DATAVERSE

Mercè Crosas, IQSS, Harvard University
@mercecrosas

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

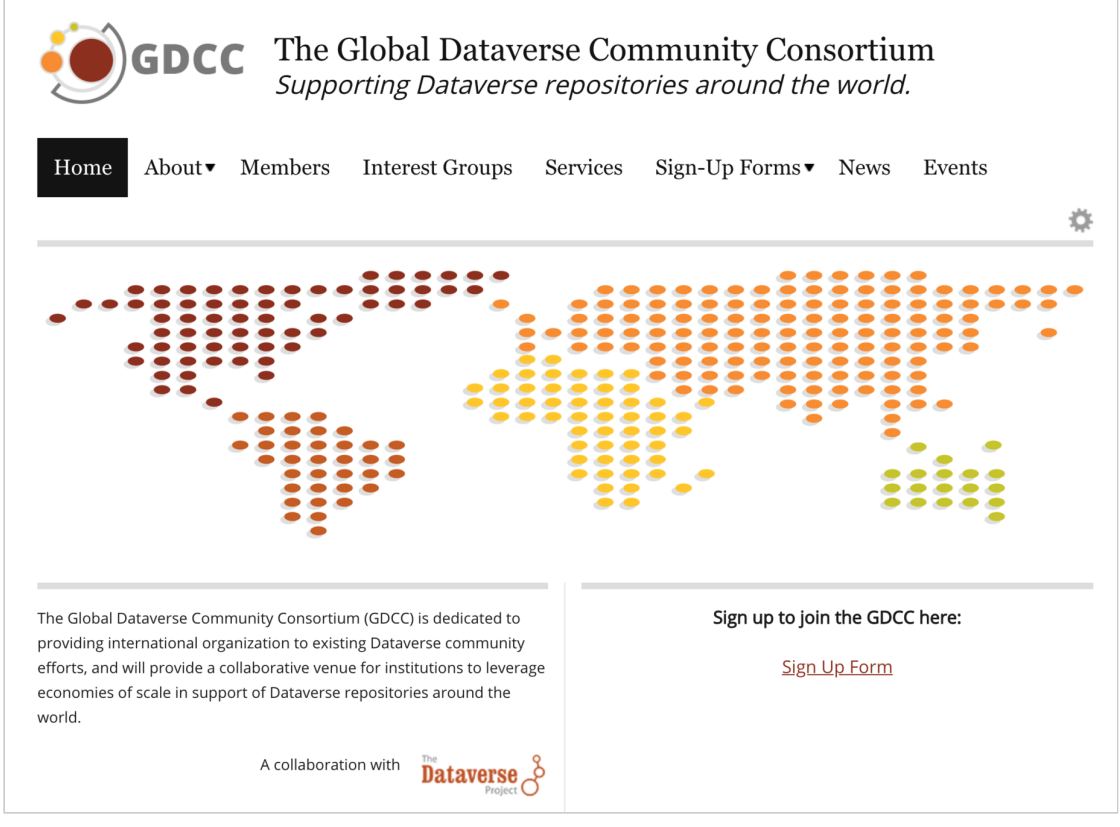
33 Dataverse Repositories sites around the world

DATAVERSE COMMUNITY GROWTH

2006	2015	2016	2017	2018
2 Dataverse sites	**14** Dataverse sites	**18** Dataverse sites	**23** Dataverse sites	**33** Dataverse sites
Dataverse Development starts at Harvard's Institute for Quantitative Social Science (IQSS) 4 developers First release	First Annual Dataverse Community Meeting			74 contributors 30 releases 12,807 commits

GLOBAL DATAVERSE COMMUNITY CONSORTIUM

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



http://dataversecommunity.global (coming soon!)

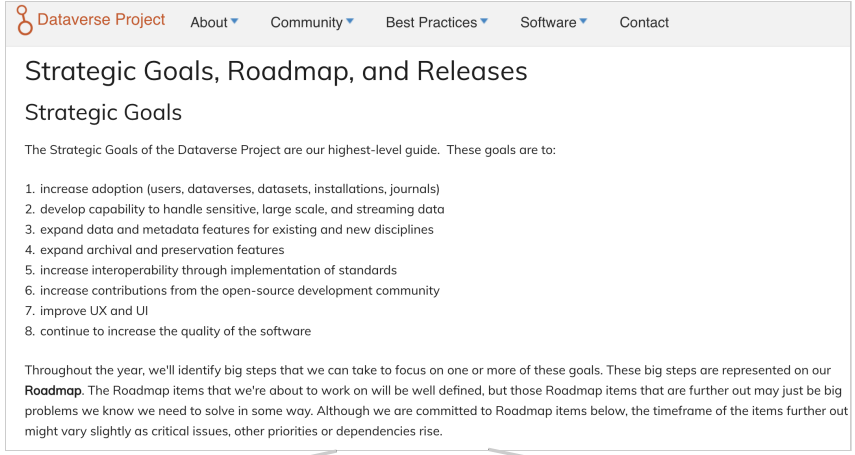
BUILDING AN ACTIVE, ENGAGED COMMUNITY WITH:

- 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)

Dataverse.org



Dataverse Project About Community Best Practices Software Contact

Strategic Goals, Roadmap, and Releases

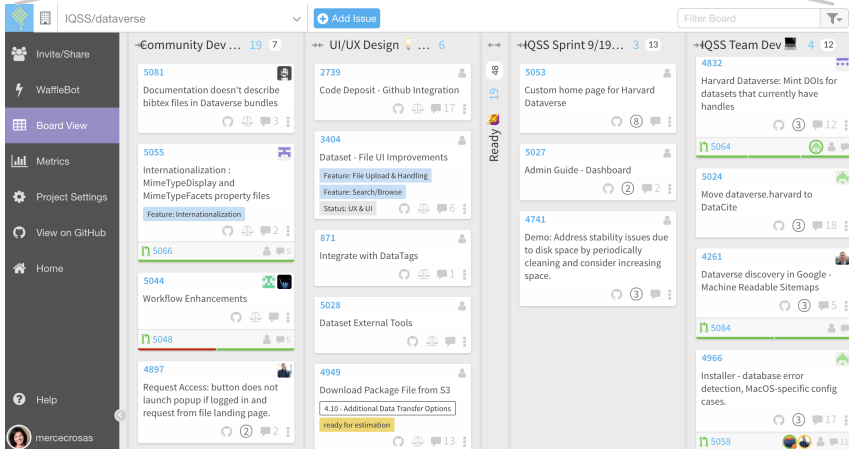
Strategic Goals

The Strategic Goals of the Dataverse Project are our highest-level guide. These goals are to:

1. increase adoption (users, dataverses, datasets, installations, journals)
2. develop capability to handle sensitive, large scale, and streaming data
3. expand data and metadata features for existing and new disciplines
4. expand archival and preservation features
5. increase interoperability through implementation of standards
6. increase contributions from the open-source development community
7. improve UX and UI
8. continue to increase the quality of the software

Throughout the year, we'll identify big steps that we can take to focus on one or more of these goals. These big steps are represented on our **Roadmap**. The Roadmap items that we're about to work on will be well defined, but those Roadmap items that are further out may just be big problems we know we need to solve in some way. Although we are committed to Roadmap items below, the timeframe of the items further out might vary slightly as critical issues, other priorities or dependencies rise.

Dataverse Waffle Board



IQSS/dataverse Add Issue Filter Board

Invite/Share
WaffleBot
Board View
Metrics
Project Settings
View on GitHub
Home
Help
mercecrosas

Community Dev … 19 7
- 5081 Documentation doesn't describe bibtex files in Dataverse bundles
- 5055 Internationalization : MimeTypeDisplay and MimeTypeFacets property files — Feature: Internationalization — 5066
- 5044 Workflow Enhancements — 5048
- 4897 Request Access: button does not launch popup if logged in and request from file landing page.

UI/UX Design … 6
- 2739 Code Deposit - Github Integration
- 3404 Dataset - File UI Improvements — Feature: File Upload & Handling — Feature: Search/Browse — Status: UX & UI
- 871 Integrate with DataTags
- 5028 Dataset External Tools
- 4949 Download Package File from S3 — 4.10 - Additional Data Transfer Options — ready for estimation

Ready 48 19

IQSS Sprint 9/19… 3 13
- 5053 Custom home page for Harvard Dataverse
- 5027 Admin Guide - Dashboard
- 4741 Demo: Address stability issues due to disk space by periodically cleaning and consider increasing space.

IQSS Team Dev 4 12
- 4832 Harvard Dataverse: Mint DOIs for datasets that currently have handles — 5064
- 5024 Move dataverse.harvard to DataCite
- 4261 Dataverse discovery in Google - Machine Readable Sitemaps — 5084
- 4966 Installer - database error detection, MacOS-specific config cases. — 5058

waffle.io + GitHub

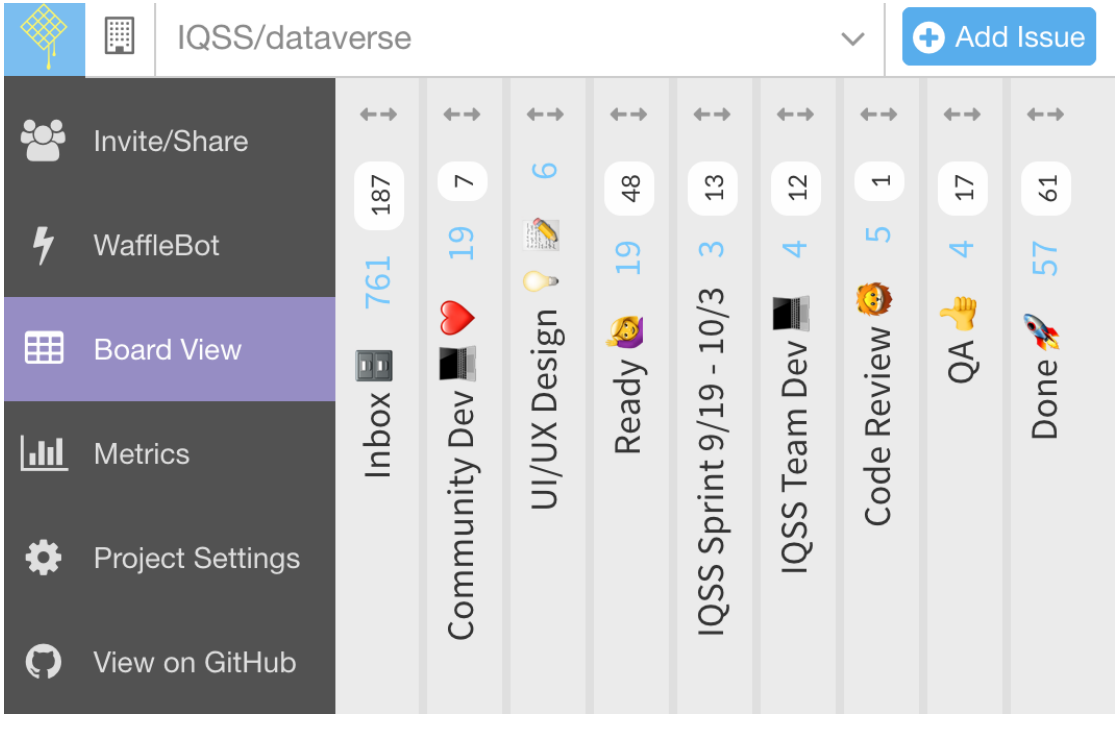
PROCESS TO SUPPORT AN AGILE DEVELOPMENT

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!

COMMUNITY

FEATURES

DATA

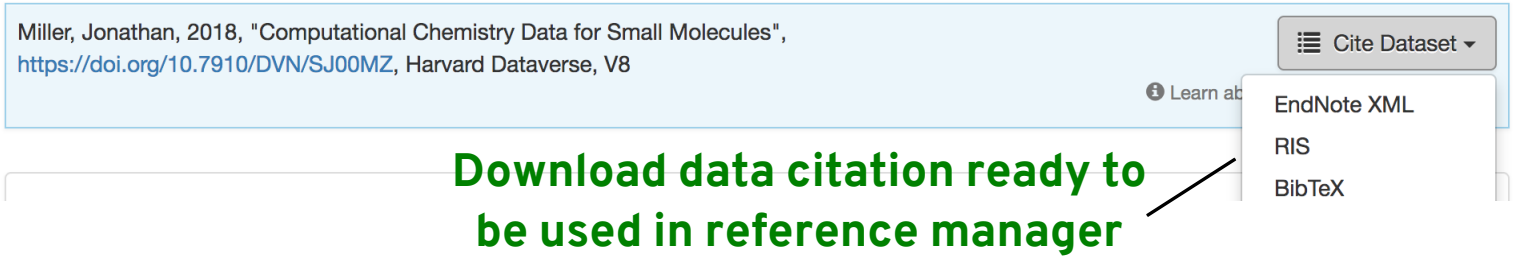
PROJECTS

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*



METADATA TO FIND AND REUSE DATA

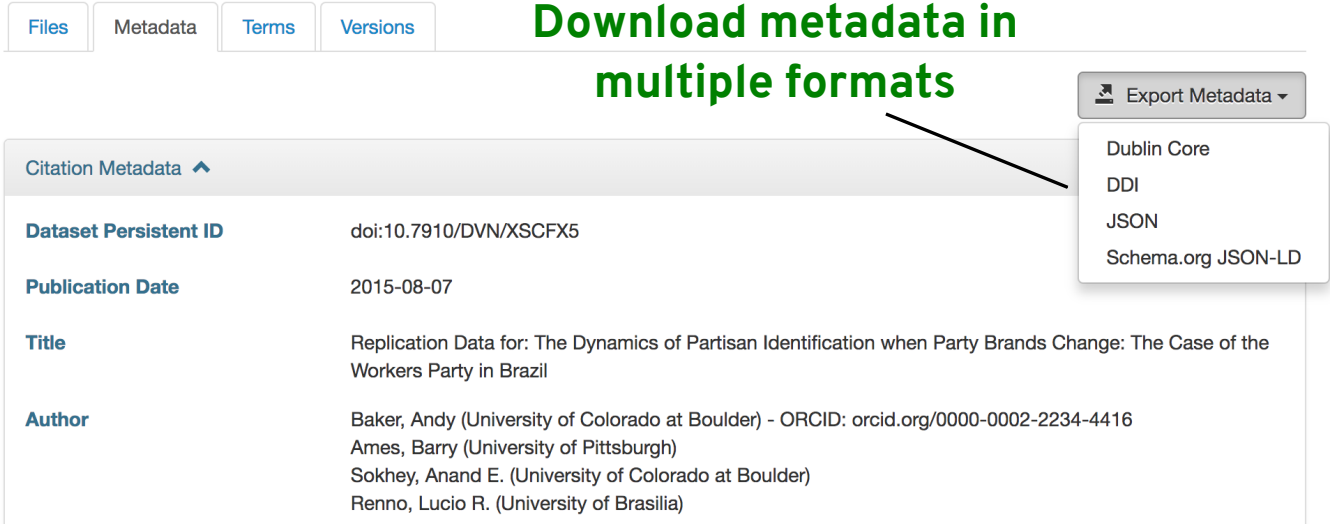
At multiple Levels:

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

With multiple Standards:

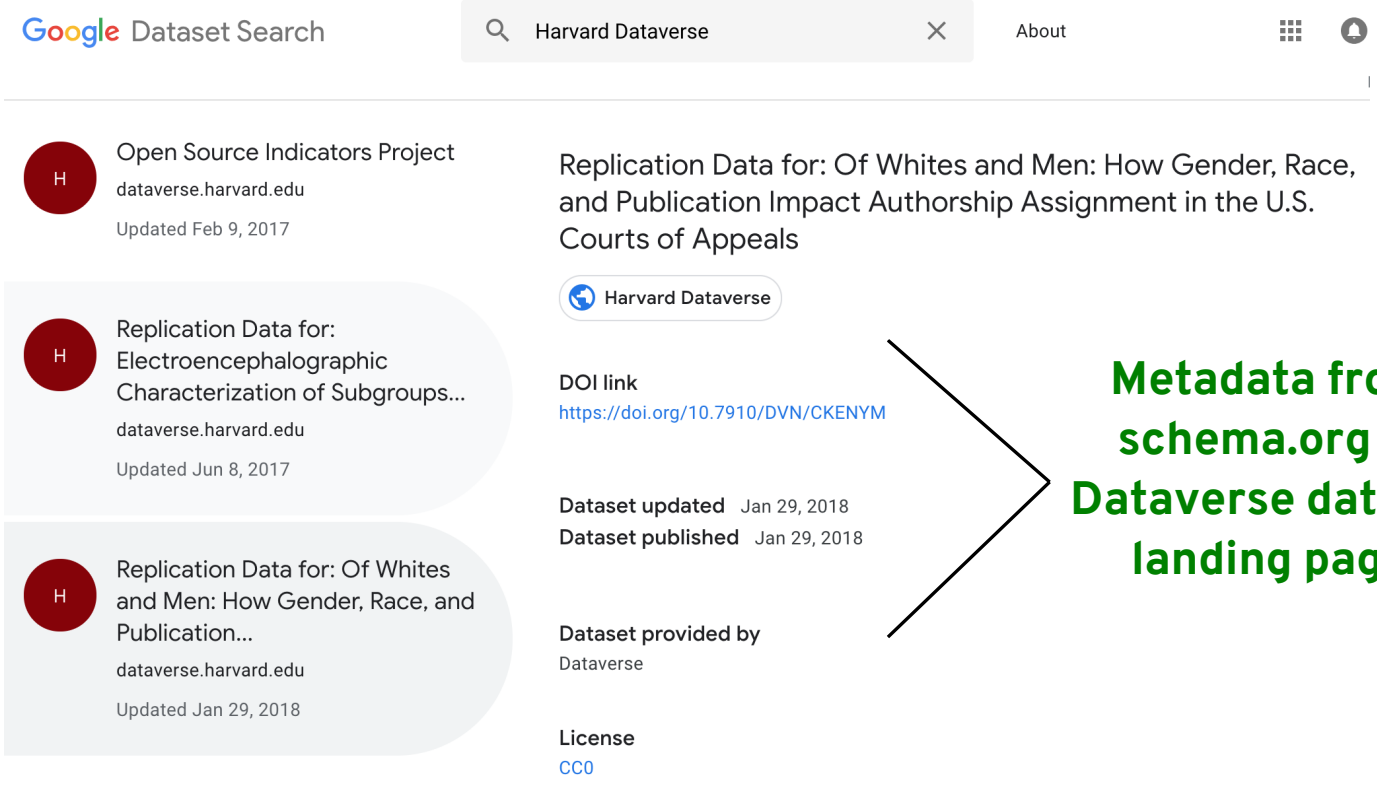
- Data Documentation Initiative (DDI)
- Dublin Core
- Schema.org

Download metadata in multiple formats



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

Files | Metadata | Terms | Versions

View Differences

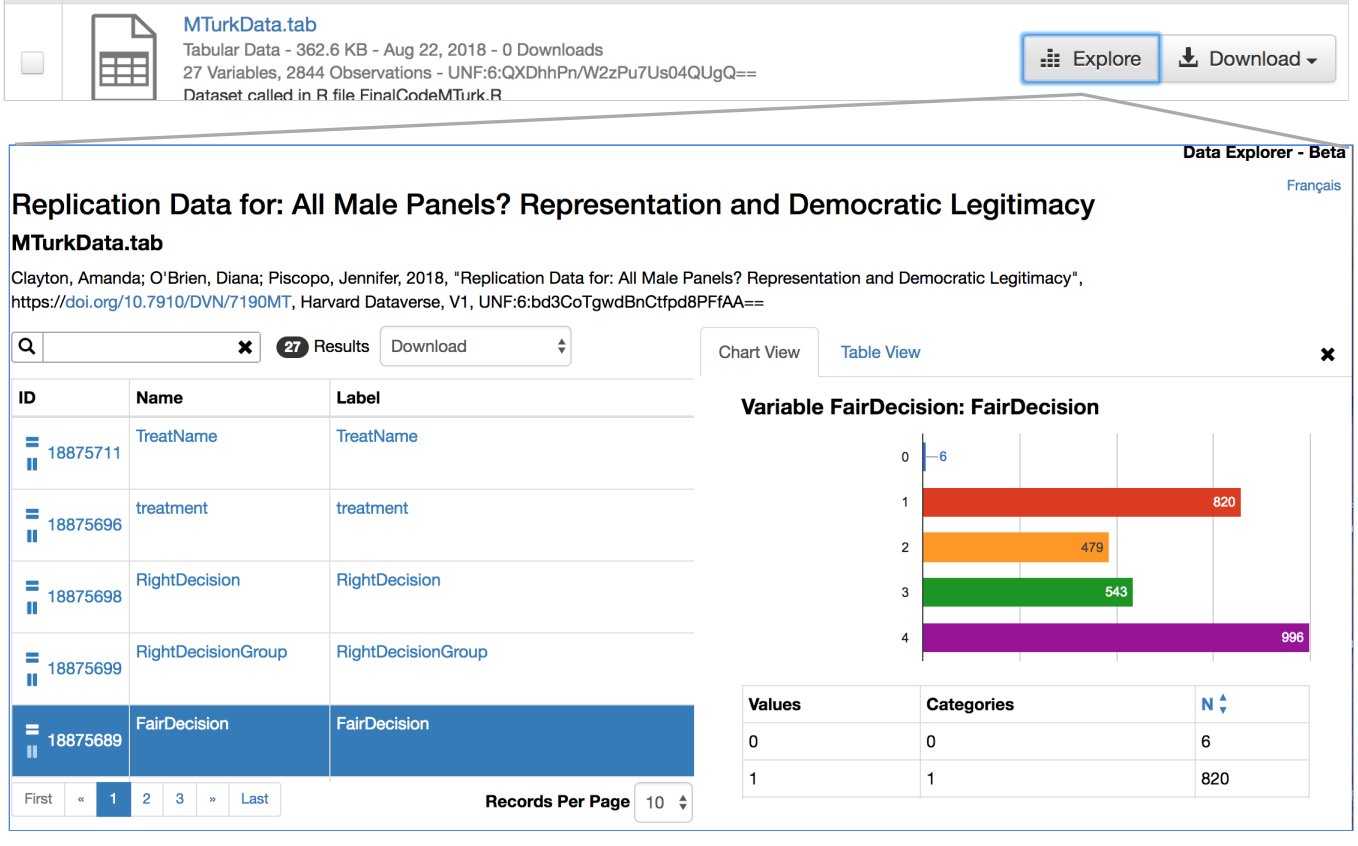
	Dataset	Summary	Contributors	Published
	8.0	**Citation Metadata:** Description (1 Changed); **Additional Citation Metadata:** (2 Added); **Files (Added: 47; Removed: 25; Changed File Metadata: 15);** View Details	Jonathan Miller	September 24, 2018
	7.0	**Citation Metadata:** Keyword (1 Changed); **Files (Added: 30);** View Details	Jonathan Miller	September 19, 2018
	6.0	**Citation Metadata:** Keyword (1 Changed); Description (1 Changed); **Files (Added: 14);** View Details	Jonathan Miller	September 16, 2018
	5.0	**Citation Metadata:** Keyword (1 Changed); **Files (Added: 6);** View Details	Jonathan Miller	September 5, 2018
	4.0	**Files (Added: 38);** View Details	Jonathan Miller	September 2, 2018
	3.0	**Files (Added: 2);** View Details	Jonathan Miller	August 30, 2018
	2.0	**Citation Metadata:** Keyword (1 Changed); **Files (Added: 6; Removed: 1);** View Details	Jonathan Miller	August 29, 2018
	1.0	This is the first published version.	Jonathan Miller	August 29, 2018

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

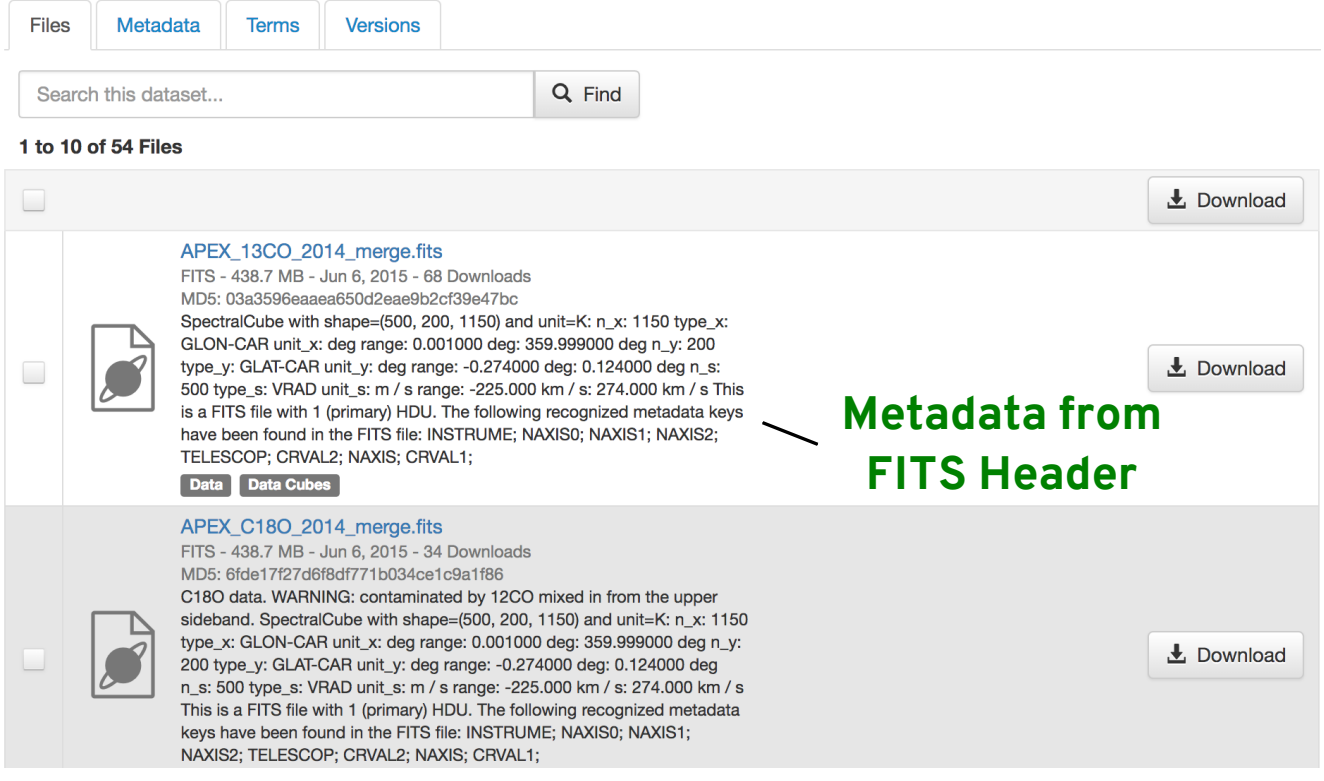
TABULAR DATA EXPLORATION

- Variable metadata automatically extracted
- Descriptive statistics automatically computed



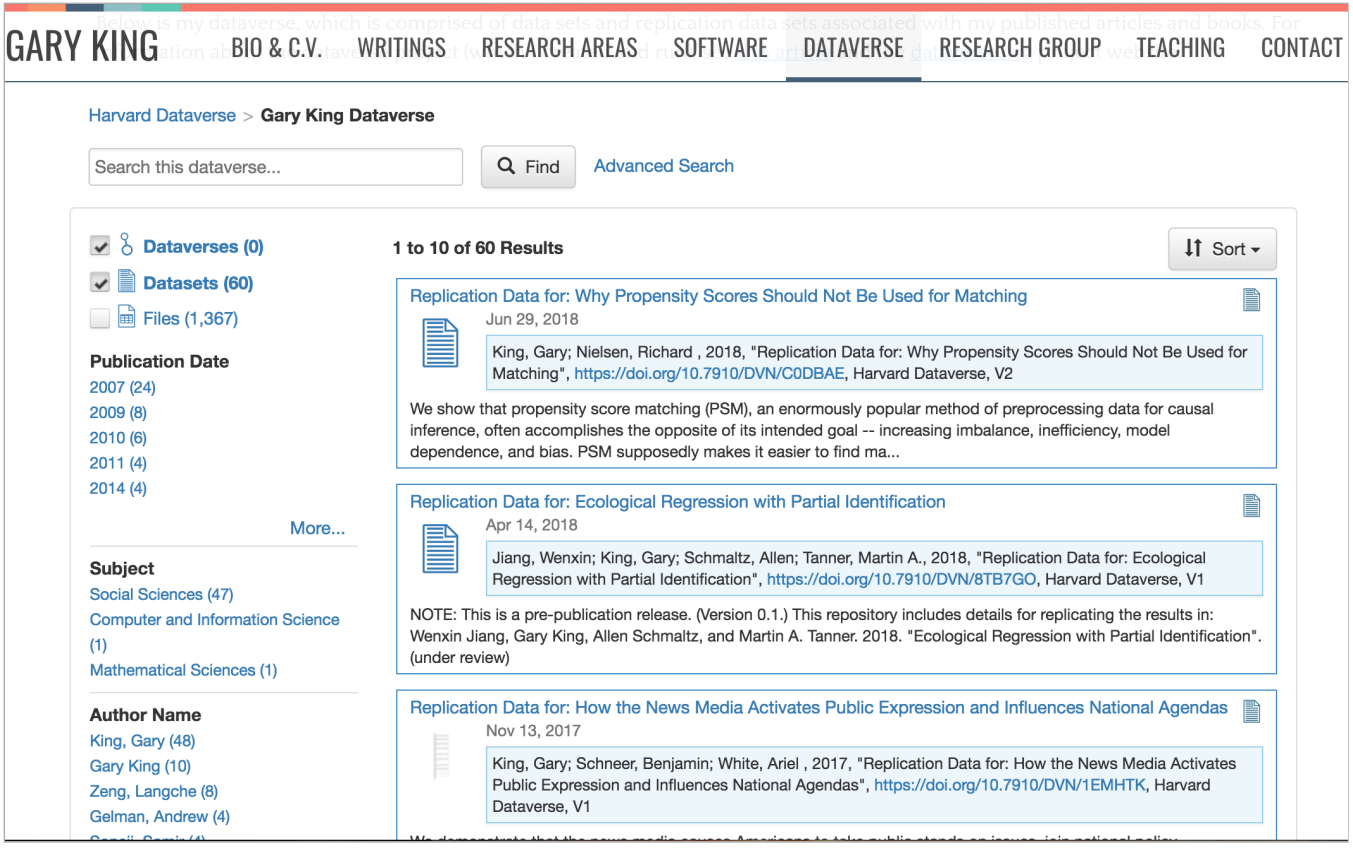
METADATA EXTRACTION FROM ASTRONOMY FILES

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



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

Open Science Framework

OJS
OPEN JOURNAL SYSTEM

RSpace

WorldMap
HARVARD UNIVERSITY

Dropbox

ORCID

COMMUNITY

FEATURES

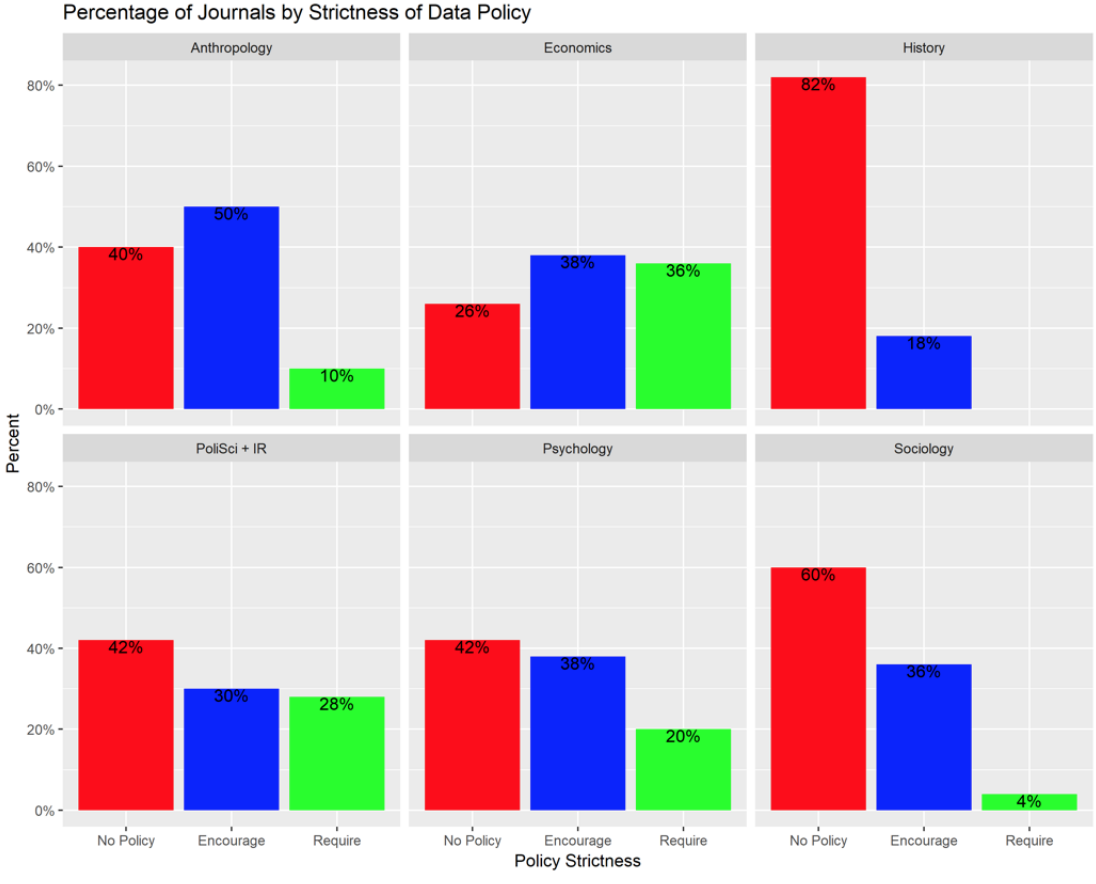
DATA

PROJECTS

A WIDE VARIETY OF DATA AND DATAVERSES

- 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.

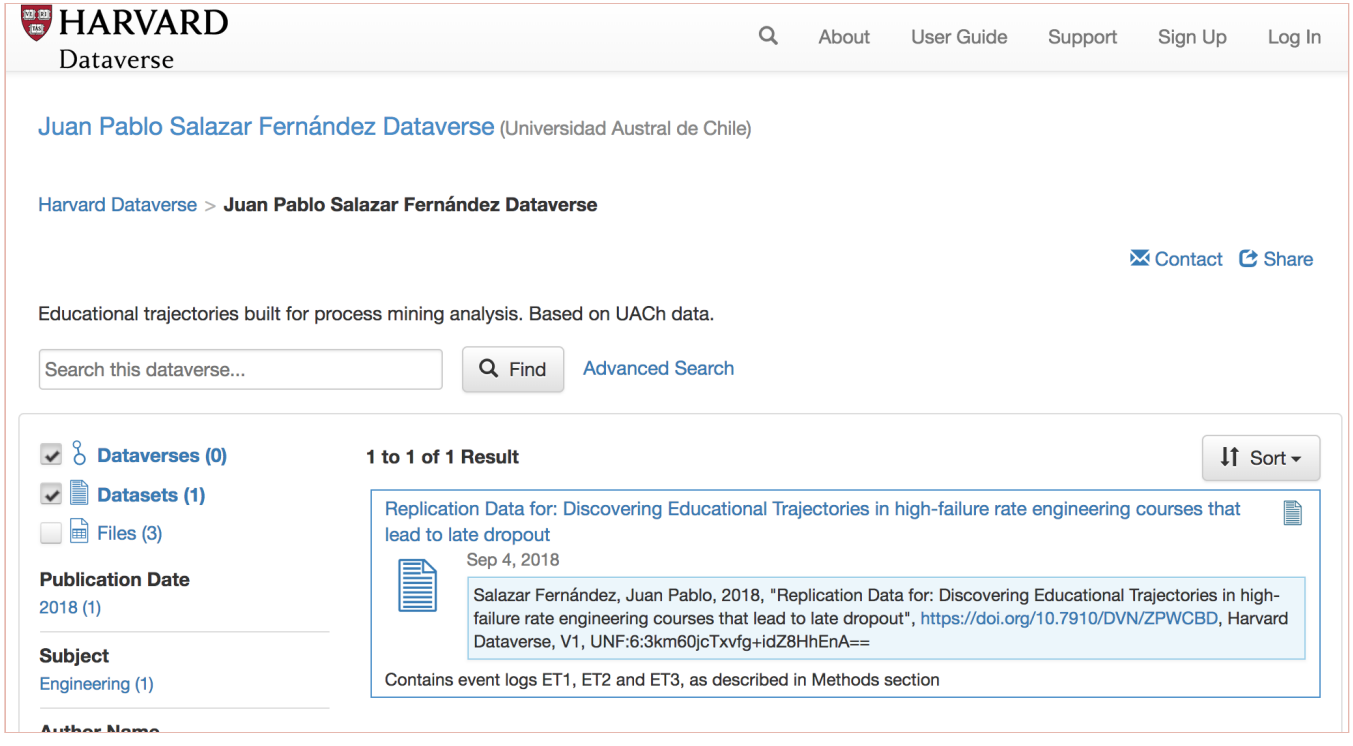
Crosas, Gautier, Karcher, Kirilova, Otalora, Schwartz, 2018.

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: STRUCURAL BIOLOGY

SBGrid Data Bank | Dataverse Pilot | Search | User Guide | Support | Log In

- Dataverses (79)
- Datasets (441)
- Files (441)

Dataverse Category
Laboratory (79)

Publication Year
2015 (210)
2016 (153)
2017 (100)
2018 (57)

Data Type
X-Ray Diffraction (418)
Micro-Electron Diffraction (13)
Structural Model (7)
XFEL Diffraction (2)
Lattice Light-Sheet Microscopy (1)

PDB ID
1V9Z (5)
1XBN (3)
3U8O (3)
3U8R (3)
5KTE (3)

1 to 10 of 520 Results | Sort

X-Ray Diffraction data from delN(40) METTL16, source of 6GFK structure
Sep 11, 2018 - Andrew McCarthy Laboratory
McCarthy, Andrew, 2018, "X-Ray Diffraction data from delN(40) METTL16, source of 6GFK structure", https://doi.org/10.15785/SBGRID/577, SBGrid Data Bank, V1
Diffraction data from crystal of del40 METTL16 MTase domain

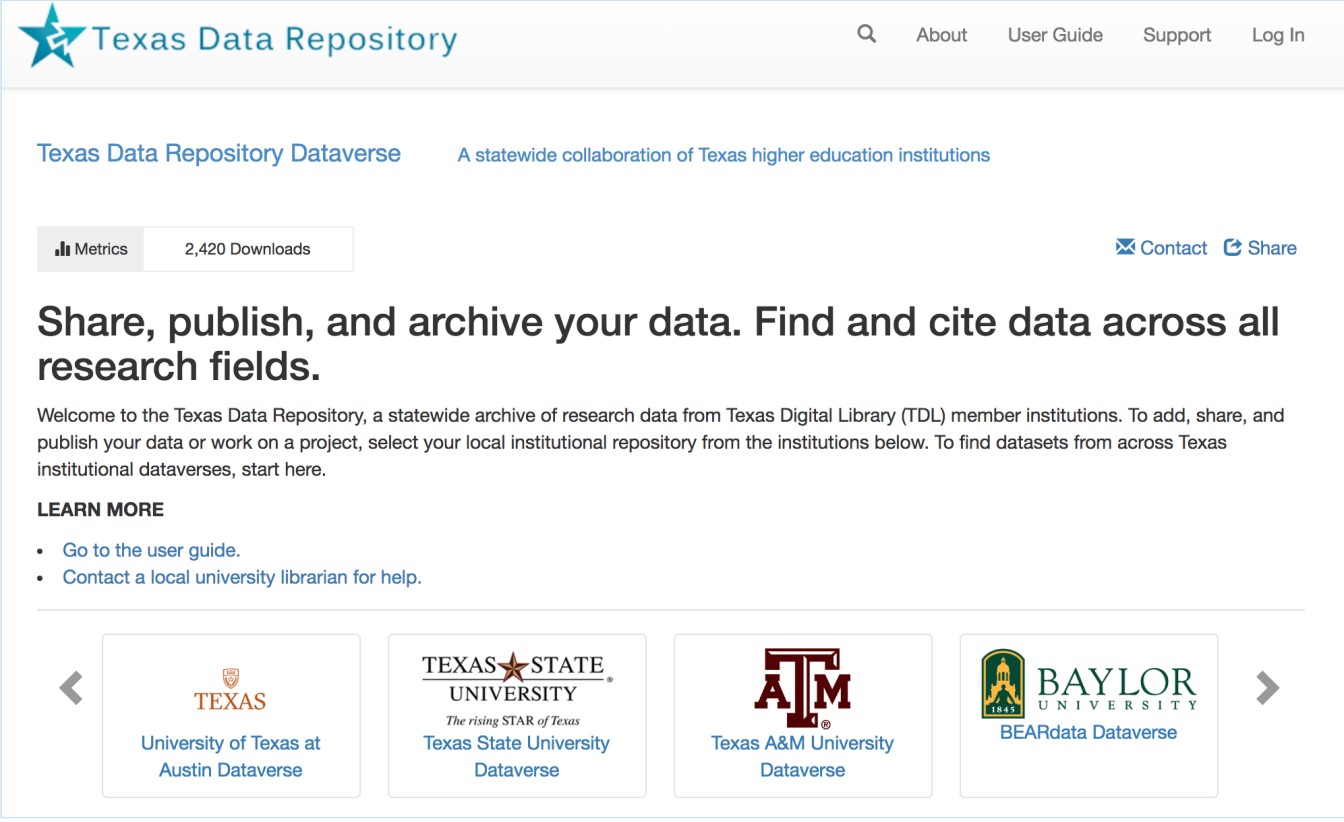
X-Ray Diffraction data from METTL16 MTase domain, source of 6GFN structure
Sep 11, 2018 - Andrew McCarthy Laboratory
McCarthy, Andrew, 2018, "X-Ray Diffraction data from METTL16 MTase domain, source of 6GFN structure", https://doi.org/10.15785/SBGRID/578, SBGrid Data Bank, V1
METTL16 MTase domain crystal form 1

X-Ray Diffraction data from METTL16 MTase domain crystal form2, source of 6GT5 structure
Sep 11, 2018 - Andrew McCarthy Laboratory
McCarthy, Andrew, 2018, "X-Ray Diffraction data from METTL16 MTase domain crystal form2, source of 6GT5 structure", https://doi.org/10.15785/SBGRID/579, SBGrid Data Bank, V1
METTL16 MTase domain crystal from 2

X-Ray Diffraction data from p190RhoGAP-A N-GTPase domain, source of 6D4G structure
Aug 31, 2018 - Titus Boggon Laboratory
Stiegler, Amy L;Boggon, Titus, 2018, "X-Ray Diffraction data from p190RhoGAP-A N-GTPase domain, source of 6D4G structure", https://doi.org/10.15785/SBGRID/575, SBGrid Data Bank, V1
8 sweeps

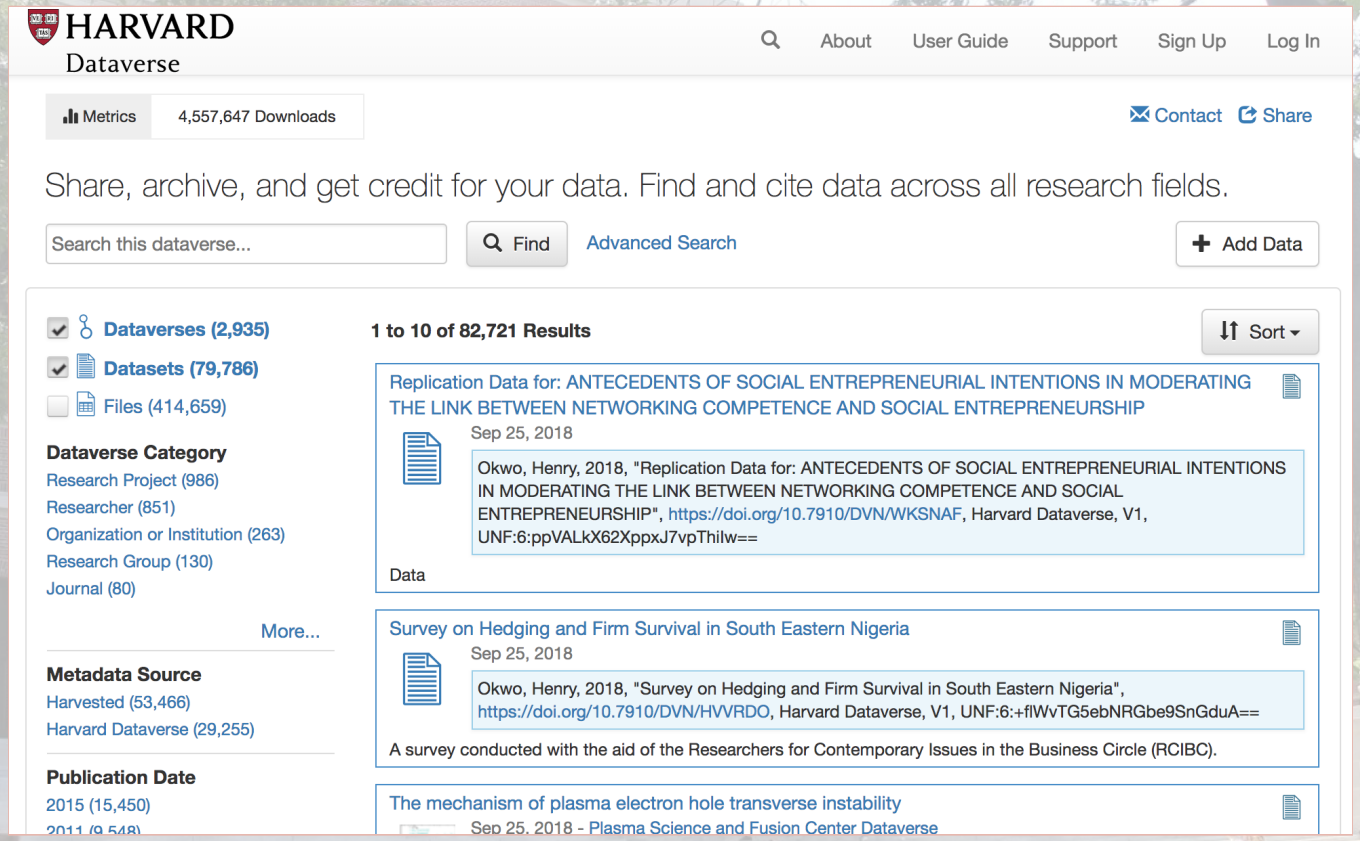
Hosted SBGrid Consortium, Harvard Medical School

DATAVERSE FOR MULTIPLE UNIVERSITIES



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

COMMUNITY

FEATURES

DATA

PROJECTS

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

SBGrid Data Bank **Dataverse Pilot**

This data file can be accessed through a terminal window, using the commands below. For more information about downloading and verifying data, see our User Guide.

Local Access

`/programs/datagrid/579`

Download Access

`rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/579` (Harvard Medical School, USA)
`rsync -av rsync://sbgrid.icm.uu.se/10.15785/SBGRID/579` (Uppsala University, Sweden)
`rsync -av rsync://sbgrid.pasteur.edu.uy/10.15785/SBGRID/579` (Institut Pasteur de Montevideo, Uruguay)
`rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/579` (Shanghai Institutes for Biological Sciences, China)

Verify Data

`cd 579 ; shasum -c files.sha`

Funding by Helmsley Charitable Trust, with focus on biomedical data, in collaboration with Piotrek Sliz

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

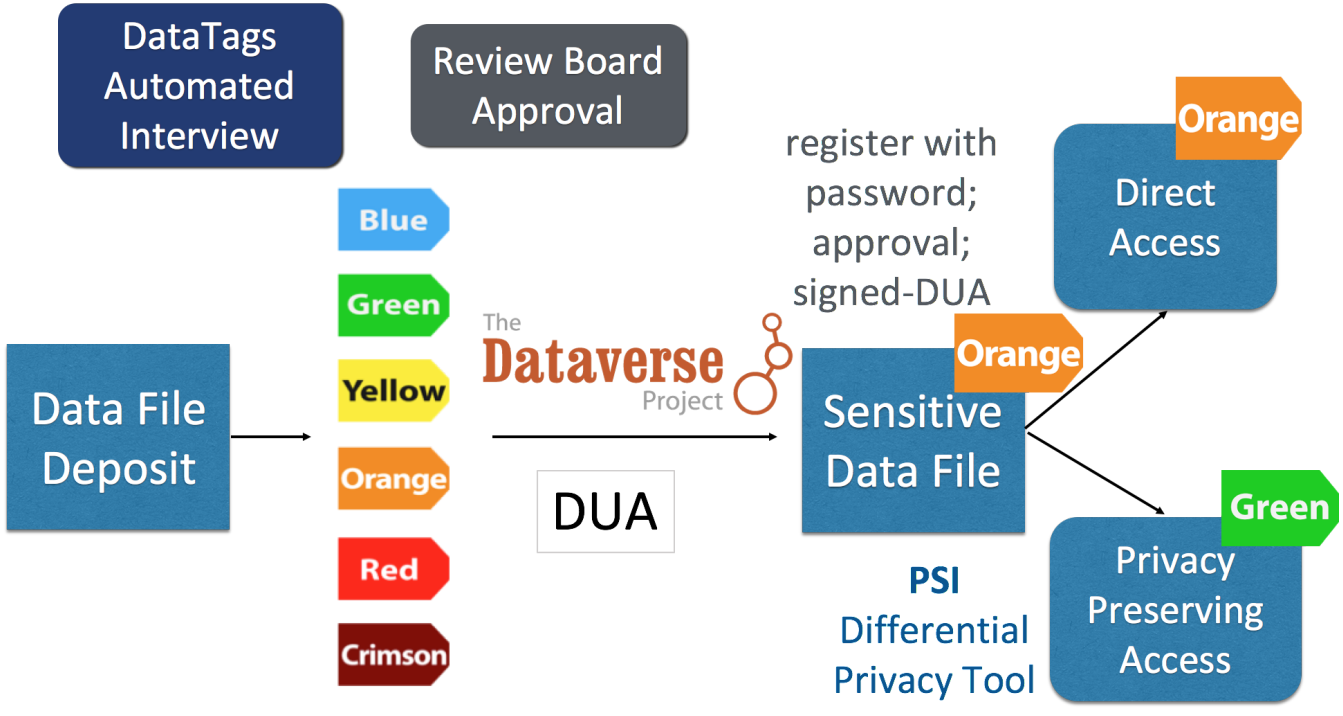
SENSITIVE DATA: DATATAGS

Standardize data security and access levels

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

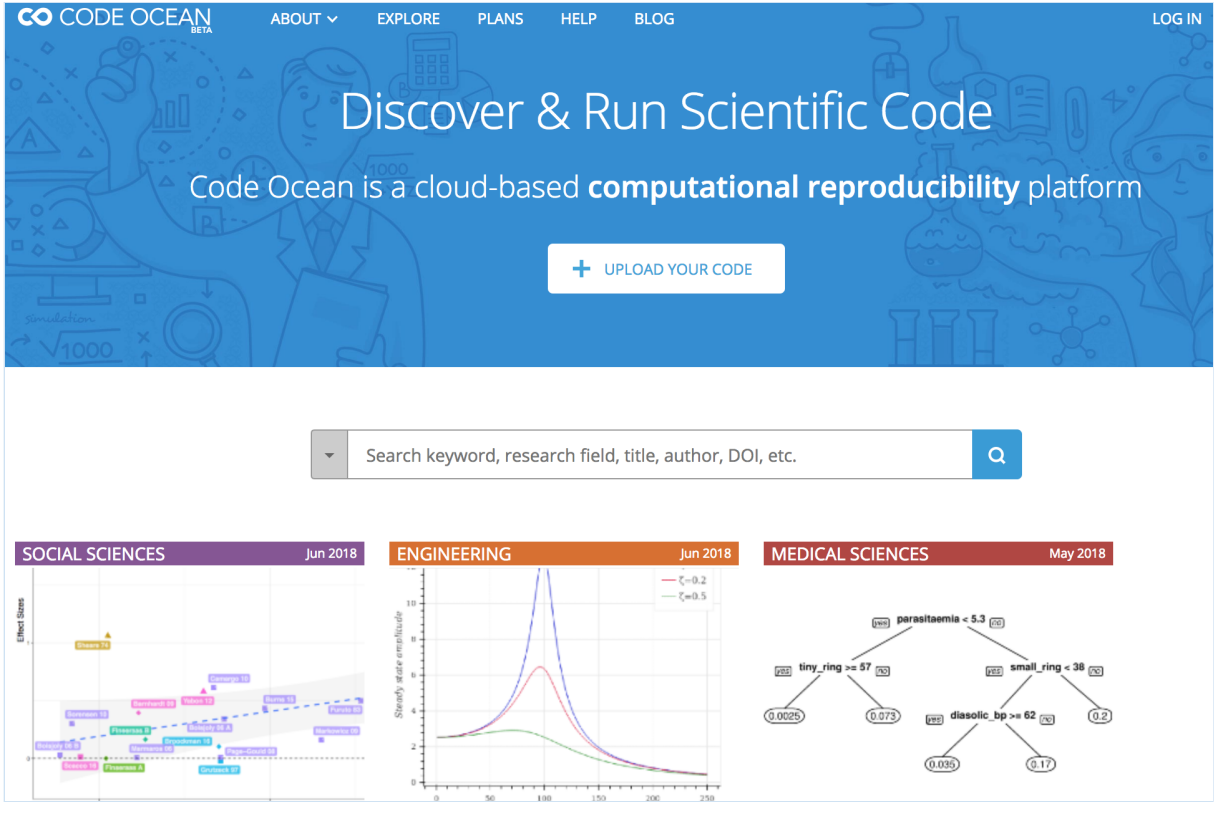
Funded by National Science Foundation,
in collaboration with Latanya Sweeney

SENSITIVE DATA: PRIVACY PRESERVING TOOLS



Funded by National Science Foundation,
in collaboration with Harvard Privacy Tools Project

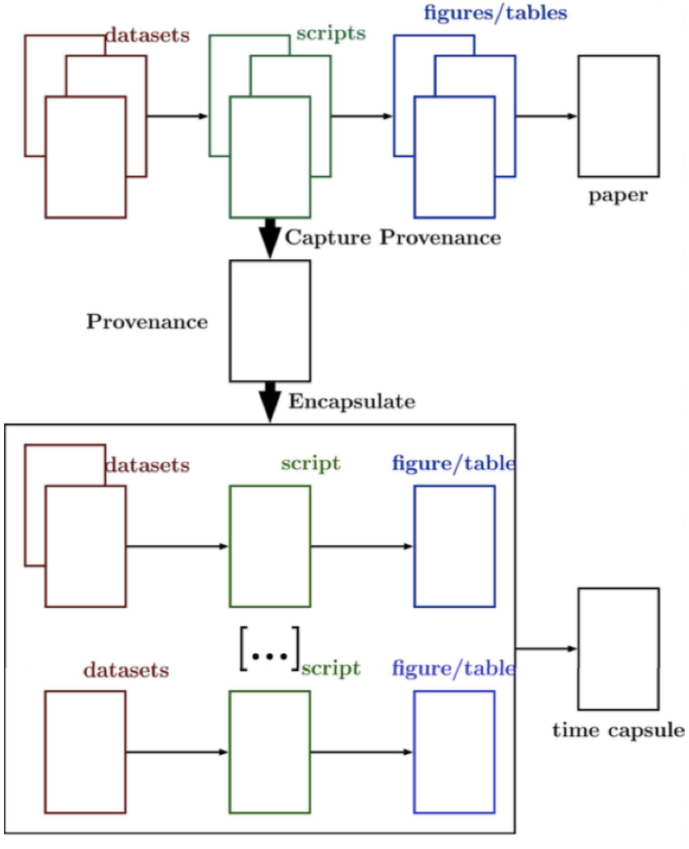
INTEGRATION WITH REPRODUCIBILITY TOOLS: CODE OCEAN



Funded by Sloan Foundation, in collaboration with CodeOcean

Alfred P. Sloan FOUNDATION

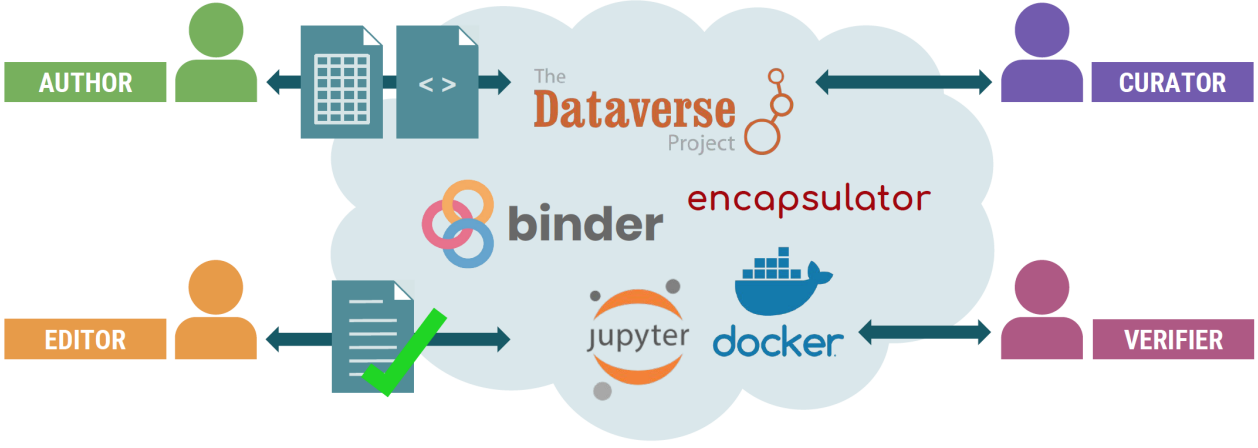
INTEGRATION WITH REPRODUCIBILITY TOOLS: ENCAPSULATOR



Funded by Sloan Foundation, in collaboration with Margo Seltzer

Alfred P. Sloan FOUNDATION

INTEGRATION WITH REPRODUCIBILITY TOOLS: CORE2



Funded by Sloan Foundation, in collaboration with the ODUM institute at UNC Chapel Hill

Alfred P. Sloan FOUNDATION

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