Advancing science with Dataverse: Publication, discovery, citation, and exploration of research data

The Dataverse Project

HARVARD Dataverse

Philip Durbin @philipdurbin

IQSS The Institute for Quantitative Social Science
What is Dataverse?
research data sharing enthusiasts: scientists, researchers, curators, librarians, etc.
Available languages:

- English (US), latest develop branch maintained by IQSS Harvard
- French (Canada), latest available 4.17 maintained by Bibliothèques Université de Montréal
- French (France), 4.9.4 maintained by Sciences Po
- German (Austria), 4.9.4 maintained by AUSSDA
- Slovenian, 4.9.4 maintained by ADP, Social Science Data Archive
- Swedish, 4.9.4 maintained by SND, Swedish National Data Service
- Ukrainian, 4.9.4 maintained by The Center for Content Analysis
- Spanish, 4.11 maintained by El Consorcio Madroño
- Italian 4.9.4 maintained by Centro Interdipartimentale UniData
- Hungarian, 4.9.4 maintained by TARKI

Open source research data repository software  http://dataverse.org
What is Dataverse for?
"All my work is built on the premise that climate change is the single biggest existential threat facing humanity."

arvindravikumar.com
Question for #AcademicTwitter.

If you collect a lot of primary data that you want to make publicly available along with the paper, what would you do?

I've always taken the SI route but #Reviewer2 is insisting on a separate DOI.

(SI stands for "supplementary information")

https://twitter.com/arvindpawan1/status/1215621920080699392

https://doi.org/10.1088/1748-9326/ab6ae1
Ravikumar, Arvind, 2020, "Replication Data for: "Repeated Leak Detection and Repair Surveys Reduce Methane Emissions Over Scale of Years"", https://doi.org/10.7910/DVN/T2ZFQN, Harvard Dataverse, V1
Cultural change
How to achieve a cultural change towards open science

Based on a tweet storm by @BrianNosek

Data take centre stage

Nature Materials 19, 1(2020) | Cite this article

864 Accesses | 41 Altmetric | Metrics

We are updating our editorial policies to further encourage authors to make their data publicly accessible. Publishing Extended Data figures and source data online will also ensure that data are given a more prominent role.

https://www.nature.com/articles/s41563-019-0574-2

https://twitter.com/BrianNosek/status/973506782063677440
Findable
Accessible
The tabular file contains information on known Harvard repositories on GitHub, such as the number of stars, programming language, day last updated, number of open issues, size, number of forks, repository URL, creation date, and description. Each repository has a corresponding JSON file that was retrieved using the GitHub API with code and a list of repositories available from https://github.com/oss/known-source-at-harvard. Each repository has a corresponding JSON file that was retrieved using the GitHub API with code and a list of repositories available from https://github.com/oss/known-source-at-harvard.

This XML file does not appear to have any style information associated with it. The document tree is shown below:

```xml
<metadata>
  <title>Open Source at Harvard</title>
  <identifier>https://doi.org/10.7910/DVN/TJCLKP</identifier>
  <creator>Durbin, Philip</creator>
  <publisher>HARVARD DATASET</publisher>
  <issued>2017-07-06</issued>
  <modified>2019-02-28T03:33:32Z</modified>
  <description>The tabular file contains information on known Harvard repositories on GitHub, such as the number of stars, programming language, day last updated, number of open issues, size, number of forks, repository URL, creation date, and description. Each repository has a corresponding JSON file that was retrieved using the GitHub API with code and a list of repositories available from https://github.com/oss/known-source-at-harvard.</description>
  <subject>Social Sciences</subject>
  <dateSubmitted>2017-07-06</dateSubmitted>
  <license>CC0</license>
</metadata>
```

https://doi.org/10.7910/DVN/TJCLKP
Interoperable
- Getting Data In
  - Dropbox
  - Open Science Framework (OSF)
  - RSpace
  - Open Journal Systems (OJS)
- Embedding Data on Websites
  - OpenScholar
- Analysis and Computation
  - Data Explorer
  - TwoRavens/Zelig
  - WorldMap
  - Compute Button
  - Whole Tale
  - Binder
- Discoverability
  - OAI-PMH (Harvesting)
  - SHARE
- Research Data Preservation
  - Archivematica
  - DuraCloud/Chronopolis

Reusable
This is your machine learning system?

Yup! You pour the data into this big pile of linear algebra, then collect the answers on the other side. What if the answers are wrong?

Just stir the pile until they start looking right.

https://xkcd.com/1838/

https://ajps.org/ajps-verification-policy/
Trisovic, Crosas, et al, 2020, working paper
FAIR Data Principles

- Findable
- Accessible
- Interoperable
- Reusable

Mercè Crosas @mercecrosas · Jan 23
The slides from my talk on the implementation of FAIR data principles in Dataverse and going beyond FAIR, at the European Dataverse Workshop @UITromso @dataverseorg #FAIRdata #dataverse2020

FAIR principles and beyond: Implementation in Dataverse
Keynote for the European Dataverse Workshop 2020 at ...

scholar.harvard.edu

https://twitter.com/mercecrosas/status/1220344995628175360

https://dx.doi.org/10.1038/sdata.2016.18
Bonus content
Searchable Linkable Open Public Indexed (SLOPI) Communication

or

Why open source projects should avoid Slack

http://blog.greptilian.com/2020/01/25/slopi-communication/
The Open Source Software Health Index Project

Fourth Quarter, 2019 update

October 11, 2019

It's been a year since we began our project to develop a framework for evaluating the health of open source software used in academic research settings by measuring different aspects or factors of OSS projects, which will help answer questions such as how easy it is for people to contribute to OSS projects and how easy it is to use and deploy the software. After initial research into software evaluation frameworks and a number of meetings and workshops with experts, we have chosen the 20 projects that we will use to evaluate our framework.

All of the projects listed below are used in academic libraries and research labs. The OSS experts we have been collaborating with this past year also contribute to many of these projects, which will make it easier to get feedback about the quality and feasibility of the factors and continue improving the framework.

- Archi
- Archivaromatica
- Bloco
- Blacklight
- CORAL
- Dataverse
- Districtbuilder
- DSpace
- Fedora Commons
- Jupyter
- Jupyter notebook
- LOCKSS
- Lots of Copies
- Keep Stuff Safe
- Mirador
- Omeka
- Open Journal Systems
- Parsl
- R Markdown
- Scikit-learn
- Stencila
- Zotero

Next steps

Our next steps include finalizing the factors in the framework and identifying potential methods for gathering data from these projects for each factor. In some cases, information about the projects can be mined from their GitHub repositories, and we've been working closely with the CHA OSS Project team, whose Augur software suite can collect and visualize GitHub data.
Thank you!

My website: http://greptilian.com

My blog: http://blog.greptilian.com

@pdurbin on GitHub.

@philipdurbin on Twitter

philip_durbin@harvard.edu

My institute at Harvard: https://www.iq.harvard.edu

For a quick conversation: https://chat.dataverse.org