DATAVERSE IS A REPOSITORY
for finding, citing, and publishing data

DATAVERSE IS A PLATFORM
for building your own data repository

DATAVERSE IS A COMMUNITY
which facilitates data access and data sharing around the world
COMMUNITY FEATURES DATA PROJECTS
A GROWING, ENGAGED COMMUNITY
DATAVERSE.ORG

33 Dataverse Repositories sites around the world
# DataVere Community Growth

<table>
<thead>
<tr>
<th>Year</th>
<th>Dataverse Sites</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>2</td>
<td>DataVere Development starts at Harvard's Institute for Quantitative Social Science (IQSS)</td>
</tr>
<tr>
<td>2015</td>
<td>14</td>
<td>First Annual DataVere Community Meeting</td>
</tr>
<tr>
<td>2016</td>
<td>18</td>
<td>4 developers</td>
</tr>
<tr>
<td>2017</td>
<td>23</td>
<td>First release</td>
</tr>
<tr>
<td>2018</td>
<td>33</td>
<td>74 contributors, 30 releases, 12,807 commits</td>
</tr>
</tbody>
</table>
GLOBAL DATAVEVERSE COMMUNITY CONSORTIUM

In 2018, a **new** international consortium is formed to support and coordinate efforts across Dataverse Repositories.

http://dataversecommunity.global (coming soon!)
BUILDING AN ACTIVE, ENGAGED COMMUNITY WITH:

• Transparency and Common Knowledge
• Process and Tools
• Human Touch
TRANSPARENCY AND COMMON KNOWLEDGE

- High-level goals and roadmap in dataverse.org site
- Development status in Waffle
- Issues discussions in GitHub
- General discussions in Google Groups (mailing list)
PROCESS TO SUPPORT AN AGILE DEVELOPMENT

Engage early with contributors on technical design and user testing:

1. Pull Request
2. Code Review
3. QA
4. Release

Tools:
- Waffle
- GitHub
- Google groups
- irc
- Slack
THE HUMAN TOUCH

- Annual Community Meeting:
  - ~150 people
  - Organizations from ~15 countries
- Quick reply to mailing list (Google groups) and IRC
- Biweekly Call (last year):
  - 23 Calls
  - 228 Participants
  - 18 Organizations

Dataverse World Cup!
A RICH SET OF USER-FRIENDLY FEATURES
DATA CITATION:
CREDIT AS AN INCENTIVE TO SHARE DATA

- A formal data citation automatically generated
- Attribution to data creators and data providers
- Persistent identifier (e.g., DOI) resolves to dataset landing page
- Version in citation
- Universal Numerical Fingerprint (UNF): a checksum independent of file format, for tabular data files
- Compliant with the *Joint Declaration of Data Citation Principles*

Download data citation ready to be used in reference manager
METADATA TO FIND AND REUSE DATA

At multiple Levels:

- Citation metadata
- Custom metadata
- File metadata
- Variable-level metadata

With multiple Standards:

- Data Documentation Initiative (DDI)
- Dublin Core
- Schema.org
SCHEMA.ORG USED BY GOOGLE DATASET SEARCH

- Schema.org JSON-LD embedded in HTML of dataset landing page
- Datasets become discoverable through Google Dataset Search

Metadata from schema.org in Dataverse dataset landing page
VERSIONING OF DATASETS AND FILES

- Major and minor versions
- Major versions show in the data citation
- Track both metadata changes and files changes
TIERED ACCESS TO DATA

- Default access is public, with CC0 waiver
- Allow public and restricted files
- Descriptive metadata always public for discoverability
- Custom Terms of Use, when needed
- Optional Guestbook to collect information from users
**TABULAR DATA EXPLORATION**

- Variable metadata automatically extracted
- Descriptive statistics automatically computed
METADATA EXTRACTION FROM ASTRONOMY FILES

Metadata (instrument information) is extracted automatically from FITS files header upon data upload.

Metadata from FITS Header
MANAGE AND CUSTOMIZE YOUR OWN DATaverse

- Create a dataverse to manage your own collection of datasets
- Brand your dataverse or embed in your website
EXTENSIVE API TO ENABLE TOOL INTEGRATION

Dataverse Project

API Guide
- Introduction
- SWORD API
- Search API
- Data Access API
- Native API
- Metrics API
- Client Libraries
- Apps

http://guides.dataverse.org
A WIDE VARIETY OF DATA AND DATAVERSE

- Dataverse for Journals
- Dataverse for Researchers
- Dataverse for Research Communities
- Dataverse for one or multiple Institutions
DATA POLICIES IN SOCIAL SCIENCE JOURNALS

More than 50% of the top 50 journals in anthropology, economics, psychology, and political sciences have data policies that either encourage or require to share the data associated with the article.

Crosas, Gautier, Karcher, Kirilova, Otalora, Schwartz, 2018.

Data Policies of highly-ranked social science journals
DATAVERSE FOR A JOURNAL

Hosted at Harvard Dataverse repository (80 journal dataverses)
DATAVERSE FOR A RESEARCHER

Hosted at Harvard Dataverse repository
DATAVERSE FOR A RESEARCH COMMUNITY: STRUCTURAL BIOLOGY

Hosted SBGrid Consortium, Harvard Medical School
DATAVERSE FOR MULTIPLE UNIVERSITIES

Texas Data Repository Dataverse
A statewide collaboration of Texas higher education institutions

Share, publish, and archive your data. Find and cite data across all research fields.

Welcome to the Texas Data Repository, a statewide archive of research data from 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.

Hosted by Texas Digital Libraries, a consortium of Texas Higher-Education Institutions
HARVARD DATaverse:
OPEN TO ALL THE RESEARCH COMMUNITY

Hosted Harvard University, in collaboration with Harvard Library, HUIT, and IQSS
http://dataverse.harvard.edu
# DATASETS DEPOSITED AT HARVARD DATaverse:

29,256

AVERAGE RELEASED DATASETS PER MONTH (IN 2018):

247

# OF TOTAL DOWNLOADS:

4M

AVERAGE DOWNLOADS PER MONTH (IN 2018):

150,000
ON-GOING PROJECTS

• Large data
• Sensitive data
• Data Quality, Reproducibility, Reusability
• Open Source 'Health' Index
LARGE DATA

More ways to upload data
- rsync

More ways to access data:
- Local access
- Compute in the cloud
- Compute in institutional research computing portals
- Integration w/ Globus?

More storages:
- Remote secure storage; data enclaves

Funding by Helmsley Charitable Trust, with focus on biomedical data, in collaboration with Piotrek Sliz
**SENSITIVE DATA: DATATAGS**

Standardize data security and access levels

<table>
<thead>
<tr>
<th>Tag Type</th>
<th>Description</th>
<th>Security Features</th>
<th>Access Credentials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>Public</td>
<td>Clear storage, Clear transmit</td>
<td>Open</td>
</tr>
<tr>
<td>Green</td>
<td>Controlled public</td>
<td>Clear storage, Clear transmit</td>
<td>Email- or OAuth Verified Registration</td>
</tr>
<tr>
<td>Yellow</td>
<td>Accountable</td>
<td>Clear storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Click-through DUA</td>
</tr>
<tr>
<td>Orange</td>
<td>More accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Password, Registered, Approval, Signed DUA</td>
</tr>
<tr>
<td>Red</td>
<td>Fully accountable</td>
<td>Encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
<tr>
<td>Crimson</td>
<td>Maximally restricted</td>
<td>Multi-encrypted storage, Encrypted transmit</td>
<td>Two-factor authentication, Approval, Signed DUA</td>
</tr>
</tbody>
</table>
SENSITIVE DATA: PRIVACY PRESERVING TOOLS

Funded by National Science Foundation, in collaboration with Harvard Privacy Tools Project
INTEGRATION WITH REPRODUCIBILITY TOOLS: CODE OCEAN

Funded by Sloan Foundation, in collaboration with CodeOcean
INTEGRATION WITH REPRODUCIBILITY TOOLS: ENCAPSULATOR

Funded by Sloan Foundation, in collaboration with Margo Seltzer
INTEGRATION WITH REPRODUCIBILITY TOOLS: CORE2

Funded by Sloan Foundation, in collaboration with the ODUM institute at UNC Chapel Hill
OPEN SOURCE 'HEALTH' INDEX

- A quantitative study to determine a health index for open source projects
- Leverage previous work (e.g., LYRASIS project and Qualification and Selection of Open Source Software (QSOS))

Funded by IMLS
THANK YOU!

dataverse.org
dataverse.harvard.edu
The Dataverse Team @IQSS
https://groups.google.com/forum/#!forum/dataverse-community

scholar.harvard.edu/mercecrosas
@mercecrosas