Research software and The Dataverse® Project

What is Dataverse?

The Dataverse® Project

Open source research data repository software

Dataverse®

Dataverse is an open source web application for sharing, citing, analyzing, and preserving research data (developed by the Data Science and Products team at the Institute for Quantitative Social Science and the Dataverse community).

dataverse.org is our home on the web and shows a map of Dataverse installations around the world, a list of features, integrations that have been made possible through REST APIs, our development roadmap, and more.

We maintain a demo site at demo.dataverse.org which you are welcome to use for testing and evaluating Dataverse.

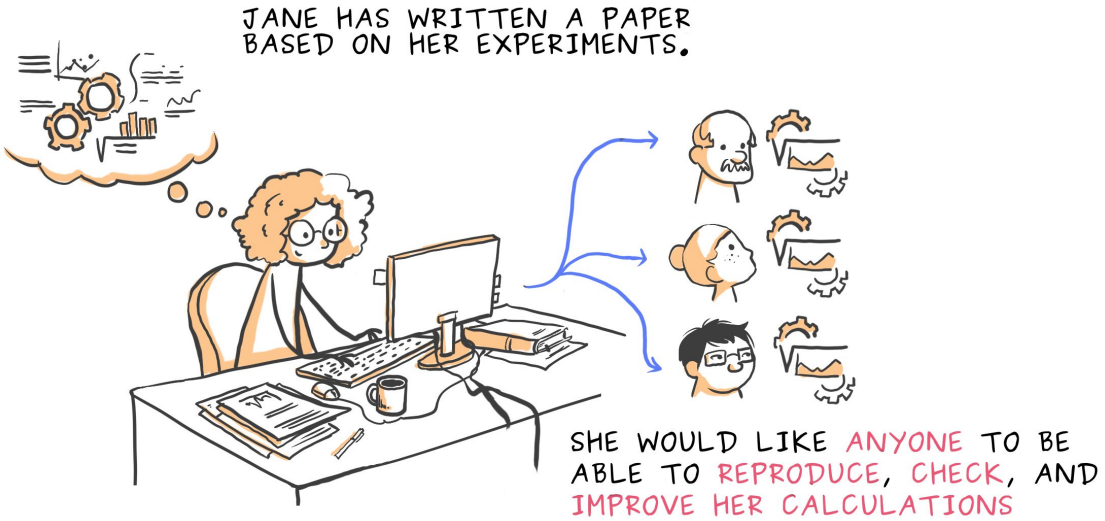
To install Dataverse, please see our Installation Guide which will prompt you to download our latest release.

To discuss Dataverse with the community, please join our mailing list, participate in a community call, chat with us at chat.dataverse.org, or attend our annual Dataverse Community Meeting.

We love contributors! Please see our Contributing Guide for ways you can help.

https://github.com/IQSS/dataverse

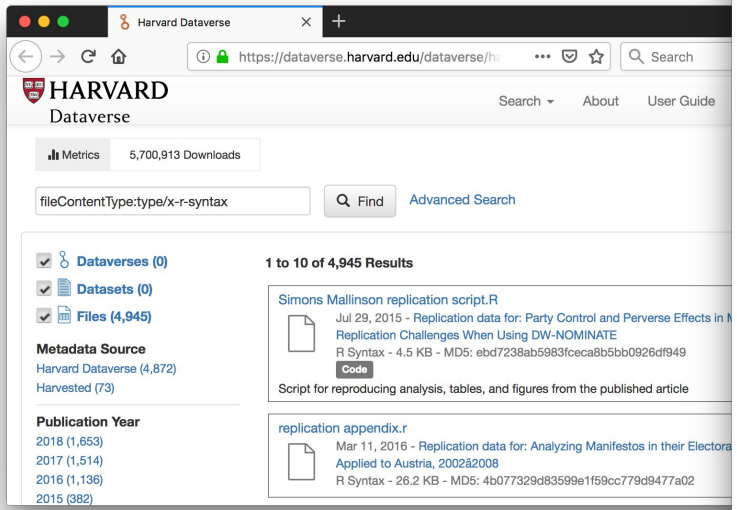
Agenda

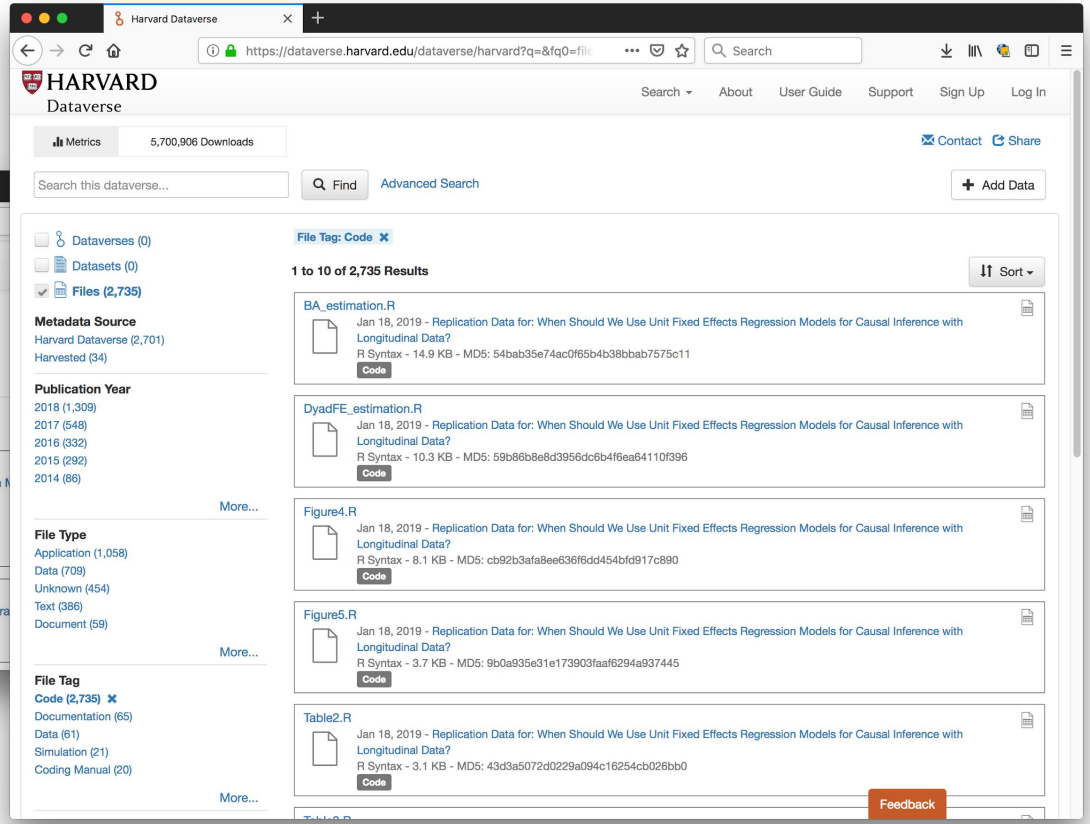


Code deposit

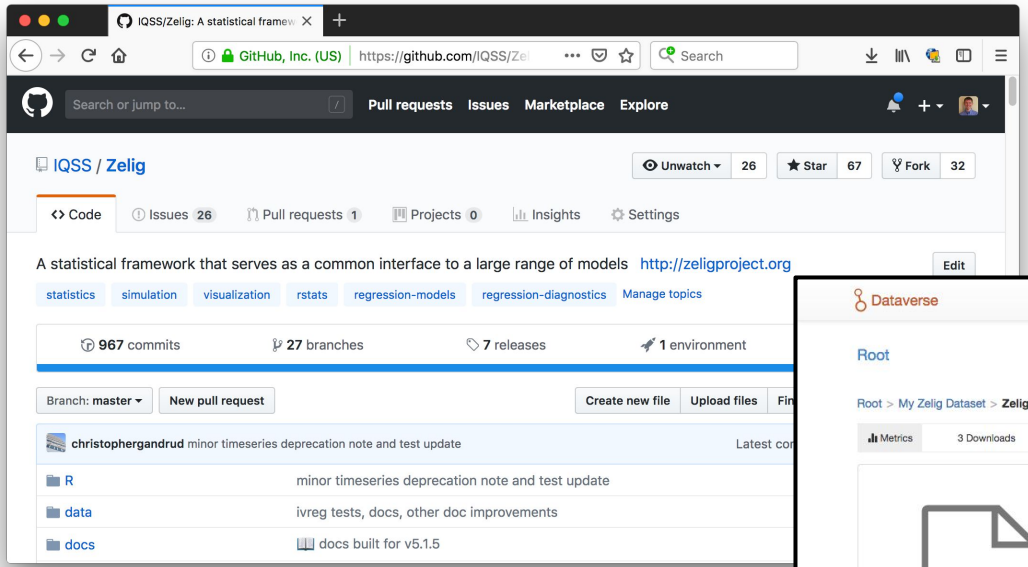
Reproducibility

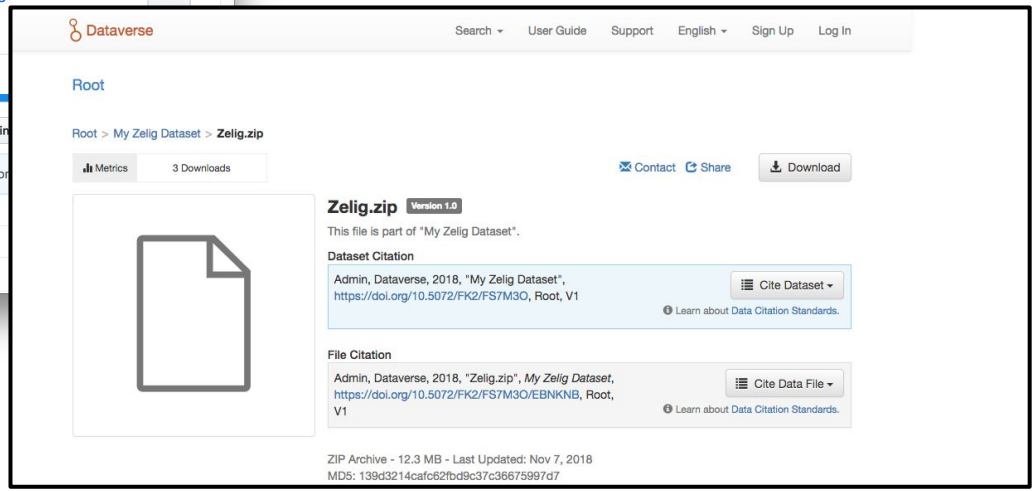
Code already in Dataverse





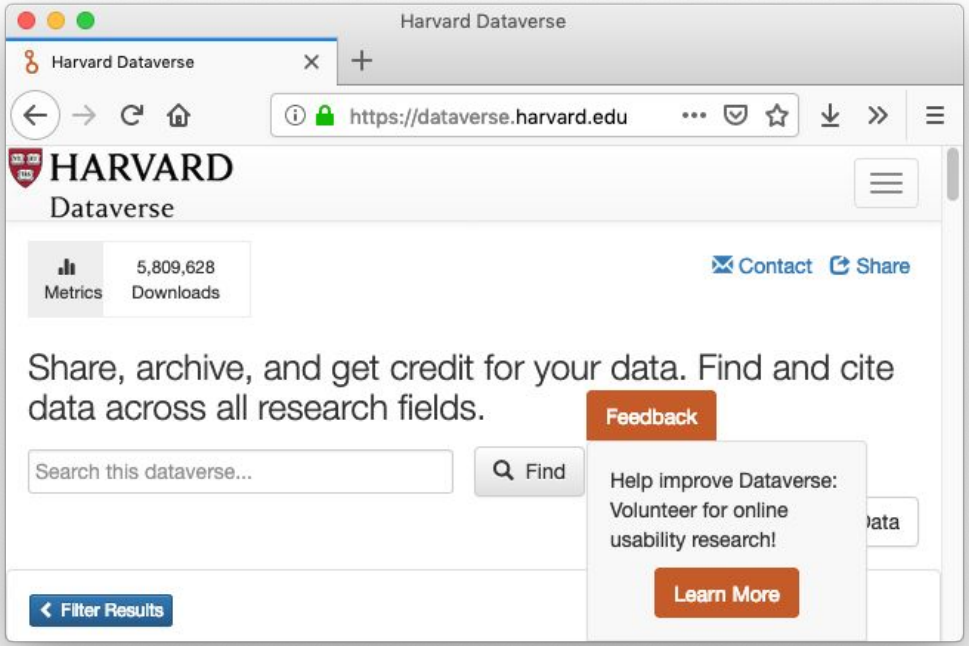
Import from GitHub prototype

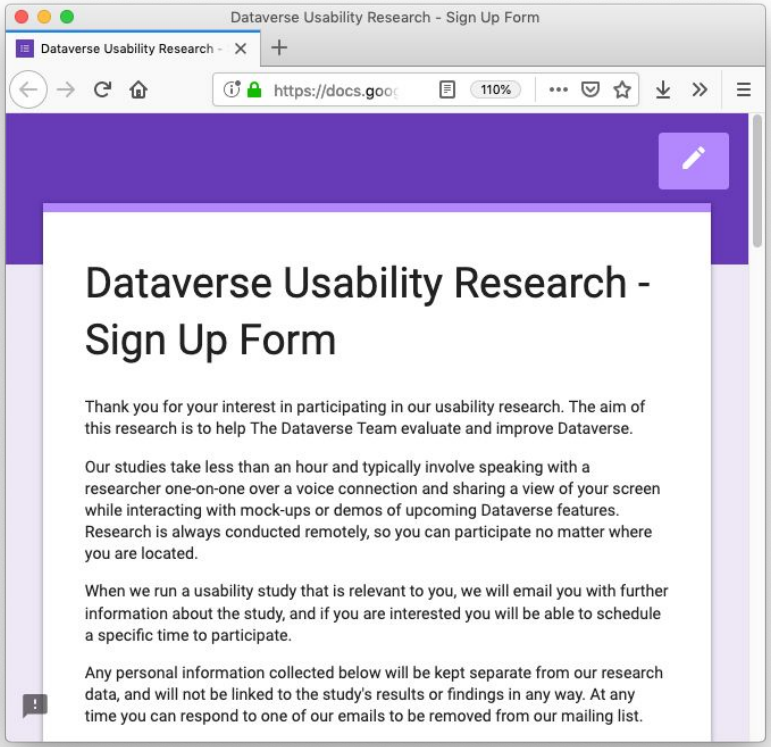




https://github.com/IQSS/dataverse/issues/2739

Usability testing





https://dataverse.harvard.edu

Software metadata

Files | Metadata | Terms | Versions

Search this dataset... Find

1 to 10 of 29 Files

Download

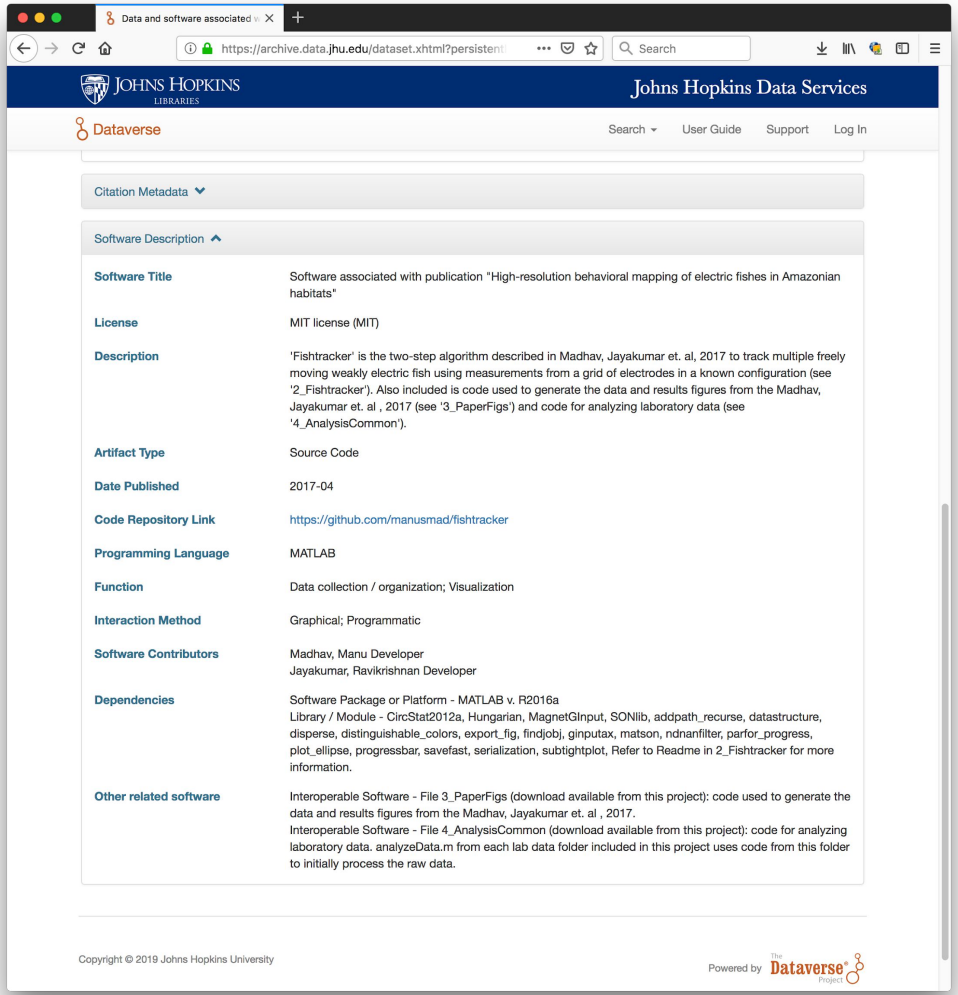
1_Readme_and_DataFormats.zip
ZIP Archive - 64.1 KB - Sep 19, 2017 - 49 Downloads
MD5: a4de02435e7eec49c422e1d45cb60758
Contains the main readme file, the data formats file, which describes the different types of data found within the other files, and a metadata html file documenting the software included with the project.
Documentation
Download

2_Fishtracker.zip
ZIP Archive - 15.2 MB - Sep 19, 2017 - 46 Downloads
MD5: bb5a1d3d3e32461796e4ee0c345472b9
Codebase for the fish tracking algorithm described in the paper. (Readme enclosed)
Code
Download

3_PaperFigs.zip
ZIP Archive - 83.3 KB - Sep 19, 2017 - 46 Downloads
MD5: fed6b315f84284945bca1cbb7070e643
Code used to generate the data and results figures from the paper. (Readme enclosed)
Code
Download

4_AnalysisCommon.zip
ZIP Archive - 388.9 KB - Sep 19, 2017 - 43 Downloads
MD5: 87867b1a13af3ba280fc3499bb8a1765
Common code for analyzing laboratory data. analyzeData.m from each lab data folder (File 6-02, 7-02, 8-02 and 9-02) uses code from this folder to initially process the raw data. (Readme enclosed)
Code
Download

http://datacurationnetwork.org



CodeMeta

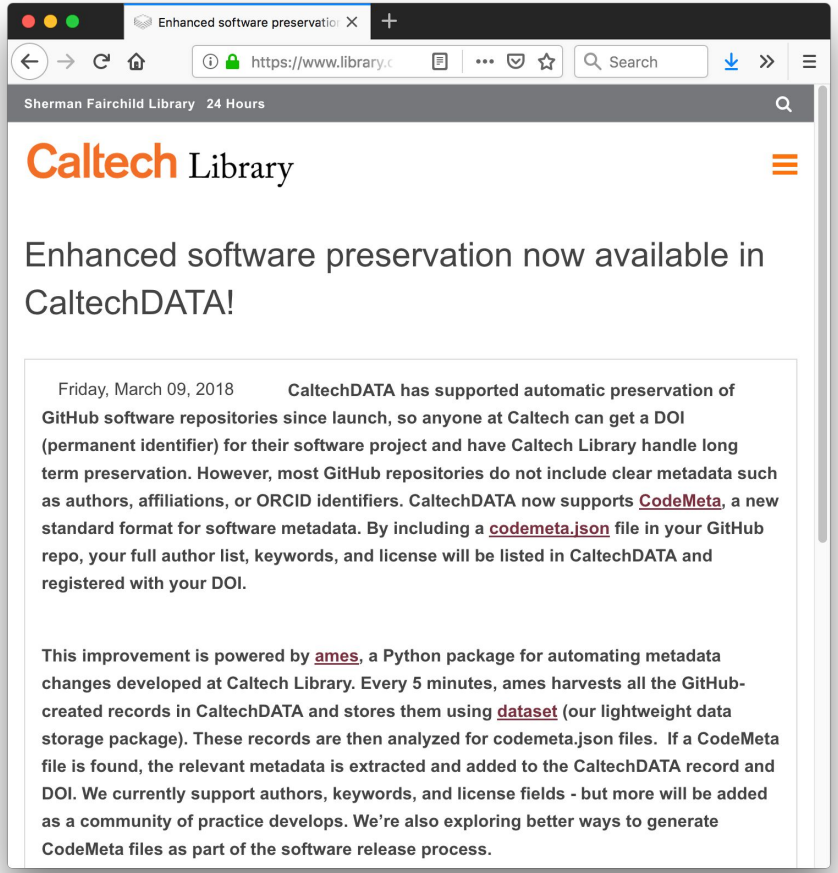
DataCite
FIND, ACCESS, AND REUSE DATA

Supported Content Types

DataCite supports a number of metadata content types:

Format	Content Type
RDF XML	application/rdf+xml
RDF Turtle	text/turtle
Citeproc JSON	application/vnd.citationstyles.csl+json
Schema.org in JSON-LD	application/vnd.schemaorg.ld+json
Codemeta	application/vnd.codemeta.ld+json
Formatted text citation	text/x-bibliography
RIS	application/x-research-info-systems
BibTeX	application/x-bibtex
DataCite XML	application/vnd.datacite.datacite+xml
DataCite JSON	application/vnd.datacite.datacite+json

https://support.datacite.org/docs/datacite-content-resolver

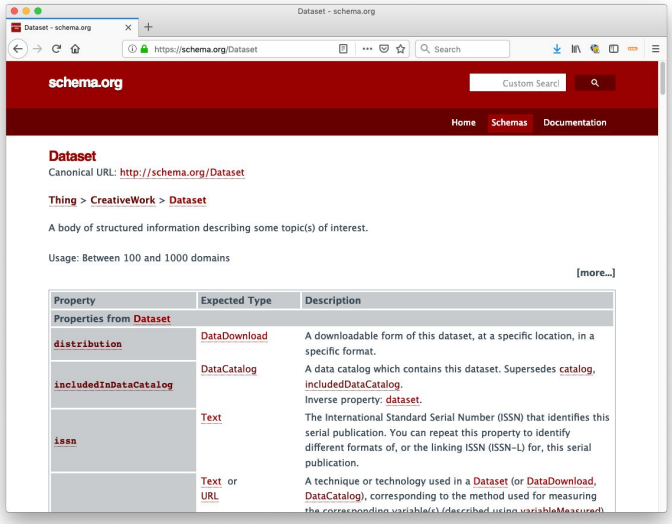


https://www.library.caltech.edu/news/enhanced-software-preservation-now-available-caltechdata

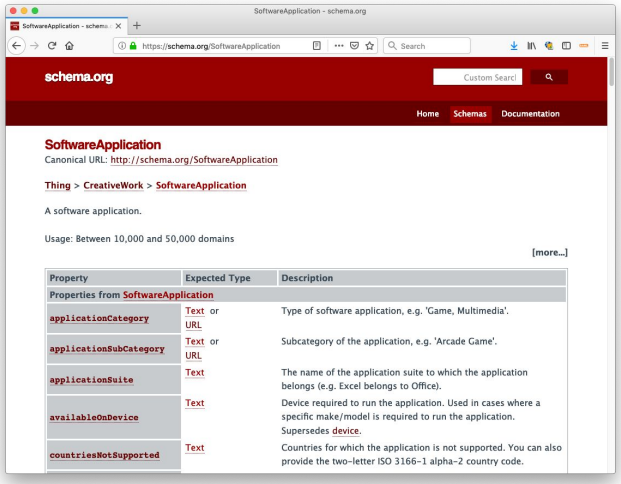
Schema.org JSON-LD

Google Dataset Search Beta

Search for Datasets

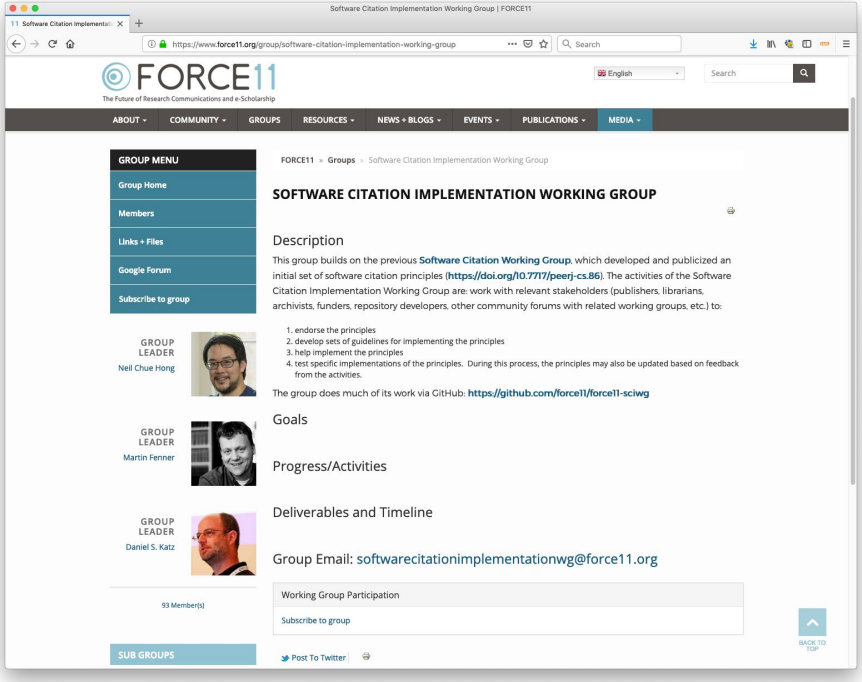


https://schema.org/Dataset



https://schema.org/SoftwareApplication

Software citation



Software Citation Principles

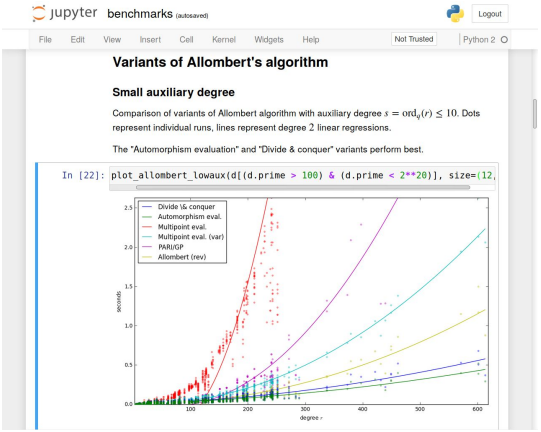
https://doi.org/10.7717/peerj-cs.86

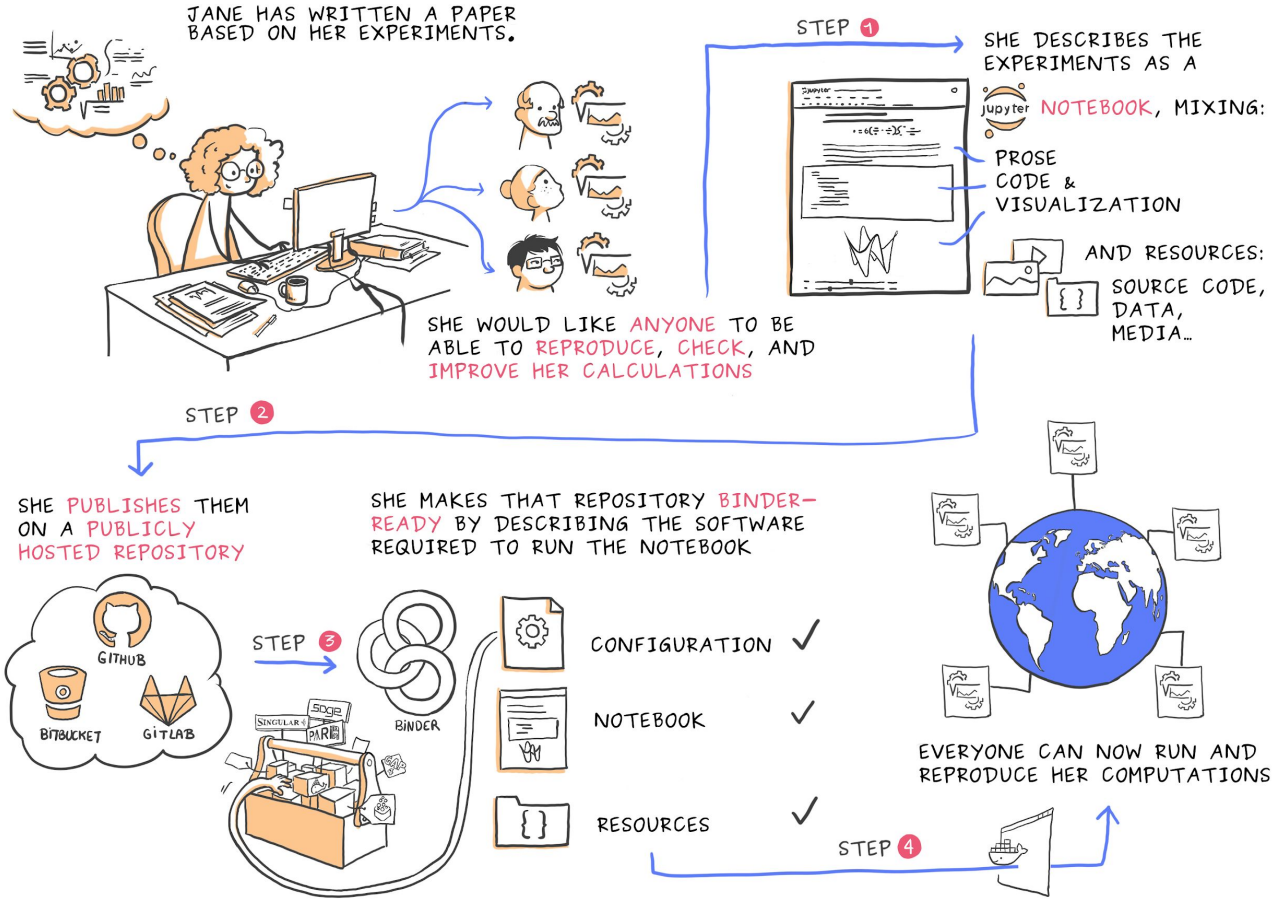
- **Importance:** Software should be considered a legitimate and citable product of research...
- **Credit and attribution:** ...normative, legal attribution to all contributors to the software...
- **Unique identification:** ...identification that is machine actionable, globally unique, interoperable, and recognized...
- **Persistence:** Unique identifiers and metadata describing the software and its disposition should persist...
- **Accessibility:** ...access to the software itself and to its associated metadata...
- **Specificity:** ...identification of, and access to, the specific version of software...

https://slides.com/dbouquin/datafest2019_citesoftware

https://www.force11.org/group/software-citation-implementation-working-group

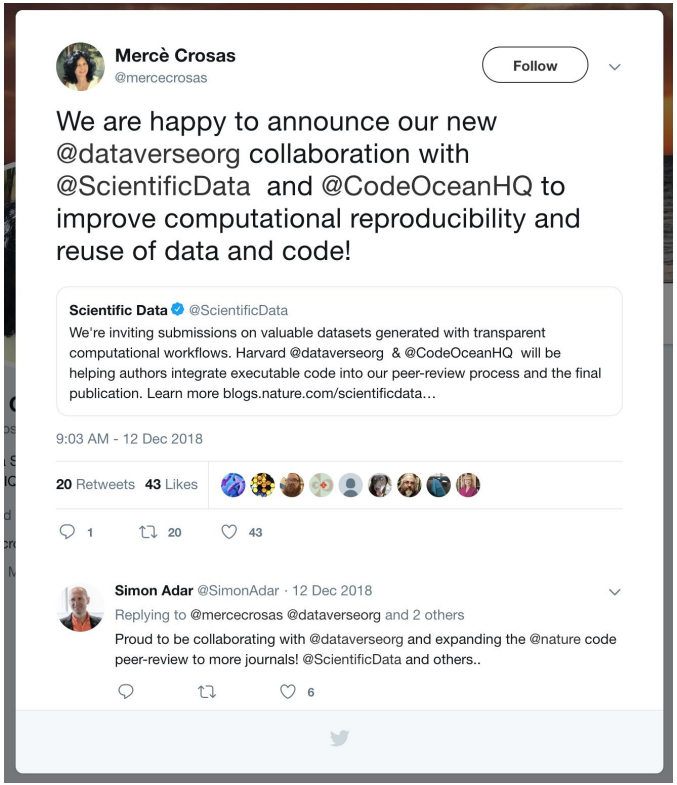
Reproducibility





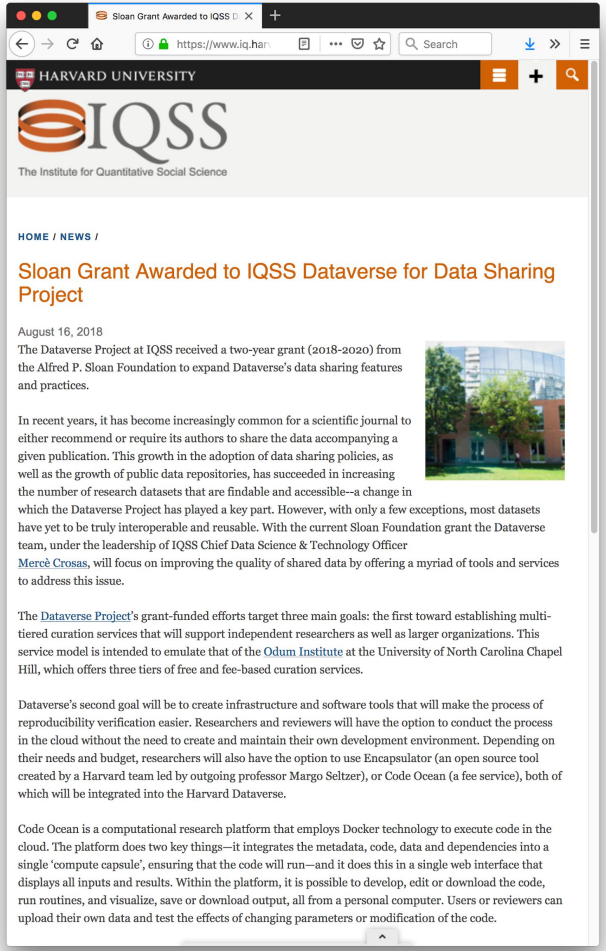
https://opendreamkit.org/2017/11/02/use-case-publishing-reproducible-notebooks/

Code Ocean (Sloan grant)



CODE OCEAN

Code Ocean is a cloud-based **computational reproducibility** platform

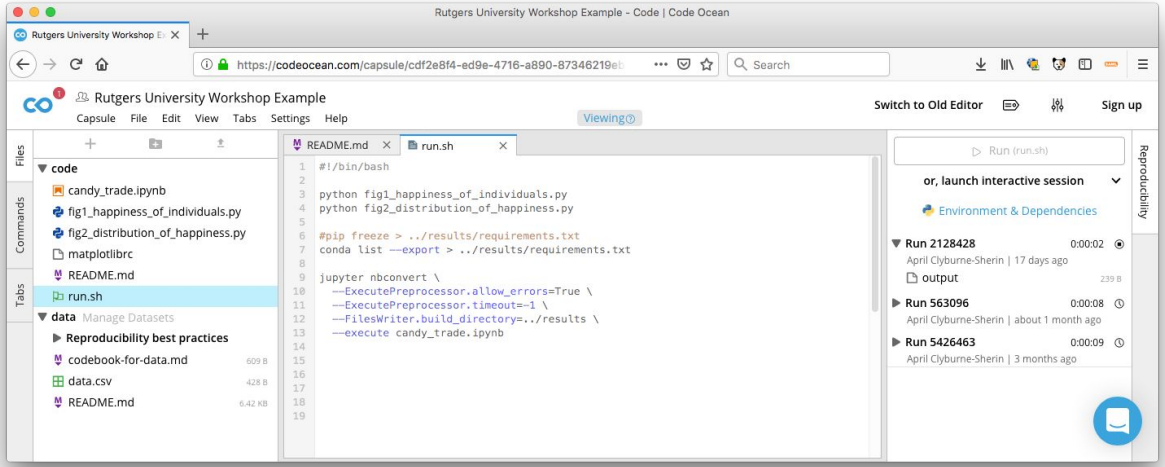


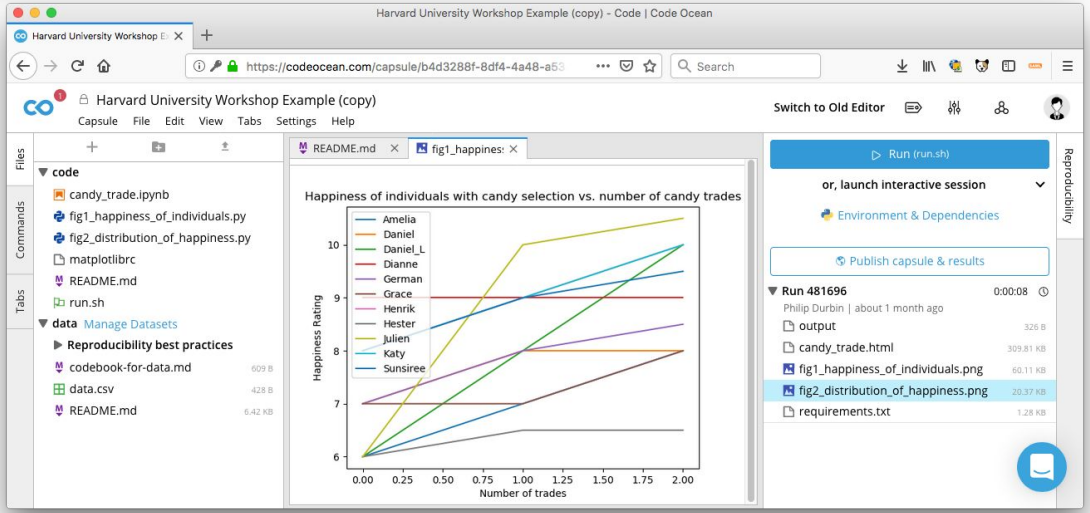
https://twitter.com/mercecrosas/status/1072899669074821122

https://www.iq.harvard.edu/news/sloan-grant-awarded-iqss-dataverse-data-sharing-project

Code Ocean

CODE OCEAN





http://bit.ly/harvard-oa-week

https://github.com/aprilcs/candy_trade

https://github.com/IQSS/dataverse/issues/5028

Replication Datasets

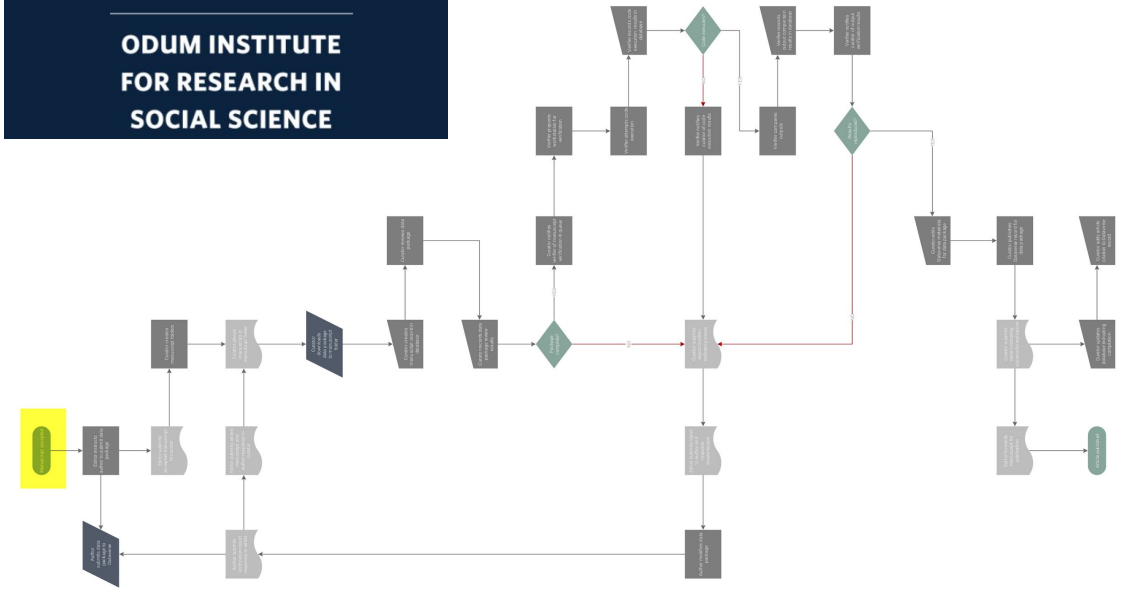
UNC
ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE

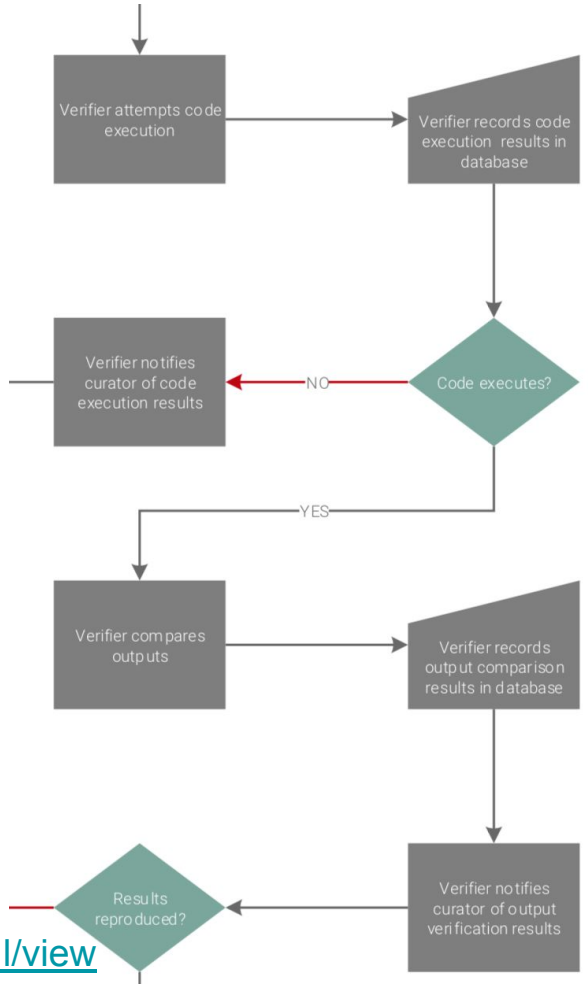
"This dataset underwent an independent verification process that **replicated the tables and figures in the primary article**. For the supplementary materials, verification was performed solely for the **successful execution of code**. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill."

6000 replication datasets in Harvard Dataverse: https://dataverse.org/best-practices/replication-dataset

Verification workflow

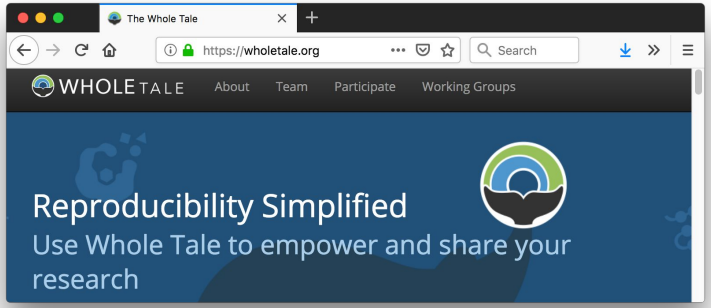
UNC
ODUM INSTITUTE
FOR RESEARCH IN
SOCIAL SCIENCE

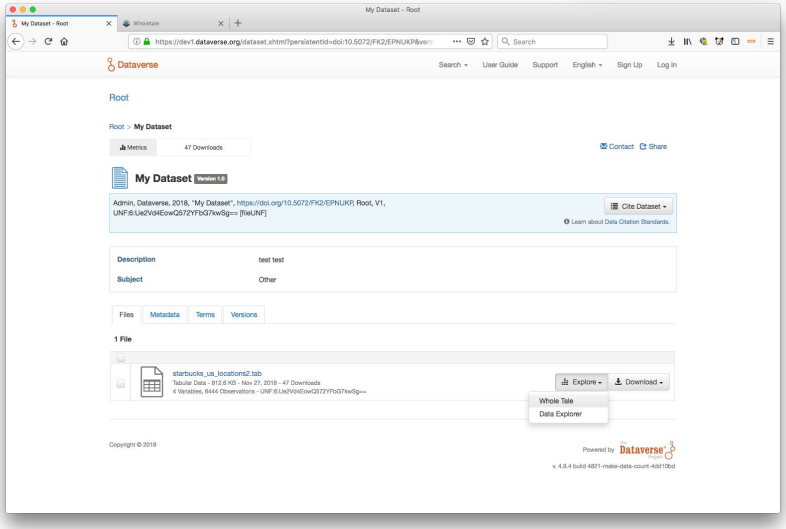


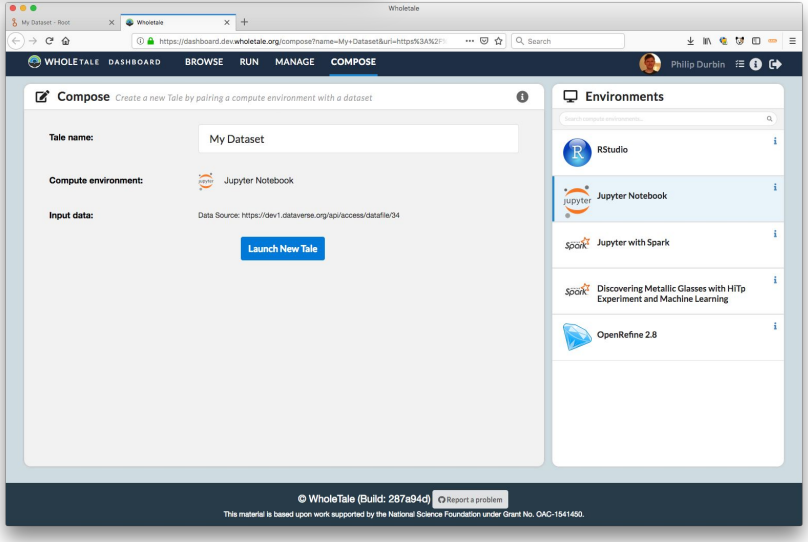


https://drive.google.com/file/d/1W4gdckYXh9fANLdTloTydJRGtb2E9PMl/view

Whole Tale







https://wholetale.org

Open Science Infrastructure Working Group

Description

The proposed working group will bring together representatives from key open science infrastructure projects related to computational reproducibility and publishing of composite research objects (e.g., code, data, and environment) to improve interoperability and to coordinate collaborative development as needed. A central goal of the project is to maximize the utility of developed tools while minimizing duplication of effort. The working group may also expand to cover technical approaches to tracking provenance.

Meeting notes

- Meeting notes
- Related Github issues:
  - https://github.com/whole-tale/whole-tale/issues/43
  - https://github.com/whole-tale/whole-tale/issues/50

Outputs

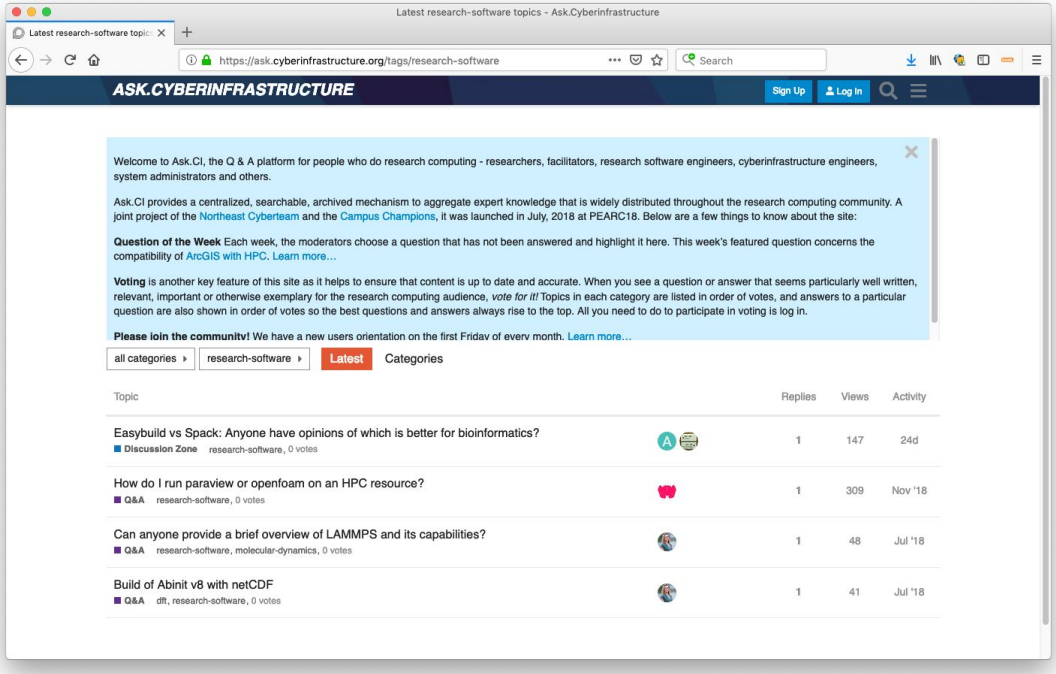
- Recommendations/best practices for publishing composite objects to research repositories for re-execution.
- Define points of interoperability across projects
- Identify areas where WT project can leverage and contribute to existing community initiatives (e.g., repo2docker)

https://github.com/whole-tale/whole-tale/tree/master/working_groups/open-infrastructure

https://dataverse.org/events

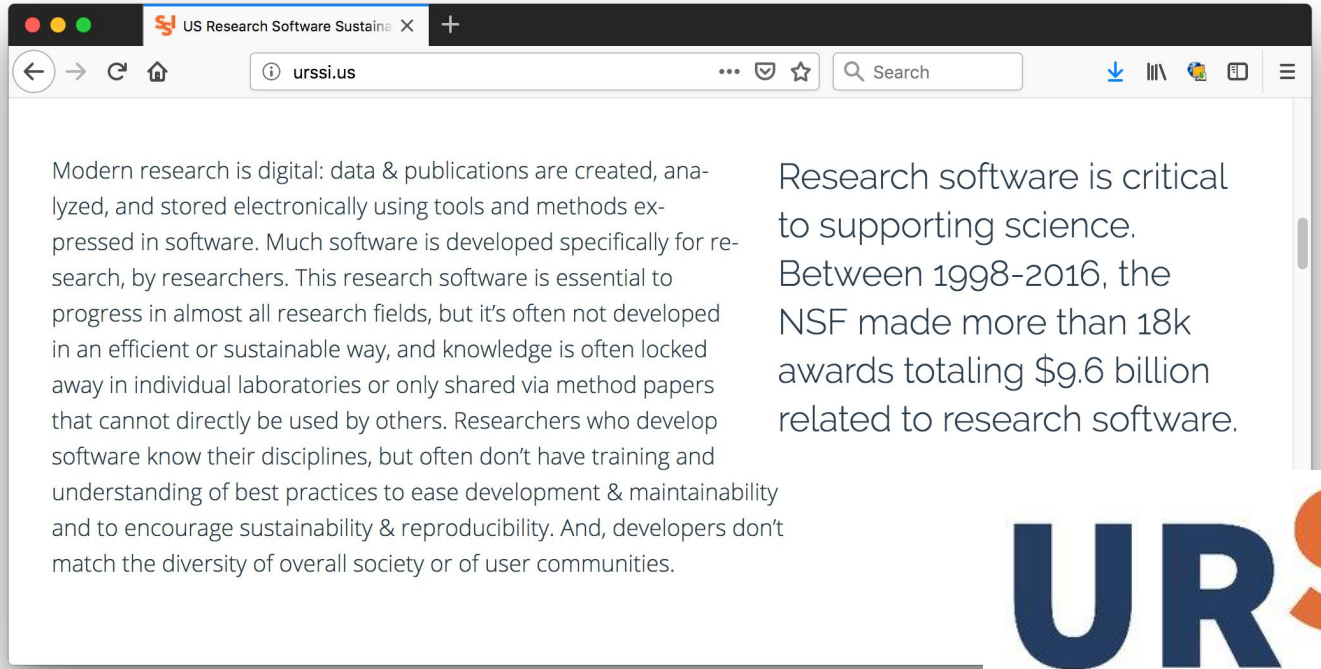
2019 Dataverse Community Meeting: June 19-21

Ask.CI research software



https://ask.cyberinfrastructure.org/tags/research-software

US Research Software Sustainability Institute



https://discuss.urssi.us/t/recent-talks-on-research-software/30

Thank you!

@philipdurbin

@dataverseorg

support@dataverse.org

The Dataverse® Project

Open source research data repository software

Researchers

Enjoy full control over your data. Receive *web visibility, academic credit,* and *increased citation counts.* A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data.* Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

Developers

Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools,* or other research and data archival systems with Dataverse. Want to contribute?

https://dataverse.org

My questions for you

- Are potential and existing code depositors to Dataverse comfortable with Dataverse's current versioning convention being applied to code? Why or why not?
- Do current or potential code depositors want to apply the version number from GitHub to the code they publish in Dataverse? Why or why not?
- Do users of code in Dataverse expect a relationship between the version in Dataverse and the releases in GitHub? Why or why not?