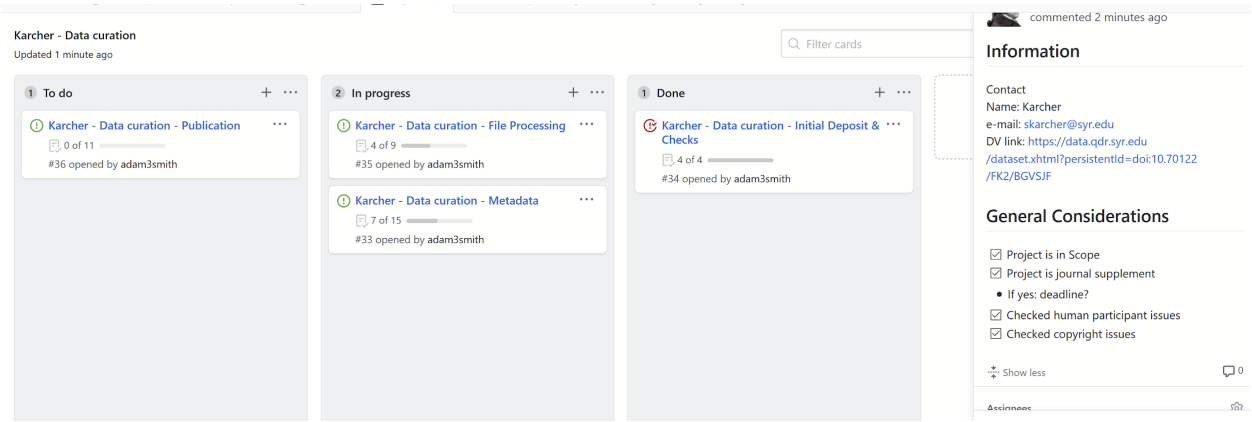

Figure 1. Github curation project board for a sample data project

dvCurator then automatically creates a local curation folder (on the curators desktop) with subfolders for Original Deposit and QDR Prepared versions of the dataset. Next, using the Dataverse API dvCurator fetches a .zip file for the full data project and populates the Original Deposit folder and an unzipped version to QDR Prepared for further curation. This includes a list of files with associated metadata as a .csv file to the same folder.

The tool also implements an entirely separate function, `datasets_byDate()` that creates an .html file with citations of all datasets published within a specified date range, designed to facilitate reporting for curation staff.

Collectively these functions allow QDR to automate most of the pre-processing of a data collection and ensure consistent application of steps for beginning a curation protocol for newly deposited data at QDR.

Data Preparation: ArchivR [7]

Submission Information Packages at QDR often contain links and references to diverse web resources. This may include, for example, policy documents that live at a particular web address, or gray literature such as blogs and other informal communications. As an authoritative repository tasked with the long-term accessibility of both data and documentation related to scholarly research QDR requires web-archiving for these resources.

ArchivR is a project developed by QDR to create persistent copies of cited web-resources and transform web-links with data documentation into persistent identifiers. The basic function is `archive` that takes a list of URLs and stores them in the Internet Archive's 'Wayback Machine' for long-term archiving. Given a URL to a web-resource, ArchivR will return a dataframe in R containing the callback data for the service. The R dataframe consists of, simply, the service used and the corresponding new archived identifier (e.g. in the code snippet below the URL 'www.example.com' is archived in the 'Way Back Machine' and a dataframe containing the new identifier is returned).

```
arc_df <- archiv(list("www.example.com", "NOTAURL", "www.github.com"))
arc_df$way_back_url
#                                                  wayback_url
# 1 http://web.archive.org/web/20190128171132/http://www.example.com
# 2                                                url not found
# 3    http://web.archive.org/web/20190128171134/https://github.com/...
```

Code Snippet 1. An example R dataframe returned by ArchivR

ArchivR will also archive all URLs in a text file (including docx, pdf and plain text formats such as html, xml, and markdown). To allow for pre-processing of URLs before archiving, Archivr also provides access to the functions used to extract URLs from a webpage (`extract_urls_from_webpage("URL")`) and from a file (`extract_urls_from_text("filepath")`), and includes an except argument as part of `archive` that excludes a URL pattern, defined by a regular expression from archiving (this could be used, for example, to avoid unnecessarily archiving digital object identifiers (DOIs). ArchivR also works with perma.cc's archiving service – and QDR principally uses perma.cc for link archiving. An API key is required for this service[8].

The ArchivR tool allows QDR to prepare data collections with hyperlinked text and web-resources to be automatically archived using available services (including for projects using hundreds of web sources such as [9] – allowing curators to focus time and attention on other publishing needs, while simultaneously guaranteeing that content in data collections remain persistent and accessible for the long-term.

Data Publication: Annotation Fetcher

Over the last two years QDR has been involved in a collaboration with the platform Hypothes.is [10] to extend the capability of using web-annotations to facilitate transparent social science research. The collaboration, Annotation for Transparent Inquiry (ATI), allows authors to persistently link data to existing or planned digital publications with a web-annotation, and archive related files (datasets, source documents, or even software) in QDR's Dataverse [11].

Curating ATI projects currently requires data curators at QDR to manually transform comments made in Microsoft Word or PDF documents into linked annotations (in HTML) using the Hypothesis web client. The annotations, when displayed in a browser, are linked to a rendered web page (see Figure 2).

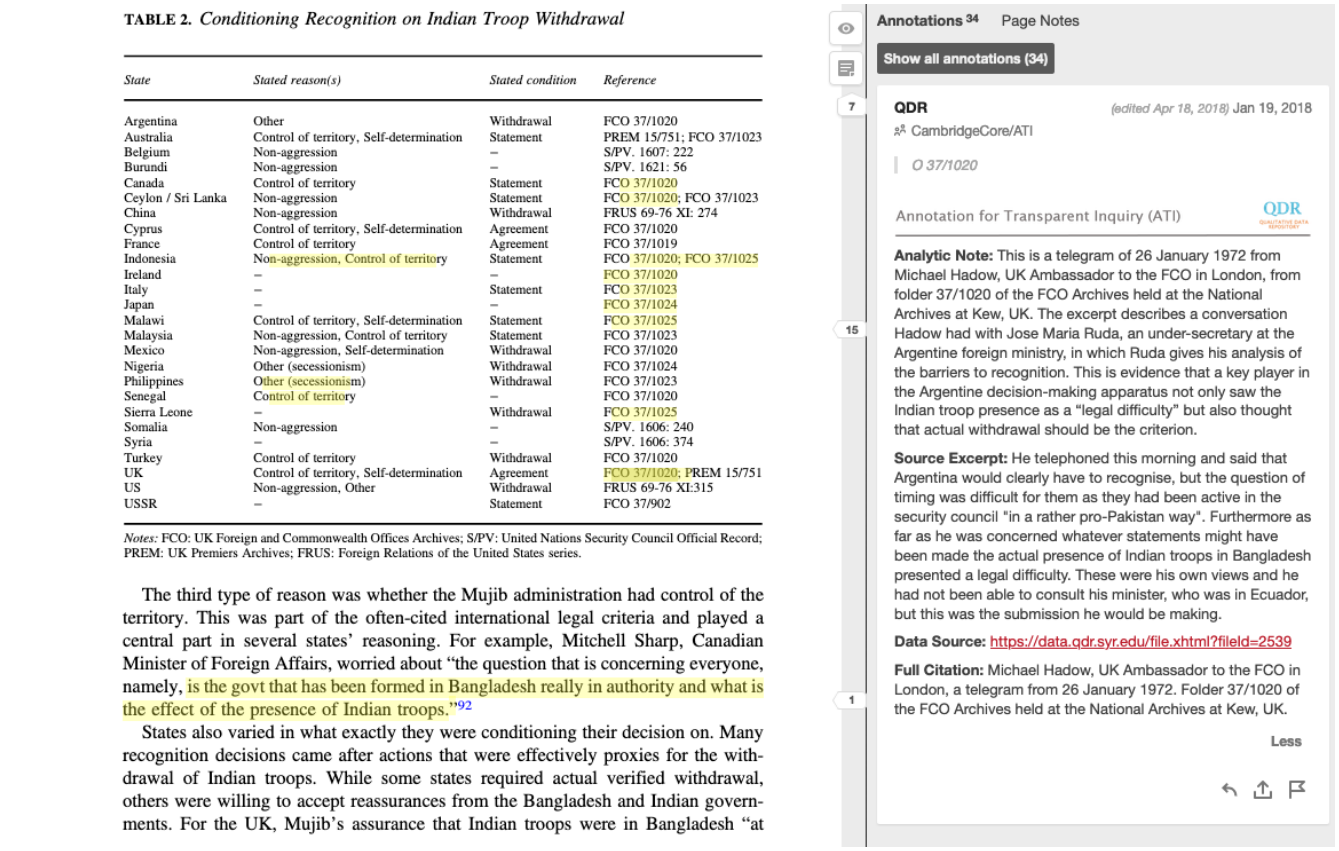

TABLE 2. *Conditioning Recognition on Indian Troop Withdrawal*

State	*Stated reason(s)*	*Stated condition*	*Reference*
Argentina	Other	Withdrawal	FCO 37/1020
Australia	Control of territory, Self-determination	Statement	PREM 15/751; FCO 37/1023
Belgium	Non-aggression	–	S/PV. 1607: 222
Burundi	Non-aggression	–	S/PV. 1621: 56
Canada	Control of territory	Statement	FCO 37/1020
Ceylon / Sri Lanka	Non-aggression	Statement	FCO 37/1020; FCO 37/1023
China	Non-aggression	Withdrawal	FRUS 69-76 XI: 274
Cyprus	Control of territory, Self-determination	Agreement	FCO 37/1020
France	Control of territory	Agreement	FCO 37/1019
Indonesia	Non-aggression, Control of territory	Statement	FCO 37/1020; FCO 37/1025
Ireland	–	–	FCO 37/1020
Italy	–	Statement	FCO 37/1023
Japan	–	–	FCO 37/1024
Malawi	Control of territory, Self-determination	Statement	FCO 37/1025
Malaysia	Non-aggression, Control of territory	Statement	FCO 37/1023
Mexico	Non-aggression, Self-determination	Withdrawal	FCO 37/1020
Nigeria	Other (secessionism)	Withdrawal	FCO 37/1024
Philippines	Other (secessionism)	Withdrawal	FCO 37/1023
Senegal	Control of territory	–	FCO 37/1020
Sierra Leone	–	Withdrawal	FCO 37/1025
Somalia	Non-aggression	–	S/PV. 1606: 240
Syria	–	–	S/PV. 1606: 374
Turkey	Control of territory	Withdrawal	FCO 37/1020
UK	Control of territory, Self-determination	Agreement	FCO 37/1020; PREM 15/751
US	Non-aggression, Other	Withdrawal	FRUS 69-76 XI:315
USSR	–	Statement	FCO 37/902

Notes: FCO: UK Foreign and Commonwealth Offices Archives; S/PV: United Nations Security Council Official Record; PREM: UK Premiers Archives; FRUS: Foreign Relations of the United States series.

The third type of reason was whether the Mujib administration had control of the territory. This was part of the often-cited international legal criteria and played a central part in several states' reasoning. For example, Mitchell Sharp, Canadian Minister of Foreign Affairs, worried about "the question that is concerning everyone, namely, is the govt that has been formed in Bangladesh really in authority and what is the effect of the presence of Indian troops."[92]

States also varied in what exactly they were conditioning their decision on. Many recognition decisions came after actions that were effectively proxies for the withdrawal of Indian troops. While some states required actual verified withdrawal, others were willing to accept reassurances from the Bangladesh and Indian governments. For the UK, Mujib's assurance that Indian troops were in Bangladesh "at

Figure 2. Annotations made to a scholarly article. In the left panel, an HTML rendering of the text that appears on the web has been annotated using the hypothesis client that appears on the right. Annotation content is displayed within the right-hand panel.

The HTML-based web-annotations contain linked documents and other data sources that are stored in QDR's Dataverse. However, the content of the annotations (as we as the anchored text that is being annotated) remains external to QDR as HTML web pages. To retrieve and store this content we designed 'Annotation Fetcher' – a GUI interface tool for Dataverse that allows a curator to call the hypothes.is API, retrieve the annotation content and anchors (as a JSON file), and archive the annotations within a Dataverse. Since annotations are stored as plain-text (retrieved via the Hypothes.is API as JSON) they can also be indexed, and searched for within QDR's Dataverse as part of a data collection (containing the original dataset, the annotations, and any documentation that is related to the collection) – thus transforming annotations into first-class scholarly research objects that are not only discoverable but also citable through Dataverse [12].

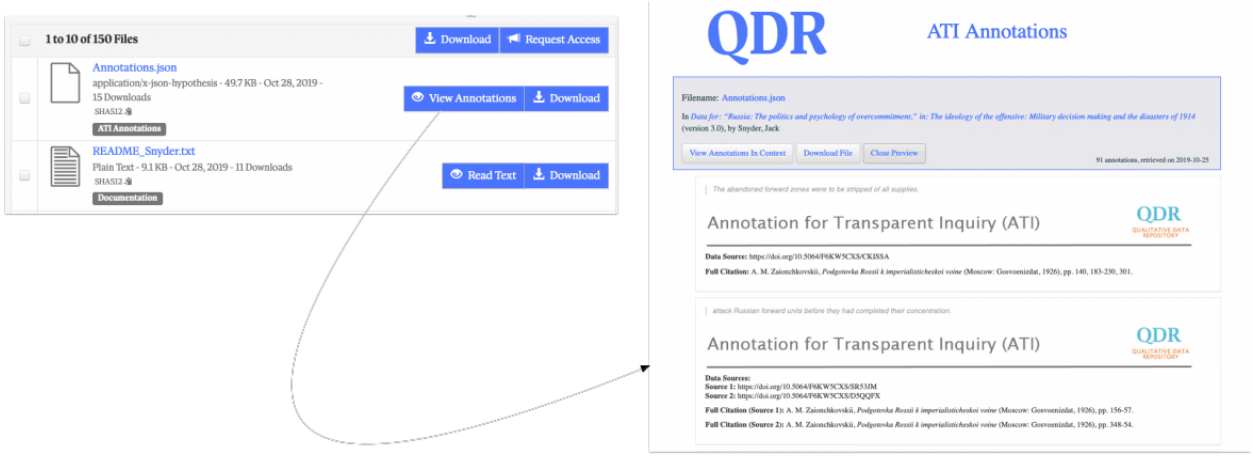

Figure 3. Annotations archived in QDR's Dataverse can be searched for and discovered as part of a data collection. Using the DV-Previewer, the annotation content and anchor can also be viewed within Dataverse.

Data Discovery: DV-Previewer [13]

Until recently, the only way for end-users to access images, audio recordings, videos, and other multi-media in a Dataverse-based repository, was to download and open files on a local desktop. This results in an undue burden for end-users when browsing or exploring new data collections. To improve access and discovery QDR developed a set of data previewers that conform to the Dataverse external tools interface. DR-Previewer tools are lightweight wrappers around standard HTML5 functionality (e.g. audio, video), or third-party libraries (pdf) or some combination (e.g. standard image displays with a third-party library to allow zooming, simple text/html displays with third-party libraries used to sanitize content to avoid security issues).

Previewers can be used without individual data repositories needing to download and configure local copies, but by simply running a set of curl commands to register the previewer with a local Dataverse instance – and there is just one command per mime-type (i.e. multiple commands to cover different types of images [14].) Importantly, to preview restricted content, a user must have permission to view the relevant dataset version and download the relevant file and must have created an API Token for themselves in Dataverse. At QDR we provide a simple button to generate an appropriately credentialed API Token corresponding to the registered user's access rights. More recently, the DVPreviewers have been adopted by the Global Dataverse Community Consortium and are now being used in multiple Dataverse installations.

In combination, both DV-Previewer and Annotation Fetcher allow curation staff to easily implement and make available unique data sources for viewing and consumption.

Conclusion

Through the development of open-source tools to facilitate meaningful curatorial interventions, we remain optimistic that challenges in preserving qualitative data, including privacy and protecting sensitive data can be overcome. This paper has focused in particular on ways in which the development team at QDR is using tools in conjunction with the Dataverse repository platform to facilitate "human-in-the-loop" curation tasks that enable efficient, reproducible, and contextually meaningful work with data. By off-loading repetitive tasks, and creating easy access to API methods, QDR's curation staff can focus time and attention on preparing sensitive data for publication, and preserving valuable data collections over the long-term.

Notes

[1] Plale, B. A., Dickson, E., Kouper, I., Liyanage, S. H., Ma, Y., McDonald, R. H., … & Withana, S. (2019). Safe Open Science for Restricted Data. *Data and Information Management*, *3*(1), 50-60.

[2] Foster, I. (2018). Research infrastructure for the safe analysis of sensitive data. *The Annals of the American Academy of Political and Social Science*, *675*(1), 102-120.

[3] Blanke, T., & Hedges, M. (2013). Scholarly primitives: Building institutional infrastructure for humanities e-Science. *Future Generation Computer Systems*, *29*(2), 654-661.

[4] Westerlund, P., Andersson, I., Päivärinta, T., & Nilsson, J. (2019). Towards automated pre-ingest workflow for bridging information systems and digital preservation services. *Records Management Journal*.

[5] Foster, I. (2018)

[6] Repository with code and documentation for dvCuratoR: https://github.com/QualitativeDataRepository/dvcurator

[7] Repository with code and documentation for ArchivR: https://github.com/QualitativeDataRepository/archivr

[8] Information on how to obtain an API key for Perma.cc can be found in the Developer Documentation: https://perma.cc/docs/developer#api-key

[9] Wedeen, L. (2019). Data for: Authoritarian apprehensions: Ideology, judgment, and mourning in Syria [Data set]. QDR Main Collection. https://doi.org/10.5064/F63776W4

[10] Hypothes.is https://web.hypothes.is/

[11] Elman, C., & Kapiszewski, D. (2018). The Qualitative Data Repository's Annotation for Transparent Inquiry (ATI) Initiative. *PS: Political Science & Politics*, *51*(1), 3-6., Karcher, S., & Weber, N. (2019). Annotation for transparent inquiry: Transparent data and analysis for qualitative research. *IASSIST Quarterly*, *43*(2), 1-9.

[12] As an additional example, the following project contains an annotation file that is discoverable on the QDR Dataverse, and is also previewable within a browser:

Snyder, Jack. 2015. "Data for: “Russia: The politics and psychology of overcommitment,” in: The ideology of the offensive: Military decision making and the disasters of 1914". Qualitative Data Repository. https://doi.org/10.5064/F6KW5CXS. QDR Main Collection. V3

[13] Repository with dv Previewers code and installation instructions: https://github.com/QualitativeDataRepository/dataverse-previewers

[14] For specific installations see https://github.com/QualitativeDataRepository/dataverse-previewers#installation

About the Authors

Nicholas Weber (nmweber@uw.edu) is an Assistant Professor at the University of Washington, and the Technical Director at the Qualitative Data Repository. He holds an MLIS and Ph.D. in Information Science from the University of Illinois. His research focuses on the design, implementation, and use of data infrastructures, and scientific software sustainability.

Sebastian Karcher (skarcher@syr.edu) is the Associate Director of the Qualitative Data Repository and a Research Assistant Professor in the Department of Political Science at Syracuse University. He holds a Ph.D. in Political Science from Northwestern University. His research interests include qualitative data management and the development of scholarly workflows.

James Myers (qqmyers@hotmail.com) is scientific software researcher and developer, and currently serves as the lead developer at the Qualitative Data Repository. He holds a Ph.D. in Experimental Chemistry from the University of California, Berkeley. Over the last two decades, Jim has worked in applied computer science, creating cutting-edge scientific "cyberinfrastructure" and working to understand the socio-technical issues that contribute to the ultimate success or failure of scientific software efforts.