Research software and Dataverse

Philip Durbin @philipdurbm
What is 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

Jane has written a paper based on her experiments. She would like anyone to be able to reproduce, check, and improve her calculations.
Code already in Dataverse
Import from GitHub prototype

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

https://dataverse.harvard.edu

Dataverse Usability Research - Sign Up Form

Thank you for your interest in participating in our usability research. The aim of this research is to help the Dataverse team evaluate and improve Dataverse.

Our studies take less than an hour and typically involve speaking with a researcher one-on-one over a voice connection and sharing a view of your screen while interacting with mock-ups or demos of upcoming Dataverse features. Research is always conducted remotely, so you can participate no matter where you are located.

When we run a usability study that is relevant to you, we will email you with further information about the study, and if you are interested you will be able to schedule a specific time to participate.

Any personal information collected below will be kept separate from our research data, and will not be linked to the study's results or findings in any way. At any time you can respond to one of our emails to be removed from our mailing list.

https://dataverse.harvard.edu
Software metadata

http://datacurationnetwork.org
CodeMeta

Supported Content Types

DataCite supports a number of metadata content types:

<table>
<thead>
<tr>
<th>Format</th>
<th>Content Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDF XML</td>
<td>application/rdf+xml</td>
</tr>
<tr>
<td>RDF Turtle</td>
<td>text/turtle</td>
</tr>
<tr>
<td>Citeproc JSON</td>
<td>application/vnd.citationstyles.csl+json</td>
</tr>
<tr>
<td>Schema.org in JSON-LD</td>
<td>application/vnd.schemacorg.id+json</td>
</tr>
<tr>
<td>Codemeta</td>
<td>application/vnd.codemeta.id+json</td>
</tr>
<tr>
<td>Formatted text citation</td>
<td>text/x-bibliography</td>
</tr>
<tr>
<td>RIS</td>
<td>application/x-research-info-systems</td>
</tr>
<tr>
<td>BibTeX</td>
<td>application/x-bibtex</td>
</tr>
<tr>
<td>DataCite XML</td>
<td>application/vnd.datacite.datacite+xml</td>
</tr>
<tr>
<td>DataCite JSON</td>
<td>application/vnd.datacite.datacite+json</td>
</tr>
</tbody>
</table>

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

Caltech Library

Enhanced software preservation now available in CaltechDATA!

Friday, March 09, 2018

CaltechDATA has supported automatic preservation of GitHub software repositories since launch, so anyone at Caltech can get a DOI (permanent identifier) for their software project and have Caltech Library handle long term preservation. However, most GitHub repositories do not include clear metadata such as authors, affiliations, or ORCID identifiers. CaltechDATA now supports CodeMeta, a new standard format for software metadata. By including a codemeta.json file in your GitHub repo, your full author list, keywords, and license will be listed in CaltechDATA and registered with your DOI.

This improvement is powered by ames, a Python package for automating metadata changes developed at Caltech Library. Every 5 minutes, ames harvests all the GitHub-created records in CaltechDATA and stores them using dataset (our lightweight data storage package). These records are then analyzed for codemeta.json files. If a CodeMeta file is found, the relevant metadata is extracted and added to the CaltechDATA record and DOI. We currently support authors, keywords, and license fields - but more will be added as a community of practice develops. We’re also exploring better ways to generate CodeMeta files as part of the software release process.

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

https://schema.org/Dataset
Software citation

Software Citation Principles

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

https://slides.com/dbouquinn/datafest2019_citesoftware

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

https://opendreamkit.org/2017/11/02/use-case-publishing-reproducible-notebooks/
Code Ocean (Sloan grant)

Mercè Crosas @mercecrosas

We are happy to announce our new @dataverseorg collaboration with @ScientificData and @CodeOceanHQ to improve computational reproducibility and reuse of data and code!

Scientific Data @ScientificData

We’re inviting submissions on valuable datasets generated with transparent computational workflows. Harvard @dataverseorg & @CodeOceanHQ will be helping authors integrate executable code into our peer-review process and the final publication. Learn more blogs.nature.com/scientificdata...

9:03 AM - 12 Dec 2018

20 Retweets 43 Likes

Simon Adar @SimonAdar - 12 Dec 2018

Replying to @mercecrosas @dataverseorg and 2 others

Proud to be collaborating with @dataverseorg and expanding the @nature code peer-review to more journals @ScientificData and others...

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

Code Ocean


https://github.com/aprilcs/candy_trade

https://github.com/IQSS/dataverse/issues/5028
6000 replication datasets in Harvard Dataverse: https://dataverse.org/best-practices/replication-dataset

"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."
Verification workflow

https://drive.google.com/file/d/1W4gdckYXh9fANLdTloTydJRGtb2E9PMI/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
2019 Dataverse Community Meeting: June 19-21

https://dataverse.org/events
Ask.CI research software

https://ask.cyberinfrastructure.org/tags/research-software
Modern research is digital: data & publications are created, analyzed, and stored electronically using tools and methods expressed in software. Much software is developed specifically for research, by researchers. This research software is essential to progress in almost all research fields, but it's often not developed in an efficient or sustainable way, and knowledge is often locked away in individual laboratories or only shared via method papers that cannot directly be used by others. Researchers who develop software know their disciplines, but often don't have training and understanding of best practices to ease development & maintainability and to encourage sustainability & reproducibility. And, developers don't match the diversity of overall society or of user communities.

Research software is critical to supporting science. Between 1998-2016, the NSF made more than 18k awards totaling $9.6 billion related to research software.

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

@philipdurbin
@dataverseorg

support@dataverse.org

---

**Open source research data repository software**

<table>
<thead>
<tr>
<th>Researchers</th>
<th>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?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals</td>
<td>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?</td>
</tr>
<tr>
<td>Institutions</td>
<td>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?</td>
</tr>
<tr>
<td>Developers</td>
<td>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?</td>
</tr>
</tbody>
</table>

[https://dataverse.org](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?