

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

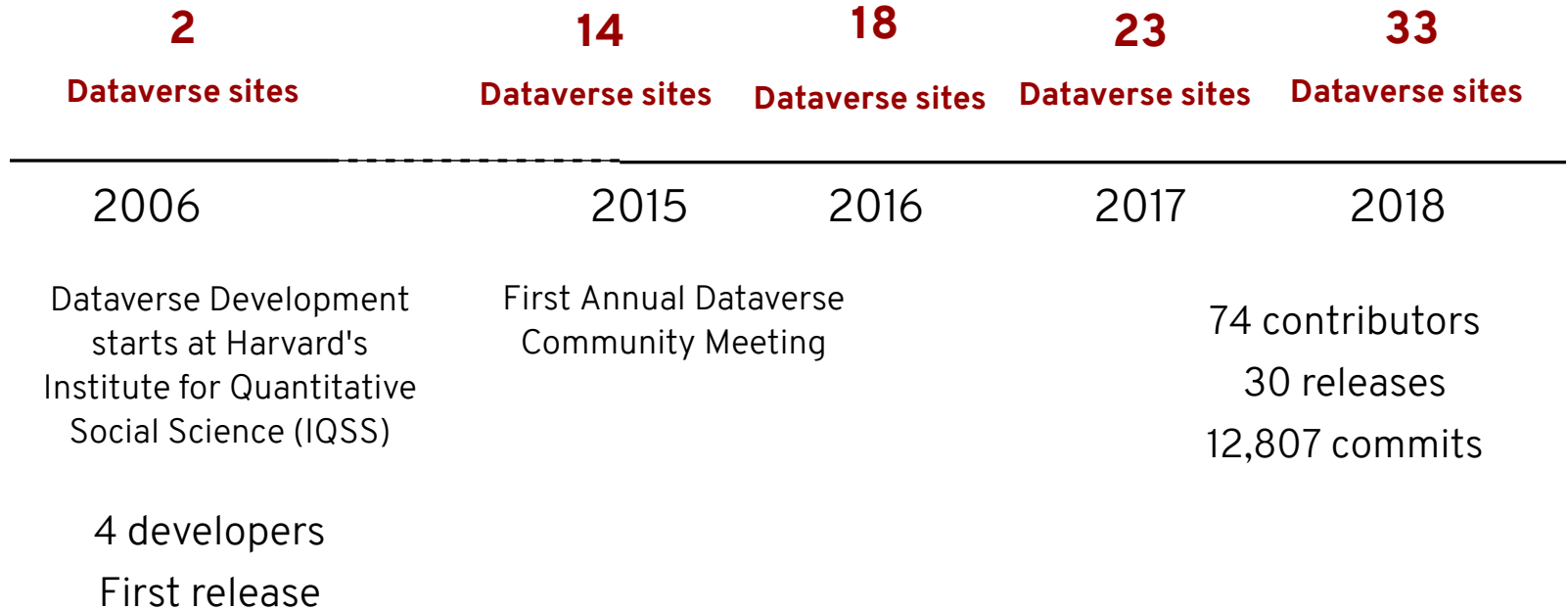
**A GROWING,
ENGAGED
COMMUNITY**

DATAVERSE.ORG

33 Dataverse Repositories sites around the world




DATAVERSE COMMUNITY GROWTH



GLOBAL DATAVERSE COMMUNITY CONSORTIUM

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



The screenshot shows the homepage of the Global Dataverse Community Consortium (GDCC). At the top left is the GDCC logo, which consists of a stylized globe made of dots in shades of orange and red, followed by the text "GDCC". To the right of the logo is the text "The Global Dataverse Community Consortium" and the tagline "Supporting Dataverse repositories around the world." Below this is a navigation menu with the following items: "Home" (highlighted in a black box), "About", "Members", "Interest Groups", "Services", "Sign-Up Forms", "News", and "Events". A gear icon for settings is located in the top right corner. The main content area features a large graphic of a world map composed of dots in various colors (orange, yellow, green). Below the map, there are two columns of text. The left column contains a paragraph: "The Global Dataverse Community Consortium (GDCC) is dedicated to providing international organization to existing Dataverse community efforts, and will provide a collaborative venue for institutions to leverage economies of scale in support of Dataverse repositories around the world." The right column contains the text "Sign up to join the GDCC here:" followed by a link labeled "Sign Up Form". At the bottom of the page, it says "A collaboration with" followed by the logo for "The Dataverse Project".

The Global Dataverse Community Consortium (GDCC) is dedicated to providing international organization to existing Dataverse community efforts, and will provide a collaborative venue for institutions to leverage economies of scale in support of Dataverse repositories around the world.

Sign up to join the GDCC here:
[Sign Up Form](#)

A collaboration with 

<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

The screenshot shows the 'Strategic Goals, Roadmap, and Releases' page on the Dataverse Project website. It lists eight strategic goals and provides a brief overview of the roadmap process.

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

The screenshot shows a Waffle Board for the 'IQSS/dataverse' project. The board is organized into columns representing different development phases: 'Community Dev', 'UI/UX Design', 'IQSS Sprint 9/19...', and 'IQSS Team Dev'. Each column contains a grid of issues, with progress bars indicating the status of each issue. The issues are numbered and include titles such as 'Documentation doesn't describe bibtext files in Dataverse bundles', 'Code Deposit - Github Integration', 'Dataset - File UI Improvements', 'Admin Guide - Dashboard', 'Demo: Address stability issues due to disk space by periodically cleaning and consider increasing space.', and 'Request Access: button does not launch popup if logged in and request from file landing page.'.



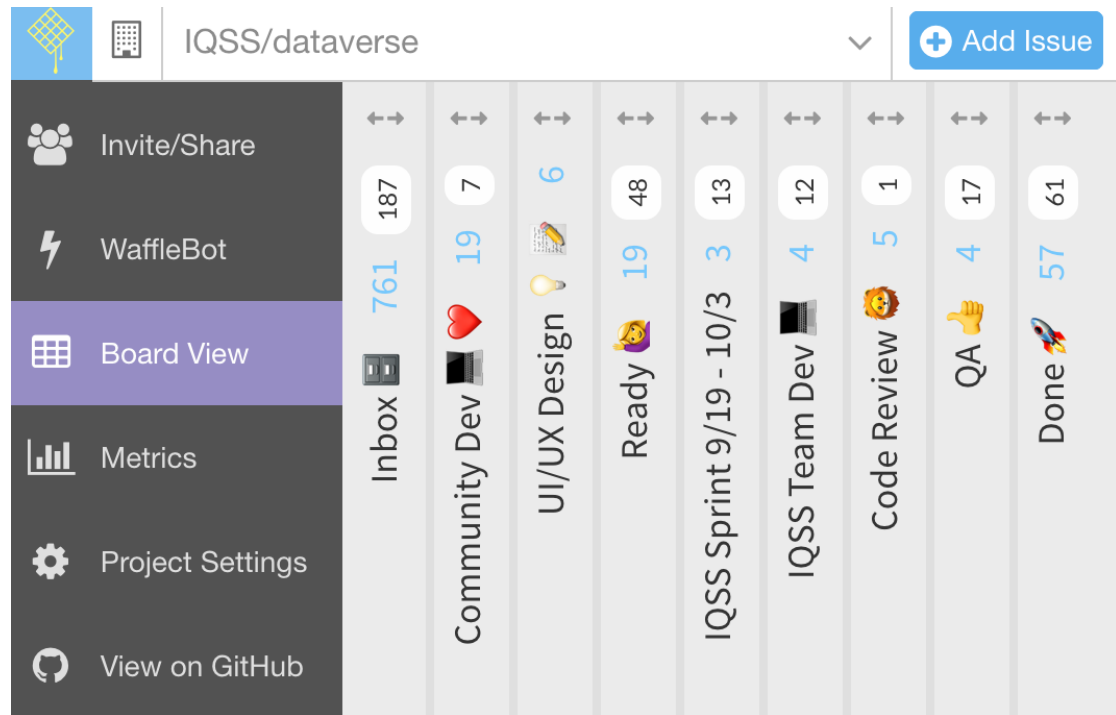
PROCESS TO SUPPORT AN AGILE DEVELOPMENT

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*

Miller, Jonathan, 2018, "Computational Chemistry Data for Small Molecules",
<https://doi.org/10.7910/DVN/SJ00MZ>, Harvard Dataverse, V8

☰ Cite Dataset ▾

Learn about

EndNote XML

RIS

BibTeX

**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

The screenshot shows a metadata page with a navigation bar containing 'Files', 'Metadata', 'Terms', and 'Versions'. A green text overlay reads 'Download metadata in multiple formats' with an arrow pointing to the 'Export Metadata' dropdown menu. The dropdown menu lists 'Dublin Core', 'DDI', 'JSON', and 'Schema.org JSON-LD'. Below the navigation bar, the 'Citation Metadata' section is expanded, showing the following information:

Dataset Persistent ID	doi:10.7910/DVN/XSCFX5
Publication Date	2015-08-07
Title	Replication Data for: The Dynamics of Partisan Identification when Party Brands Change: The Case of the Workers Party in Brazil
Author	Baker, Andy (University of Colorado at Boulder) - ORCID: orcid.org/0000-0002-2234-4416 Ames, Barry (University of Pittsburgh) Sokhey, Anand E. (University of Colorado at Boulder) Renno, Lucio R. (University of Brasilia)

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

The image shows a screenshot of the Google Dataset Search interface. The search bar contains 'Harvard Dataverse'. Below the search bar, three search results are listed. The first result is 'Open Source Indicators Project' from dataverse.harvard.edu, updated Feb 9, 2017. The second result is 'Replication Data for: Electroencephalographic Characterization of Subgroups...' from dataverse.harvard.edu, updated Jun 8, 2017. The third result is 'Replication Data for: Of Whites and Men: How Gender, Race, and Publication...' from dataverse.harvard.edu, updated Jan 29, 2018. A callout box highlights the second and third results. To the right of the callout box, a detailed view of the third result is shown, including a 'Harvard Dataverse' badge, a DOI link (<https://doi.org/10.7910/DVN/CKENYM>), 'Dataset updated Jan 29, 2018', 'Dataset published Jan 29, 2018', 'Dataset provided by Dataverse', and 'License CC0'. A large green arrow points from the callout box to the text 'Metadata from schema.org in Dataverse dataset landing page'.

Google Dataset Search

Harvard Dataverse

About

Open Source Indicators Project
dataverse.harvard.edu
Updated Feb 9, 2017

Replication Data for: Electroencephalographic Characterization of Subgroups...
dataverse.harvard.edu
Updated Jun 8, 2017

Replication Data for: Of Whites and Men: How Gender, Race, and Publication...
dataverse.harvard.edu
Updated Jan 29, 2018

Replication Data for: Of Whites and Men: How Gender, Race, and Publication Impact Authorship Assignment in the U.S. Courts of Appeals

Harvard Dataverse

DOI link
<https://doi.org/10.7910/DVN/CKENYM>

Dataset updated Jan 29, 2018
Dataset published Jan 29, 2018

Dataset provided by
Dataverse

License
CC0

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
<input type="checkbox"/>	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
<input type="checkbox"/>	7.0	Citation Metadata: Keyword (1 Changed); Files (Added: 30); View Details	Jonathan Miller	September 19, 2018
<input type="checkbox"/>	6.0	Citation Metadata: Keyword (1 Changed); Description (1 Changed); Files (Added: 14); View Details	Jonathan Miller	September 16, 2018
<input type="checkbox"/>	5.0	Citation Metadata: Keyword (1 Changed); Files (Added: 6); View Details	Jonathan Miller	September 5, 2018
<input type="checkbox"/>	4.0	Files (Added: 38); View Details	Jonathan Miller	September 2, 2018
<input type="checkbox"/>	3.0	Files (Added: 2); View Details	Jonathan Miller	August 30, 2018
<input type="checkbox"/>	2.0	Citation Metadata: Keyword (1 Changed); Files (Added: 6; Removed: 1); View Details	Jonathan Miller	August 29, 2018
<input type="checkbox"/>	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

<input type="checkbox"/>	 <p>00613Newberger-Childrearing-StudyDescription.pdf Adobe PDF - 17.7 KB - Nov 27, 2007 - 11 Downloads MD5: 7f06518a663078310cf8399b9cbf5340 Overview: abstract, research methodology, publications, and other info.</p> <p>1. Documentation</p>	Public	 Download
<input type="checkbox"/>	 <p>00613Newberger-ChildRearing-Subjects_CP-01 thru CP-09.pdf Adobe PDF - 2.5 MB - Sep 14, 2018 - 0 Downloads MD5: c5411b2cc2b4cbacd8f9d6c2d8b073e6 Subjects CP-01, 02, 03, 04, 05, 06, 07, 08, and 09</p> <p>2. Data</p>	Restricted	
<input type="checkbox"/>	 <p>00613Newberger-ChildRearing-Subjects_P-01 thru P-06.pdf Adobe PDF - 1.3 MB - Sep 14, 2018 - 0 Downloads MD5: ada399d924ec47a2c0a1c127d97971f5 Subjects P-01, 02, 04, 05, and 06</p> <p>2. Data</p>		
<input type="checkbox"/>	 <p>00613Newberger-ChildRearing-Subjects_P-07 thru P-11.pdf Adobe PDF - 2.1 MB - Sep 14, 2018 - 0 Downloads MD5: 3ad3dc5aa1bad7deb31da4b32af4747d Subjects P-07, 08, 09, 10, and 11</p> <p>2. Data</p>		

TABULAR DATA EXPLORATION

- Variable metadata automatically extracted
- Descriptive statistics automatically computed


MTurkData.tab
 Tabular Data - 362.6 KB - Aug 22, 2018 - 0 Downloads
 27 Variables, 2844 Observations - UNF:6:QXDhhPn/W2zPu7Us04QUgQ==
 Dataset called in R file FinalCodeMTurk.R

[Explore](#)
[Download](#)

Data Explorer - Beta [Français](#)

Replication Data for: All Male Panels? Representation and Democratic Legitimacy

MTurkData.tab

Clayton, Amanda; O'Brien, Diana; Piscopo, Jennifer, 2018, "Replication Data for: All Male Panels? Representation and Democratic Legitimacy", <https://doi.org/10.7910/DVN/7190MT>, Harvard Dataverse, V1, UNF:6:bd3CoTgwdBnCtfd8PFfAA==

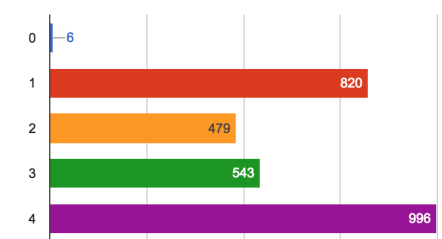
27 Results
Download

Chart View
Table View

ID	Name	Label
18875711	TreatName	TreatName
18875696	treatment	treatment
18875698	RightDecision	RightDecision
18875699	RightDecisionGroup	RightDecisionGroup
18875689	FairDecision	FairDecision

First « 1 2 3 » Last
 Records Per Page 10

Variable FairDecision: FairDecision



Values	Categories	N
0	0	6
1	1	820




METADATA EXTRACTION FROM ASTRONOMY FILES

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

Files Metadata Terms Versions

Search this dataset... Find

1 to 10 of 54 Files

<input type="checkbox"/>	 Download
<input type="checkbox"/>	<p>APEX_13CO_2014_merge.fits FITS - 438.7 MB - Jun 6, 2015 - 68 Downloads MD5: 03a3596eaaaa650d2eae9b2cf39e47bc SpectralCube with shape=(500, 200, 1150) and unit=K: n_x: 1150 type_x: GLON-CAR unit_x: deg range: 0.001000 deg: 359.999000 deg n_y: 200 type_y: GLAT-CAR unit_y: deg range: -0.274000 deg: 0.124000 deg n_s: 500 type_s: VRAD unit_s: m / s range: -225.000 km / s: 274.000 km / s This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: INSTRUME; NAXIS0; NAXIS1; NAXIS2; TELESCOP; CRVAL2; NAXIS; CRVAL1;</p> <p>Data Data Cubes</p>  Download
<input type="checkbox"/>	<p>APEX_C18O_2014_merge.fits FITS - 438.7 MB - Jun 6, 2015 - 34 Downloads MD5: 6fde17f27d6f8df771b034ce1c9a1f86 C18O data. WARNING: contaminated by 12CO mixed in from the upper sideband. SpectralCube with shape=(500, 200, 1150) and unit=K: n_x: 1150 type_x: GLON-CAR unit_x: deg range: 0.001000 deg: 359.999000 deg n_y: 200 type_y: GLAT-CAR unit_y: deg range: -0.274000 deg: 0.124000 deg n_s: 500 type_s: VRAD unit_s: m / s range: -225.000 km / s: 274.000 km / s This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: INSTRUME; NAXIS0; NAXIS1; NAXIS2; TELESCOP; CRVAL2; NAXIS; CRVAL1;</p>  Download

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

The screenshot shows the Harvard Dataverse interface for a user named Gary King. The navigation menu at the top includes: GARY KING, BIO & C.V., WRITINGS, RESEARCH AREAS, SOFTWARE, DATAVERSE, RESEARCH GROUP, TEACHING, and CONTACT. The main content area is titled "Harvard Dataverse > Gary King Dataverse". It features a search bar with the text "Search this dataverse...", a "Find" button, and a link to "Advanced Search". On the left sidebar, there are filters for "Dataverses (0)", "Datasets (60)", and "Files (1,367)". Below these are filters for "Publication Date" (2007 (24), 2009 (8), 2010 (6), 2011 (4), 2014 (4)) and "Subject" (Social Sciences (47), Computer and Information Science (1), Mathematical Sciences (1)). The main content area displays "1 to 10 of 60 Results" and a "Sort" button. Three dataset entries are visible:

- Replication Data for: Why Propensity Scores Should Not Be Used for Matching**
Jun 29, 2018
King, Gary; Nielsen, Richard , 2018, "Replication Data for: Why Propensity Scores Should Not Be Used for Matching", <https://doi.org/10.7910/DVN/C0DBAE>, Harvard Dataverse, V2
We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal -- increasing imbalance, inefficiency, model dependence, and bias. PSM supposedly makes it easier to find ma...
- Replication Data for: Ecological Regression with Partial Identification**
Apr 14, 2018
Jiang, Wenxin; King, Gary; Schmaltz, Allen; Tanner, Martin A., 2018, "Replication Data for: Ecological Regression with Partial Identification", <https://doi.org/10.7910/DVN/8TB7GO>, Harvard Dataverse, V1
NOTE: This is a pre-publication release. (Version 0.1.) This repository includes details for replicating the results in: Wenxin Jiang, Gary King, Allen Schmaltz, and Martin A. Tanner. 2018. "Ecological Regression with Partial Identification". (under review)
- Replication Data for: How the News Media Activates Public Expression and Influences National Agendas**
Nov 13, 2017
King, Gary; Schneer, Benjamin; White, Ariel , 2017, "Replication Data for: How the News Media Activates Public Expression and Influences National Agendas", <https://doi.org/10.7910/DVN/1EMHTK>, Harvard Dataverse, V1

EXTENSIVE API TO ENABLE TOOL INTEGRATION



API Guide

Introduction

SWORD API

Search API

Data Access API

Native API

Metrics API

Client Libraries

Apps

<http://guides.dataverse.org>



Open Science Framework



COMMUNITY

FEATURES

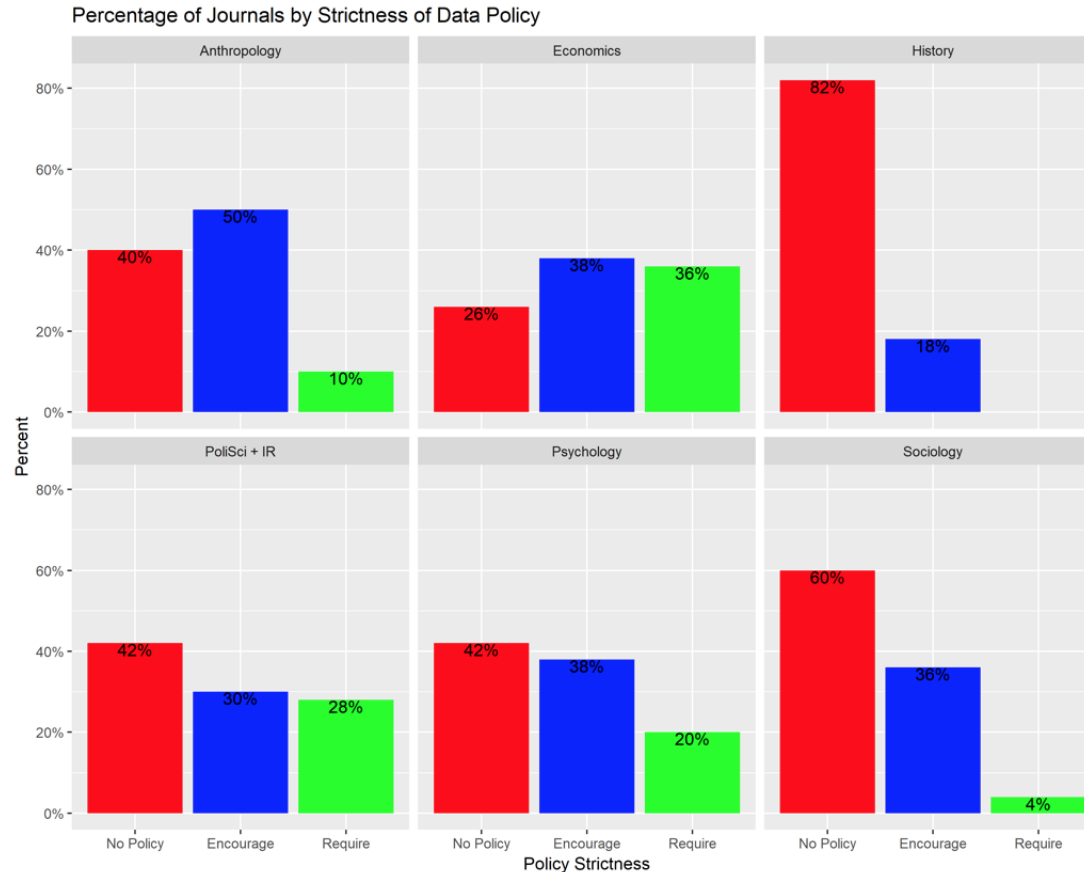
DATA

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

The screenshot shows the Harvard Dataverse interface for the American Journal of Political Science (AJPS) Dataverse. At the top, the Harvard logo and 'Dataverse' text are on the left, and navigation links (About, User Guide, Support, Sign Up, Log In) are on the right. A large green banner features the 'AJPS' logo and the text 'AMERICAN JOURNAL of POLITICAL SCIENCE'. Below the banner, it identifies the 'American Journal of Political Science (AJPS) Dataverse (Michigan State University)' with the URL 'ajps.org'. A breadcrumb trail reads 'Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse'. On the right, there are 'Contact' and 'Share' icons. A paragraph describes the journal's commitment to significant advances in knowledge and understanding of citizenship, governance, and politics, and provides a link to their website. Below this is a search bar with the text 'Search this dataverse...', a 'Find' button, an 'Advanced Search' link, and an 'Add Data' button. On the left, there are filters for 'Dataverses (0)', 'Datasets (334)', and 'Files (4,418)'. A 'Publication Date' filter shows counts for 2014 (67), 2013 (64), and 2015 (61). The main content area shows '1 to 10 of 334 Results' and a 'Sort' dropdown. The first result is titled 'Replication Data for: Racial or Spatial Voting? The Effects of Candidate Ethnicity and Ethnic Group Endorsements in Local Elections', dated Sep 20, 2018. The authors listed are Boudreau, Cheryl; Elmendorf, Christopher; MacKenzie, Scott. The abstract text is partially visible: 'With the growth of Election and Asian American populations, candidates frequently must appeal to diverse electorates'.

Hosted at Harvard Dataverse repository **(80 journal dataverses)**

DATAVERSE FOR A RESEARCHER

The screenshot shows the Harvard Dataverse website. At the top left is the Harvard logo and the text "HARVARD Dataverse". To the right are navigation links: "About", "User Guide", "Support", "Sign Up", and "Log In". Below the header, the user's name "Juan Pablo Salazar Fernández Dataverse" is displayed, followed by "(Universidad Austral de Chile)". A breadcrumb trail reads "Harvard Dataverse > Juan Pablo Salazar Fernández Dataverse". On the right side, there are "Contact" and "Share" buttons. The main content area starts with the description "Educational trajectories built for process mining analysis. Based on UCh data." Below this is a search bar with the placeholder "Search this dataverse...", a "Find" button, and a link to "Advanced Search". On the left side, there are filters for "Dataverses (0)", "Datasets (1)", and "Files (3)". Under "Publication Date", it shows "2018 (1)". Under "Subject", it shows "Engineering (1)". The main results area shows "1 to 1 of 1 Result" and a "Sort" button. The single result is a dataset titled "Replication Data for: Discovering Educational Trajectories in high-failure rate engineering courses that lead to late dropout", dated "Sep 4, 2018". The dataset description includes the author "Salazar Fernández, Juan Pablo, 2018", the title, a DOI link "https://doi.org/10.7910/DVN/ZPWCBBD", and the version "Harvard Dataverse, V1, UNF:6:3km60jcTxvfg+idZ8HhEnA==". A note at the bottom of the result box states "Contains event logs ET1, ET2 and ET3, as described in Methods section".

Hosted at Harvard Dataverse repository

DATAVERSE FOR A RESEARCH COMMUNITY: STRUCTURAL BIOLOGY

The screenshot displays the SBGrid Data Bank Dataverse Pilot interface. The header includes the SBGrid logo, 'Data Bank', and 'Dataverse Pilot' in a red box. Navigation links for 'Search', 'User Guide', 'Support', and 'Log In' are visible. The left sidebar contains filters for 'Dataverses (79)', 'Datasets (441)', and 'Files (441)'. Under 'Dataverse Category', 'Laboratory (79)' is selected. The 'Publication Year' filter shows counts for 2015 (210), 2016 (153), 2017 (100), and 2018 (57). The 'Data Type' filter lists 'X-Ray Diffraction (418)', 'Micro-Electron Diffraction (13)', 'Structural Model (7)', 'XFEL Diffraction (2)', and 'Lattice Light-Sheet Microscopy (1)'. The 'PDB ID' filter lists '1V9Z (5)', '1XBN (3)', '3U8O (3)', '3U8R (3)', and '5KTE (3)'. The main content area shows '1 to 10 of 520 Results' with a 'Sort' button. Four search results are displayed, each with a thumbnail, title, date, author, citation text with a DOI link, and a description. The results are: 1. 'X-Ray Diffraction data from delN(40) METTL16, source of 6GFK structure' by Andrew McCarthy Laboratory, dated Sep 11, 2018. 2. 'X-Ray Diffraction data from METTL16 MTase domain, source of 6GFN structure' by Andrew McCarthy Laboratory, dated Sep 11, 2018. 3. 'X-Ray Diffraction data from METTL16 MTase domain crystal form2, source of 6GT5 structure' by Andrew McCarthy Laboratory, dated Sep 11, 2018. 4. 'X-Ray Diffraction data from p190RhoGAP-A N-GTPase domain, source of 6D4G structure' by Titus Boggon Laboratory, dated Aug 31, 2018.

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

The screenshot shows the Texas Data Repository website. At the top left is the logo, a blue star with a white 'T' and 'D' inside, followed by the text 'Texas Data Repository'. To the right of the logo are navigation links: a magnifying glass icon, 'About', 'User Guide', 'Support', and 'Log In'. Below the navigation bar, the main heading is 'Texas Data Repository Dataverse' with a subtitle 'A statewide collaboration of Texas higher education institutions'. A metrics box shows 'Metrics' and '2,420 Downloads'. To the right of the metrics are 'Contact' and 'Share' links. The main content area features a large heading: 'Share, publish, and archive your data. Find and cite data across all research fields.' Below this is a paragraph: '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.' Underneath is a 'LEARN MORE' section with two bullet points: 'Go to the user guide.' and 'Contact a local university librarian for help.' At the bottom, there is a carousel of four institutional dataverse logos: 'University of Texas at Austin Dataverse', 'Texas State University Dataverse' (with the tagline 'The rising STAR of Texas'), 'Texas A&M University Dataverse', and 'Baylor University BEARdata Dataverse'. Navigation arrows are visible on either side of the carousel.

Hosted by Texas Digital Libraries, a consortium of Texas Higher-Education Institutions

HARVARD DATAVERSE: OPEN TO ALL THE RESEARCH COMMUNITY

The screenshot displays the Harvard Dataverse website. At the top left is the Harvard logo and the text "HARVARD Dataverse". To the right are navigation links: "About", "User Guide", "Support", "Sign Up", and "Log In". Below the header, there is a "Metrics" section showing "4,557,647 Downloads" and links for "Contact" and "Share". A main message reads: "Share, archive, and get credit for your data. Find and cite data across all research fields." Below this is a search bar with the placeholder "Search this dataverse...", a "Find" button, a link to "Advanced Search", and an "Add Data" button. The main content area shows search results for "1 to 10 of 82,721 Results". On the left, there are filters for "Dataverses (2,935)", "Datasets (79,786)", and "Files (414,659)". Under "Dataverse Category", there are links for "Research Project (986)", "Researcher (851)", "Organization or Institution (263)", "Research Group (130)", and "Journal (80)". Under "Metadata Source", there are links for "Harvested (53,466)" and "Harvard Dataverse (29,255)". Under "Publication Date", there are links for "2015 (15,450)" and "2011 (9,548)". The search results list includes:

- Replication Data for: ANTECEDENTS OF SOCIAL ENTREPRENEURIAL INTENTIONS IN MODERATING THE LINK BETWEEN NETWORKING COMPETENCE AND SOCIAL ENTREPRENEURSHIP**
Sep 25, 2018
Okwo, Henry, 2018, "Replication Data for: ANTECEDENTS OF SOCIAL ENTREPRENEURIAL INTENTIONS IN MODERATING THE LINK BETWEEN NETWORKING COMPETENCE AND SOCIAL ENTREPRENEURSHIP", <https://doi.org/10.7910/DVN/WKSNAF>, Harvard Dataverse, V1, UNF:6:ppVALkX62XppxJ7vpThilw==
- Survey on Hedging and Firm Survival in South Eastern Nigeria**
Sep 25, 2018
Okwo, Henry, 2018, "Survey on Hedging and Firm Survival in South Eastern Nigeria", <https://doi.org/10.7910/DVN/HVVRDO>, Harvard Dataverse, V1, UNF:6:~fIWVTG5ebNRGbe9SnGduA==
A survey conducted with the aid of the Researchers for Contemporary Issues in the Business Circle (RCIBC).
- The mechanism of plasma electron hole transverse instability**
Sep 25, 2018 - Plasma Science and Fusion Center Dataverse

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

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



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
```



SENSITIVE DATA: DATATAGS

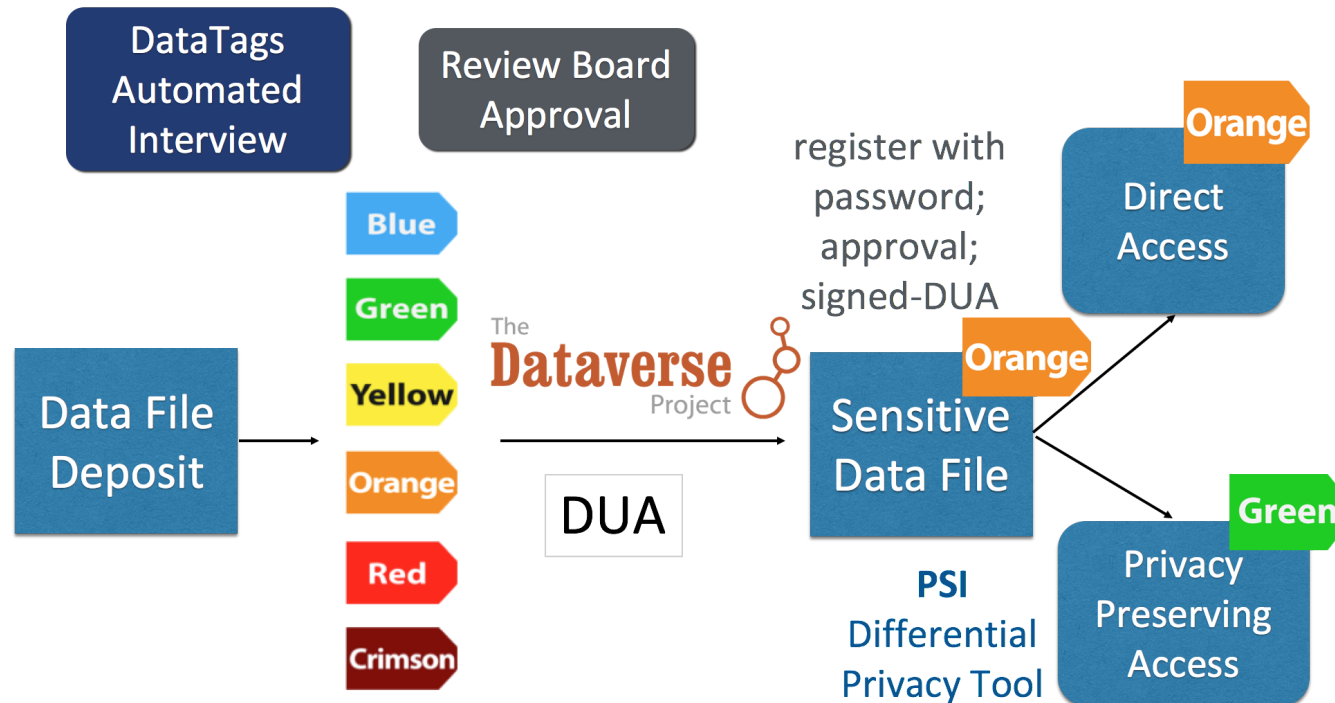
Standardize data security and access levels

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

CODE OCEAN BETA ABOUT EXPLORE PLANS HELP BLOG LOG IN

Discover & Run Scientific Code

Code Ocean is a cloud-based **computational reproducibility** platform

+ UPLOAD YOUR CODE

Search keyword, research field, title, author, DOI, etc.

SOCIAL SCIENCES Jun 2018

ENGINEERING Jun 2018

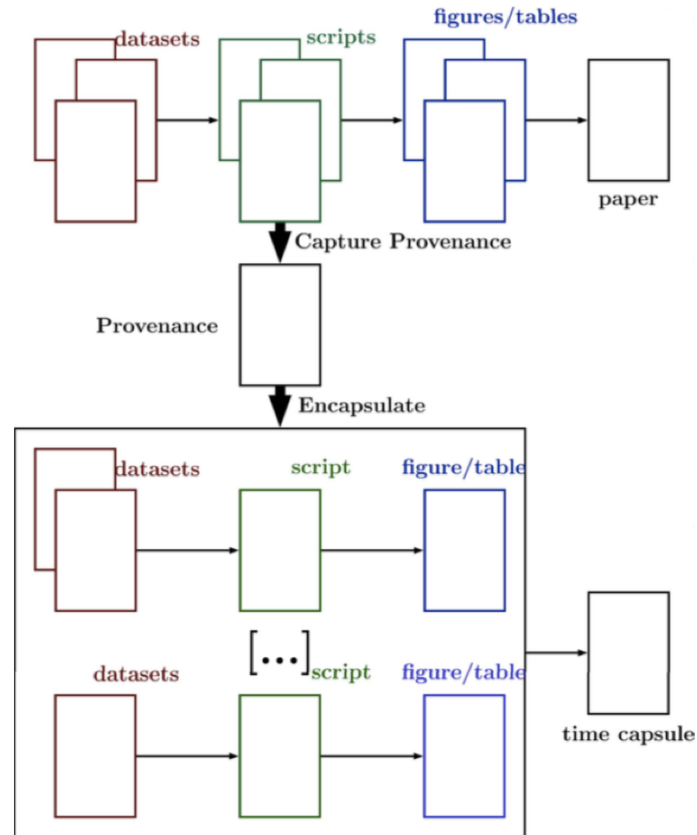
MEDICAL SCIENCES May 2018

Funded by Sloan Foundation, in collaboration with CodeOcean



Alfred P. Sloan
FOUNDATION

INTEGRATION WITH REPRODUCIBILITY TOOLS: ENCAPSULATOR

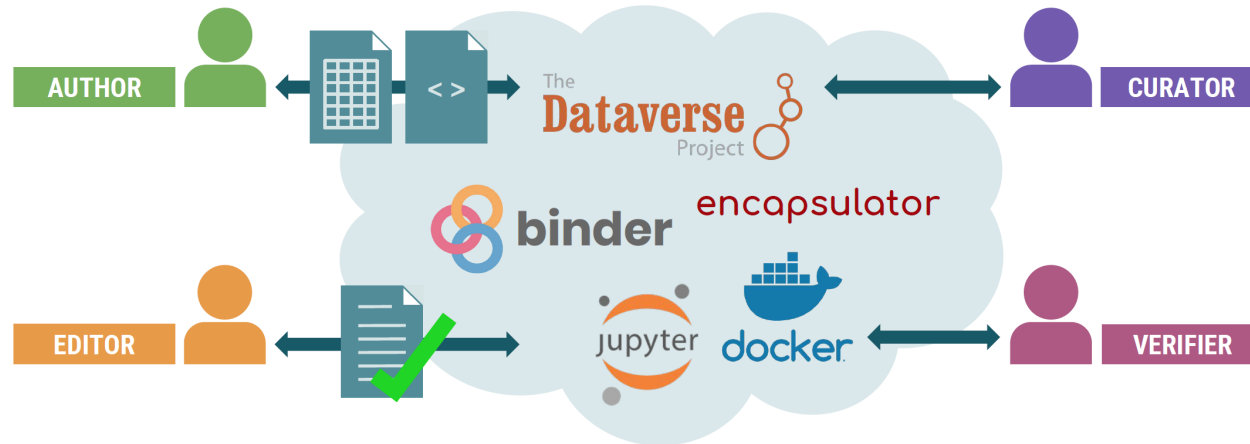


Funded by Sloan Foundation, in collaboration with Margo Seltzer



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