Dataverse
A repository platform for sharing data and code

Workshop of the OECD’s Committee for Scientific and Technological Policy
Revision of Recommendation concerning access to research data from public funding
OECD Conference Center, Paris, October 15 2019

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer, IQSS
Harvard University’s Research Data Officer, HUIT
@mercecrosas
Dataverse – Achievements

Facilitating data sharing with technology, standards, and incentives
15 years of Dataverse

• Launched in 2006 by Harvard’s Institute for Quantitative Social Science

• Now 48 Dataverse installations sites throughout 6 continents

• Each site can support multiple Universities or research organizations

• Harvard Dataverse is the largest Dataverse repository site with:
  • 3,400 dataverses (collections of datasets)
  • 91,000 datasets
  • 560,000 files
  • 8.9 million downloads

• An active, growing open-source community with 100 contributors, 41 releases
Sites harvest metadata from each other

- 48 Dataverse sites can be federated via the Open Archives Initiative Protocol for Metadata Harvesting.
- Data **searchable** in one Dataverse site while hosted in another.
Two Examples: Consortium Dataverse sites

- **8 universities in Norway** as members; the other 3 to join soon
- Policies and guidelines common to all DataverseNO members
- Global and local support
- Applied for Core Trust Seal certificate
  - [https://site.uit.no/dataverseno/](https://site.uit.no/dataverseno/)

- **11 universities in Texas** as members
- Led by Texas Digital Library Consortium
- Texas Data Repository steering committee focuses on outreach
- Working on Core Trust Seal
- 800 datasets created in last 2 years
  - [https://www.tdl.org/texas-data-repository/](https://www.tdl.org/texas-data-repository/)
Combines technology, standards, incentives

- A **data citation** with a globally unique persistent identifier and credit for data author
- Standard **metadata**, with variable-level metadata, plus rich custom metadata
  - Schema.org JSON-LD, DataCite, Dublin Core, DDI, OAI-ORE, OpenAire, PROV
- **Tiered data access**, depending on data restrictions:
  - Fully Open, CC0; Guestbook; Restricted w/ Data Use Agreement
- **Data publishing workflows**, with anonymous review, and reviewer/curator roles
- Multiple **versions** of a dataset
- **Branding and customization** for each dataverse (collection of datasets)
- Overall, follows **FAIR** guiding principles
Data Citation Implementation

- Global unique persistent identifiers (GUID)
- Allow to cite specific version or subset
- GUID resolves to dataset landing page
- Landing page w/ instructions to access data
- Machine-readable citation metadata
- Schema.org JSON-LD in landing page
- Standard Bibliographic formats
Dataset Landing Page

- Dataset and file citation
- Data files review and downloads
  - Automatic transforms tabular files to multiple formats
  - Extracts variable metadata
- Dataset Metadata
- Metrics: Make Data Count
- Terms and data use agreements
- Versions and Provenance
- APIs to access metadata & data
- Machine-actionable
Dataverse — Next

Improving research reproducibility and data reuse
8,000 of the 90,000 datasets in Harvard Dataverse contain the files to reproduce the publish results.

documentation

data

code
Re-execution of R Code in published datasets

84.4% R code files fail to execute

77.5% of datasets with R files contain non-executable code

Source: Current study on code execution for datasets published in Dataverse by Ana Trisovic (Harvard’s IQSS) funded by Sloan Foundation
New Features: computational reproducibility

- Include **reproducibility as part of review** workflow
- Integrate Dataverse with computational tools to **facilitate code execution**
- Deposit a **capsule** (container with data and code) once verified for reproducibility
- When possible, **automate code execution** upon publishing the data and code

Also, in the process of evaluating integration with:

- Data Curation Tools
- Research Objects
- Citation and metadata for software
Integration with computational tools

Computational & Reproducibility Environments

Pull data from Dataverse

Archive input, output, and workflow to Dataverse

Data + Code

Execute and curate code

The Dataverse Project

Reproducible container; research object

Reproducibility verification

@mercecrosas
Thanks

dataverse.org | dataversecommunity.global/ | scholar.harvard.edu/mercecrosas