Role of Generalist Repositories to Enhance Data Discoverability and Reuse
NIH, Feb. 11-12
DATAVERSE SOFTWARE PLATFORM USED WORLDWIDE

- 54 installations
- In 6 continents
- 6000 dataverses
- 130K datasets
- Metadata is shared across the Network of Dataverses
A GROWING AND ENGAGED COMMUNITY

“The community is the one thing that makes this work!”

Sherry Lake, Scholarly Repository Librarian, University of Virginia

- Core development at Harvard
- > 100 code contributors
- Regular, popular Community meetings, calls, discussions
- Fast adoption (e.g., in Europe) w/ multiple language support
- Governance and assistance from the Global Dataverse Community Consortium
Dataverse repositories come in various types and sizes.

The **Harvard Dataverse** is a generalist repository open to all, where NIH funded data may be published.

NIH funded data may also be published in other US **Dataverse repositories**, including domain-specific and institutional repositories.
EXAMPLES
HARVARD DATaverse: Free and Open to all

- 90K discoverable datasets
- > 3000 in biomedicine
- > 80 journal dataverses
- Datasets up to TB
- Files up to 10 GB
- Offers new fee-based data curation services
SBGRID DATA:
A Repository for Structural Biology primary data

- Helmsley grant to PIs Sliz and Crosas to expand Dataverse for SBGrid Data and other biomedical repositories
- Added support for large datasets, in-place computation, ORCID auth
- Leverages Dataverse community for longer term repository sustainability

Beta: not in production yet

This book seeks to make the sources of its inferences as transparent as possible. In doing so it draws from best practices outlined by the American Political Science Association in its 2012 Council Statement on Openness in Political Science and in a 2013 Guideline for Data Access and Research Transparency for Qualitative Research in Political Science. This data project and its associated Transparency Appendix (TRAX) are key components of this effort. Where permitted by copyright, the data project provides digital copies of grey literatures, policy documents, and media reports referenced in the book or in the book's
UNC DATAVERSE: Supports repositories for others through partnerships
Share, publish, and manage your data. Find and cite data across all research fields.

Welcome to the Texas Data Repository Dataverse, a research data management system for Texas Digital Library (TDL) member institutions. To add, share, and publish your data or work on a project, select your local institutional repository from the institutions below. To find datasets from across Texas institutional dataverses, start here.

LEARN MORE

- Go to the user guide.
- Contact a local university librarian for help.
UVA DATAVERSE and JOHNS HOPKINS DATAVERSE: Institutional repositories serving their community
FAIR AND CORE TRUST SEAL
Dataverse substantially facilitates supporting FAIR data principles and Core Trust Seal certification.

Best archival practices, dedicated support, and data reviews and curation contribute to improve data quality and trustworthiness.
• Strong support for Findable, Accessible, and Reusable principles

• (F3 was fixed after test)

• Interoperable principles more difficult to follow; often requires input from data authors or curators

• Emphasis in machine-readability

• FAIR is not a “standard”, it’s a path
FINDABILITY: Full, Standard Data Citation Automatically Generated

Data Citation, with DataCite DOI, fully compliant with Joint Declaration of Data Citation Principles
FINDABILITY AND REUSABILITY: Support Many Metadata Standards

ACCESSIBILITY:
Metadata Always Available

Deaccession reason in dataset landing page when data not longer available
INTEROPERABLE:
Rich Metadata Derived from each Variable in data file

Extensive variable metadata (summary statistics) automatically derived from tabular data file
INTEROPERABLE: Facilitating Use of Standard Vocabularies

FAIR Controlled Vocabularies can be added in metadata template
REUSABLE: Licenses, Terms, and Tier-access to Data

- CC0 as default waiver for open data;
- Other Licenses and custom Terms;
- Tier-access for restricted data
BEYOND FAIR, CURRENT DATAVERSE EFFORTS

- Responsible FAIR for **sensitive data:**
  DataTags, Differential Privacy tools,
  Remote Trusted Storages (**larger data**)

- **Data curation for data quality:**
  New data curation tools and services

- **Capsules for reproducibility:**
  Integration with computational platforms, capsules and research objects

Thanks!