Principles of Data Sharing
(and what this means for generalist repositories)

Stefano M. Iacus, Senior Research Scientist
Director of Data Science and Product Research @ IQSS, Harvard University
Affiliate Faculty of the Kempner Institute for the Study of Natural and Artificial Intelligence

"Research Data Roadmap - Guiding Wisdom in the Management Mode" workshop @ National Yang Ming Chao Tung University (NYCU), 1st February 2024, Hsinchu City, Taiwan

HARVARD Dataverse | IQSS The Institute for Quantitative Social Science | HARVARD LIBRARY | HARVARD UNIVERSITY Information Technology | The Dataverse Project

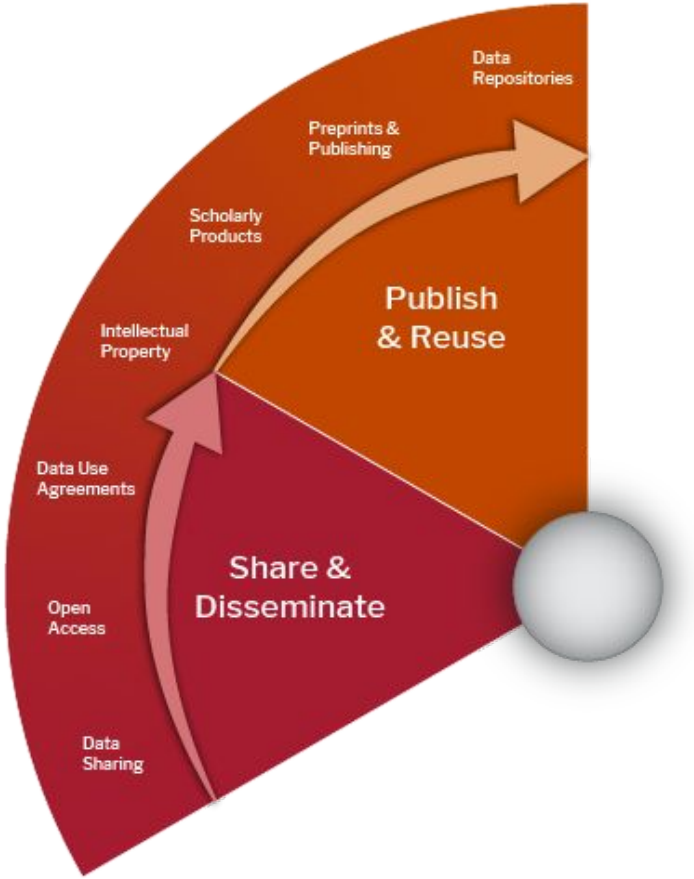
What is the RDM?

"The ***active and ongoing*** management of data ***through its lifecycle*** of interest and usefulness to scholarship, science, and education."

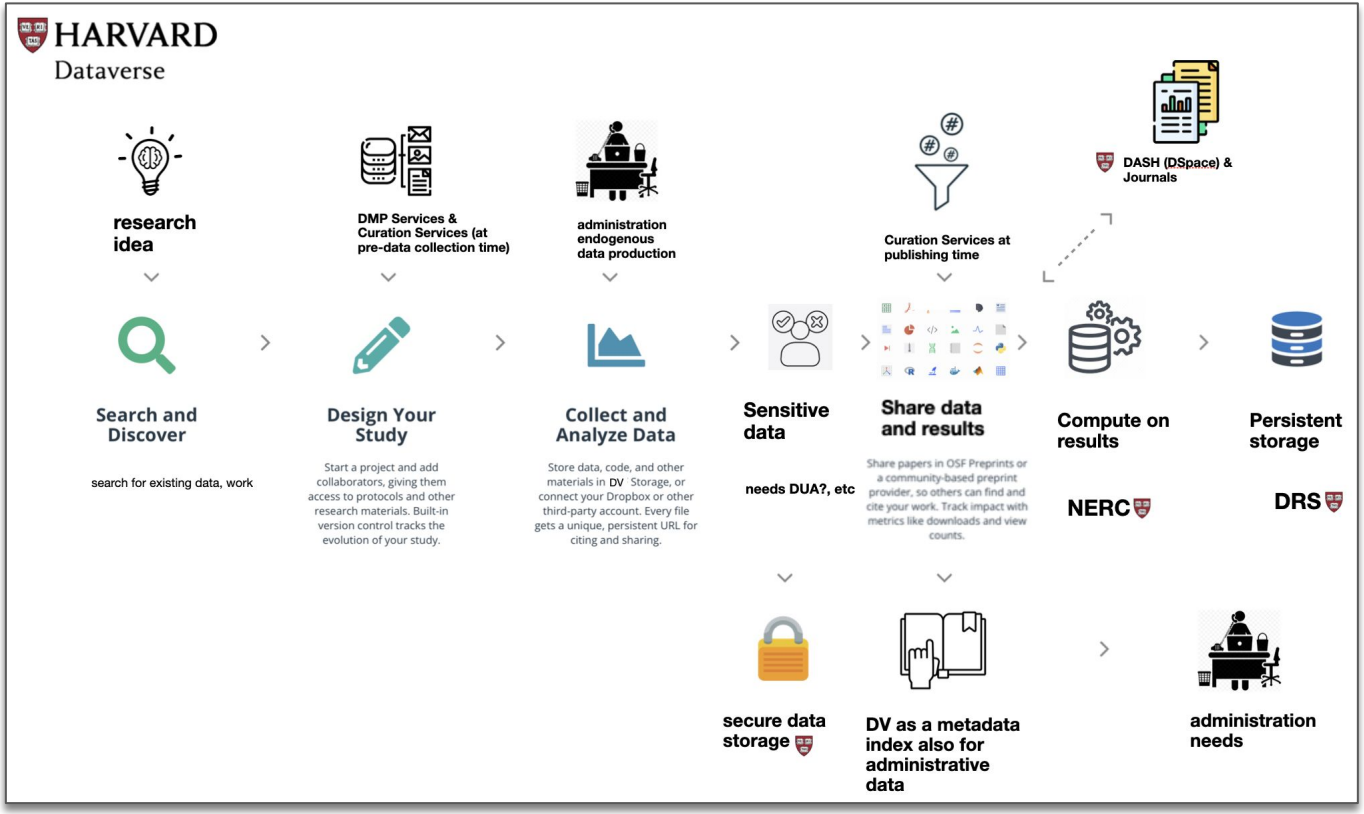
The University of Illinois' Graduate School of Library and Information Science

Data Sharing

- Depositing data in a repository
- Choosing data licenses
- Applying metadata to make published data more findable
- Write good documentation so shared data is actually reusable
- Steward shared data over its useful term

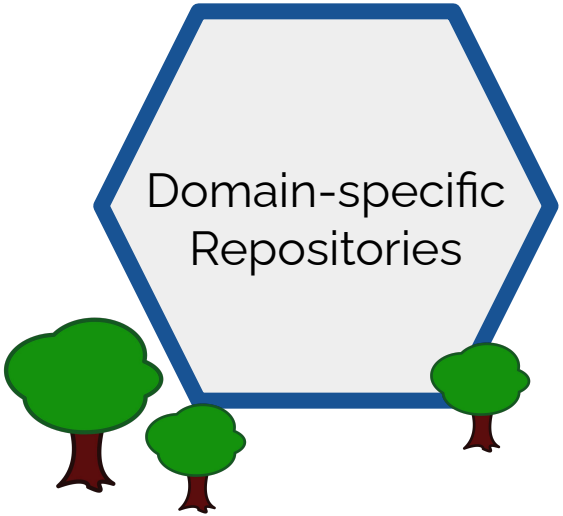


Harvard Dataverse through the data life cycle (future and current applications)



Research Data Repository Ecosystem

Different trees in the same forest



e.g. GenBank

e.g. Dataverse

e.g. WorldBank

Desirable Characteristics of Research Data Repositories

- Unique Persistent Identifiers
- Long-Term Sustainability
- Metadata
- Curation and Quality Assurance
- Free and Easy Access
- Broad and Measured Reuse
- Clear User Guidance
- Security and Integrity
- Confidentiality
- Common Format
- Provenance
- Retention Policy

Guidance set forth by NIH

And by The National Science and Technology Council, cited in OSTP guidance

GREI Generalist Ecosystem Initiative

Re

NIH National Institute of Health

The Dataverse® Project

OSF

figshare

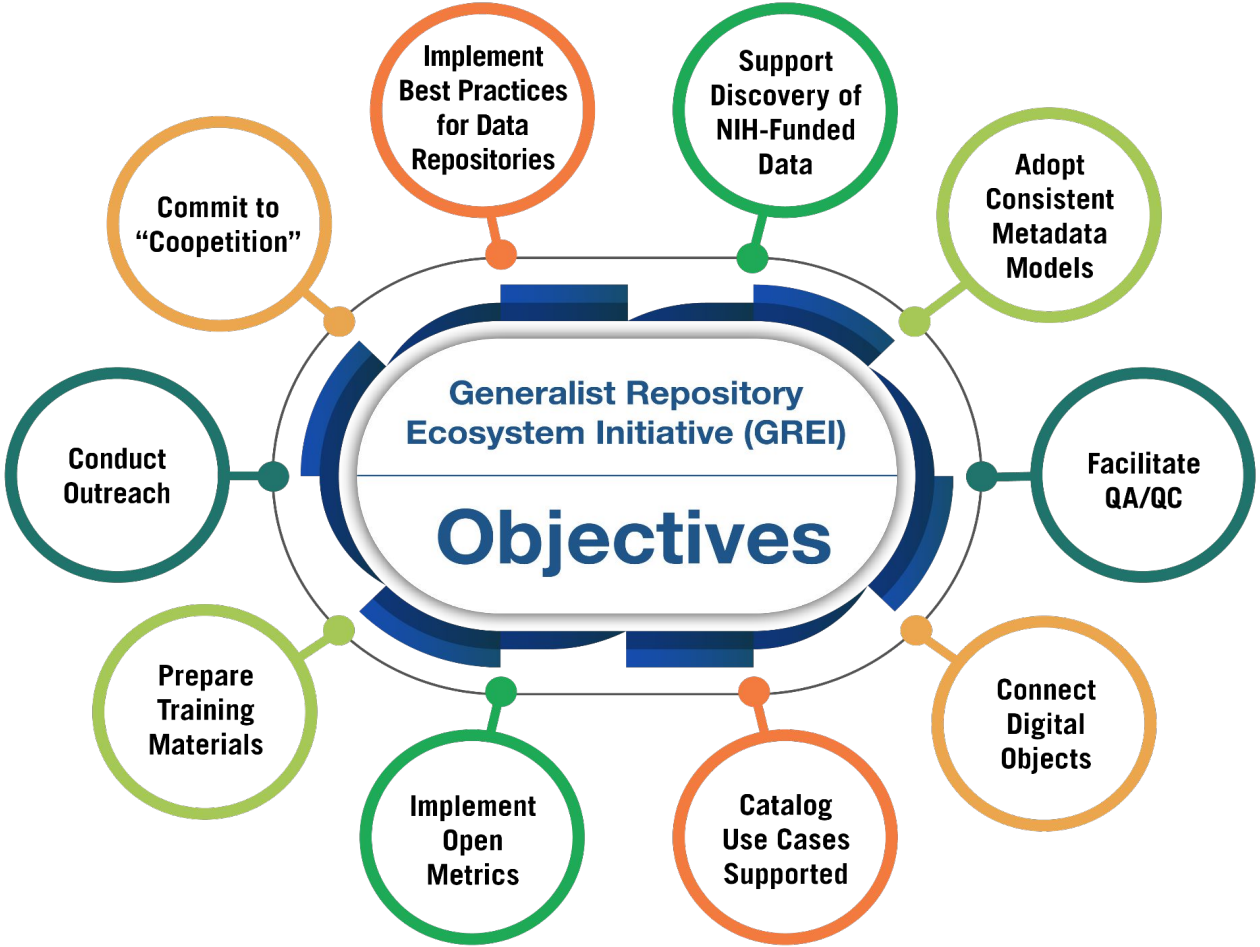
DRYAD

MENDELEY DATA

Vivli
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

zenodo

*“**Coopetition**” towards better research data sharing world !*



https://datascience.nih.gov/data-ecosystem/exploring-a-generalist-repository-for-nih-funded-data

Generalist Repository Comparison Chart

This chart is designed to assist researchers in finding a generalist repository should no domain repository be available to preserve their research data. Generalist repositories accept data regardless of data type, format, content, or disciplinary focus. For this chart, we included a repository available to all researchers specific to clinical trials (Vivli) to bring awareness to those in this field.

doi: 10.5281/zenodo.3946720

https://fairsharing.org/collection/GeneralRepositoryComparison

TOPIC	HARVARD DATAVERSE	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Brief Description	Harvard Dataverse is a free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data.	Open-source, community-led data curation, publishing, and preservation platform for CC0 publicly available research data Dryad is an independent non-profit that works directly with: • researchers to publish datasets utlizing best practices for discovery and reuse • publishers to support the integration of data availability statements and data citations into their workflows • institutions to enable scalable campus support for research data managment best practices at low cost	A free, open access, data repository where users can make all outputs of their research available in a discoverable, reusable, and citable manner. Users can upload files of any type and are able to share diverse research products including datasets, code, multimedia files, workflows, posters, presentations, and more. With discoverable metadata supporting FAIR principles, file visualizations, and integrations, researchers can make their work more impactful and move research further faster.	Mendeley Data is a free repository specialized for research data. Search more than 20+ million datasets indexed from 1000s of data repositories and collect and share datasets with the research community following the FAIR data principles.	OSF is a free and open source project management tool that supports researchers throughout their entire project lifecycle in open science best practices.	Vivli is an independent, non-profit organization that has developed a global data-sharing and analytics platform. Our focus is on sharing individual participant-level data from completed clinical trials to serve the international research community.	Powering Open Science, built on Open Source. Built by reserachers for researchers. Run from the CERN data centre, whose purpose is long term preservation for the High Energy Physics discipline, one of the largest scientific datasets in the world
Size limits	No byte size limit per dataset. Harvard Dataverse currently sets a file size limit of 2.5GB.	300GB/dataset	Soft limit of 20GB/file for free accounts. System limit of 5000GB/file. Unlimited storage of public data but 20GB storage for private data for free accounts. Email info@figshare.com to have upload and storage limits raised.	10GB per dataset	Projects currently have not storage limit. There is a 5GB/file upload limit for native OSF Storage. There is no limit imposed by OSF for the amount of storage used across add-ons connected to a given project.	If more than 10GB per study data, reach out to us	50GB per dataset, contact us via https://zenodo.org/support for higher limits
Storage space per researcher	1 TB per researcher	No limit	No limit	No limit	No limit	No limit	No limit
Persistent, Unique Identifier Support	DOI, Handle	DOI	DOI	DOI	DOI	DOI	DOI

https://doi.org/10.5281/zenodo.3946719 (Updated version 2!)

Common features and unique features

Common:

Core Metadata
Persistent Identifiers
Discoverable
Flexibility
Open access, FAIR
Metrics

Unique:

Output types
Storage, size limits
Licenses
Review
Controlled Access
Visualization
Costs

FINDABLE

Increases visibility, citations, and impact of research

Supports knowledge discovery and innovation

ACCESSIBLE

Streamlines and maximizes ability to build upon previous research results

Attracts partnerships with researchers and business in allied disciplines

FAIR

INTEROPERABLE

Supports and promotes inter- and cross-disciplinary data and reuse

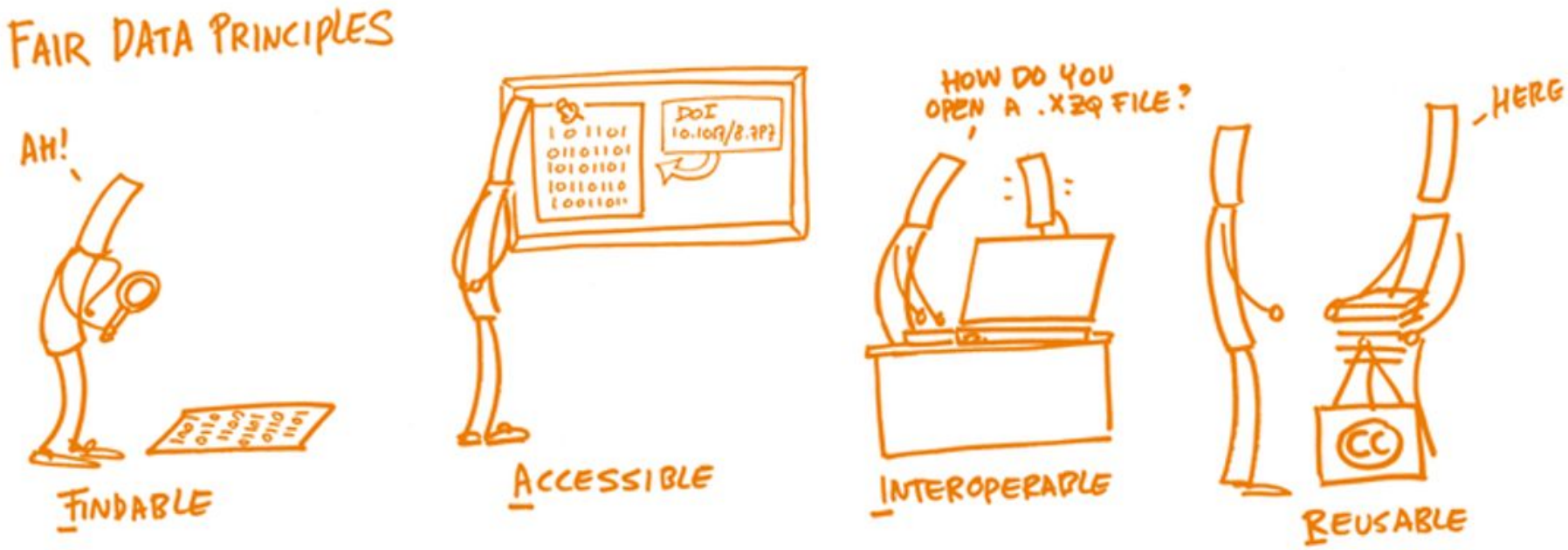
REUSABLE

Promotes use and reuse of data allowing resources to be allocated wisely

Improves reproducibility and reliability of research results

The FAIR Guiding Principles (Wilkinson et al. 2016)

in practice…



Who makes data FAIR?

Who makes data FAIR?

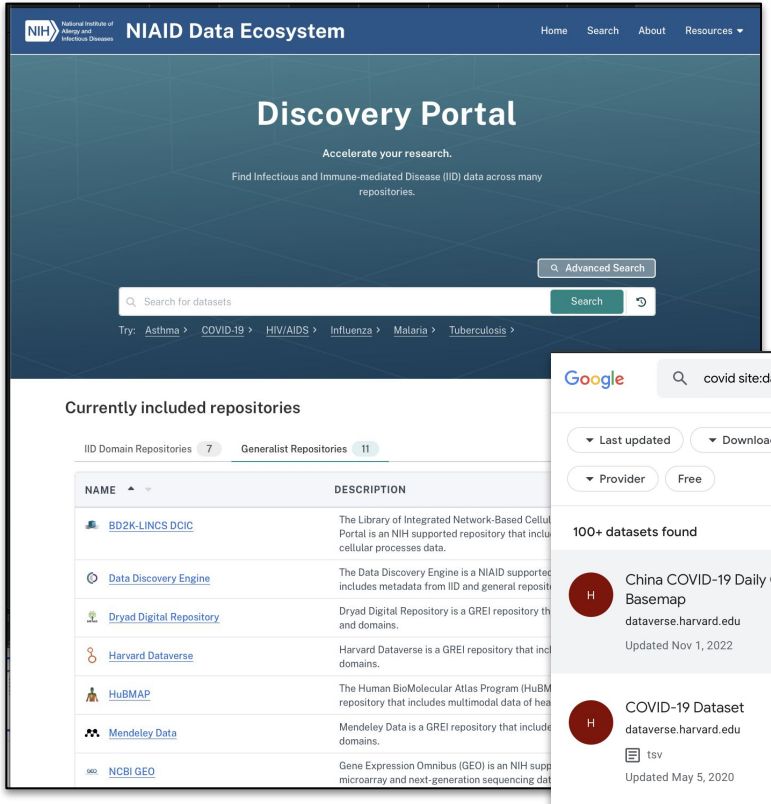
Repositories !

- Assign **persistent identifiers**
- Structure **metadata** records according to a **disciplinary standard** or **schema**
- **Index data** as searchable resources
- Retrieve datasets according to an **open protocol** that supports authentication
- **Preserves** data files and metadata
- **Provenance** and **versions** are tracked

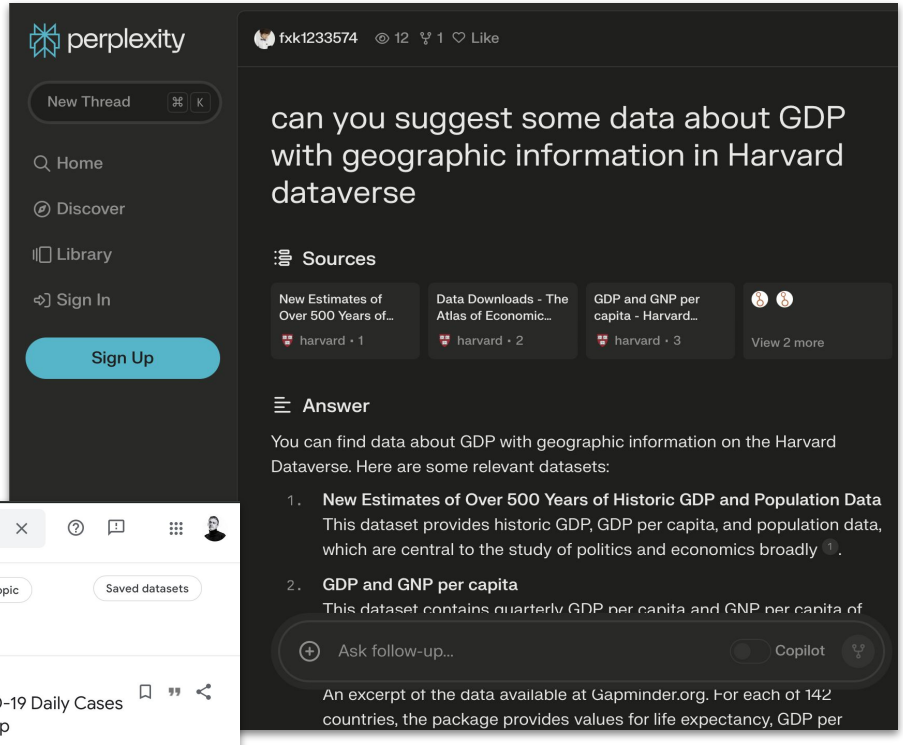
and Researchers too:

- Structure data clearly and apply good data management practices
- Document data and software
- Richly describe data using standardized metadata fields
- Apply license and/or clear terms of use

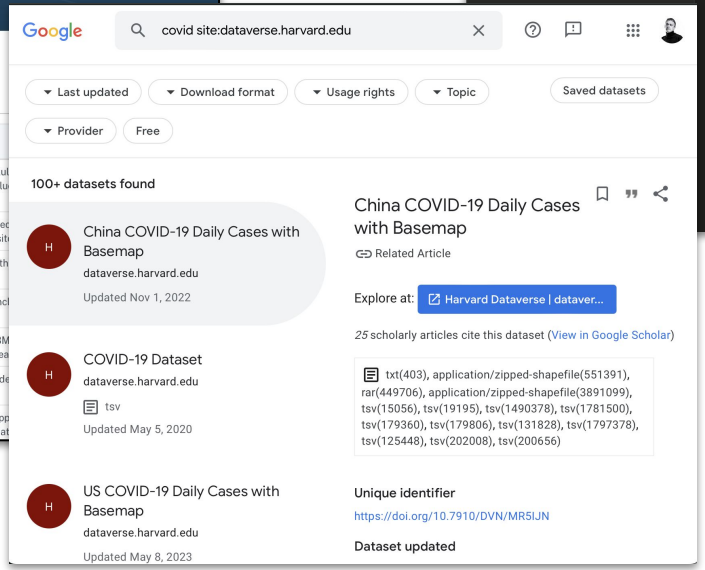
Discoverability & Interoperability



https://data.niaid.nih.gov/



https://www.perplexity.ai/



https://datasetsearch.research.google.com/

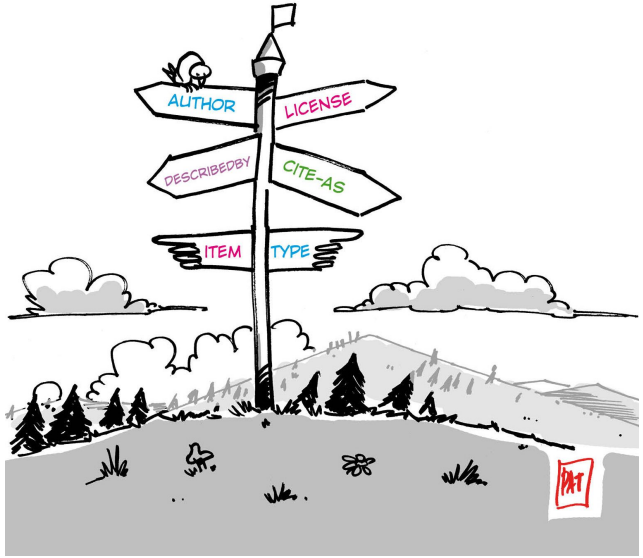
Signposting and Discoverability

Repositories **web pages** are **not optimized for** use by **machine agents** that navigate the scholarly web.

How can a robot determine which link on a landing page lead to content and which to metadata?

How can a bot distinguish those links from the myriad of other links on the page?

Signposting exposes these info to bots in in a standards-based way.



Release 5.14 added Signposting support to Dataverse to improve machine discoverability of datasets and files.

More discoverability features here: https://guides.dataverse.org/en/5.14/admin/discoverability.html

FAIR Signposting “Level 1”

Table 1: Link Relations used by FAIR Signposting	
Relation	**Usage**
cite-as	A one-to-one relationship between the entity and its globally unique identifier
describedby	A one-to-many relationship between the entity and all known metadata records about that entity
item	A one-to-many relationship between an entity representing a deposit and the data file(s) it contains.

HTTP
Link Headers

```
Link rel="cite-as"
https://upload.wikimedia.org/wikipedia/commons/9/91/Mona_Lisa_vectorized.svg

Link rel="described-by"
https://commons.wikimedia.org/wiki/File:Mona_Lisa_vectorized.svg#metadata
```

Starting Point:

- Web Search
- Bookmark
- DOI resolution
- Other ID resolution
- …

Sebastian Wallroth, CC0, via Wikimedia Commons

zenodo

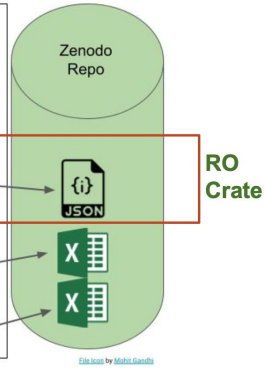
```
Link rel="cite-as"
http://doi.org/10.123/456.78

Link rel="described-by"
http://data.crosscite.org/10.123/456.78

Link rel="described-by"
https://zenodo.org/record/6438032/files/ro-crate-metadata.jsonld

Link rel="item"
https://zenodo.org/record/6438032/files/frequent_bigrams.csv

Link rel="item"
https://zenodo.org/record/6438032/files/frequent_terms.csv
```



More on this by

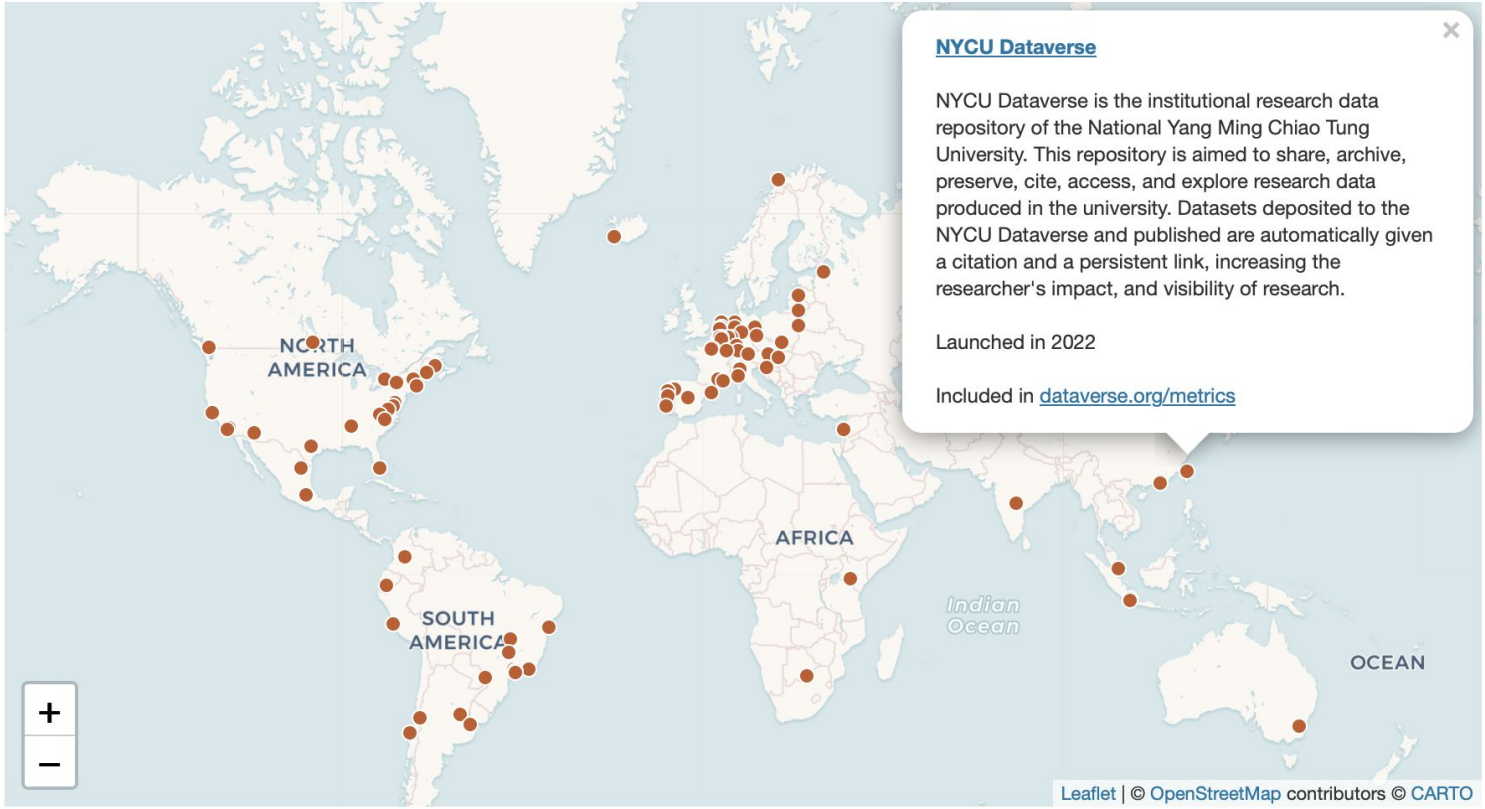
eosc
FAIR metrics and Data Quality
Task Force

https://tinyurl.com/FAIR-Signposting-GREI

What is and Why Dataverse ?

DATAVERSE REPOSITORIES - A WORLD VIEW

111 Installations



NYCU Dataverse

NYCU Dataverse is the institutional research data repository of the National Yang Ming Chiao Tung University. This repository is aimed to share, archive, preserve, cite, access, and explore research data produced in the university. Datasets deposited to the NYCU Dataverse and published are automatically given a citation and a persistent link, increasing the researcher's impact, and visibility of research.

Launched in 2022

Included in dataverse.org/metrics

Desirable Characteristics of Research Data Repositories

- Unique Persistent Identifiers
- Long-Term Sustainability
- Metadata
- Curation and Quality Assurance
- Free and Easy Access
- Broad and Measured Reuse
- Clear User Guidance
- Security and Integrity
- Confidentiality
- Common Format
- Provenance
- Retention Policy

Dataverse has all of them

plus more

What is and Why Dataverse ?

- An open-source platform to **publish, cite,** and **archive research data**
- Built to support **multiple types** of data, **users**, and **workflows**
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) since 2006
- Development funded by IQSS and with **grants**, in collaboration with institutions around the world
- Core team
  - @ IQSS - developers, designers, UX/UI, metadata specialists, curation team, leadership team
  - key contributors from the community

What are the features of a Dataverse Repository?

Create a “**dataverse/collection**”

Create a “**dataset**,” with extensive **metadata**/**files**

Upload data files/documentation, with metadata

Publish and share your dataverse and dataset

Link your datasets to coauthors/link other data to your page

Dataset and file level **DOIs**

Full dataset and file level **citations**

Set “**terms of access**”

Export metadata in several format

Export data citation in several formats

Private URL to share your dataset in draft format

Data Analysis, File Previewer

File folder hierarchy preservation

Restrict/Open files for **access**/request access

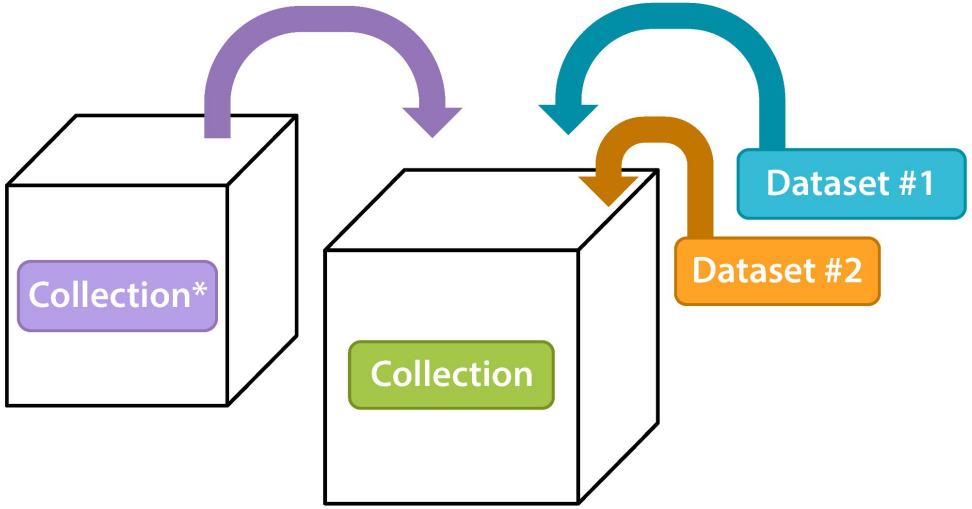
Workflows for data deposit and publishing

Custom metadata blocks

Dataverse Collections

- Ability to create Dataverse **collections** to organize datasets according to your needs
- Dataverses collections can also contain other collections, enabling any **hierarchical** structure
- **Different rules** can be applied for different Dataverse collections, e.g. for Metadata, Permissions, etc.

Schematic Diagram of a **Collection** in Dataverse Software 5.0



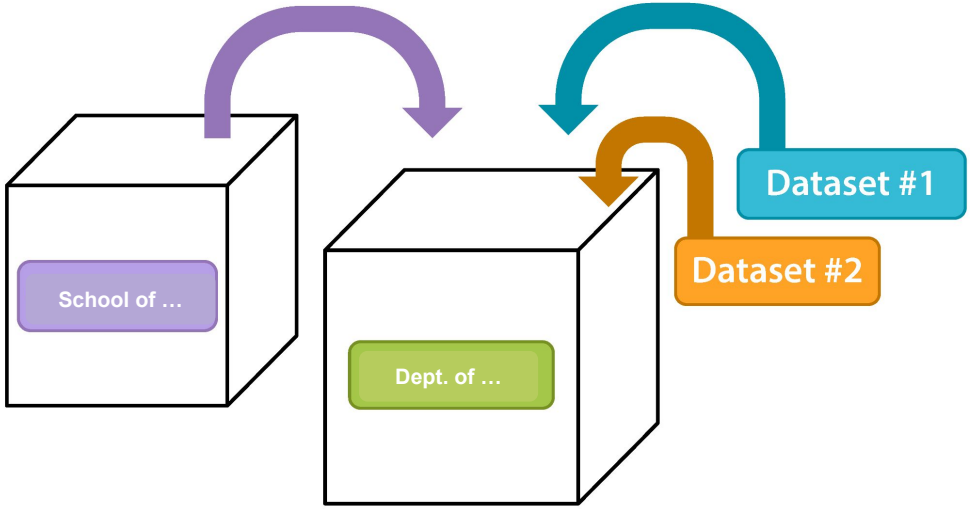
Container for your **Datasets** and/or **Collections***

* Collections can contain other Collections

Dataverse Collections

- Ability to create Dataverse **collections** to organize datasets according to your needs
- Dataverses collections can also contain other collections, enabling any **hierarchical** structure
- **Different rules** can be applied for different Dataverse collections, e.g. for Metadata, Permissions, etc.

Schematic Diagram of a **Collection** in Dataverse Software 5.0



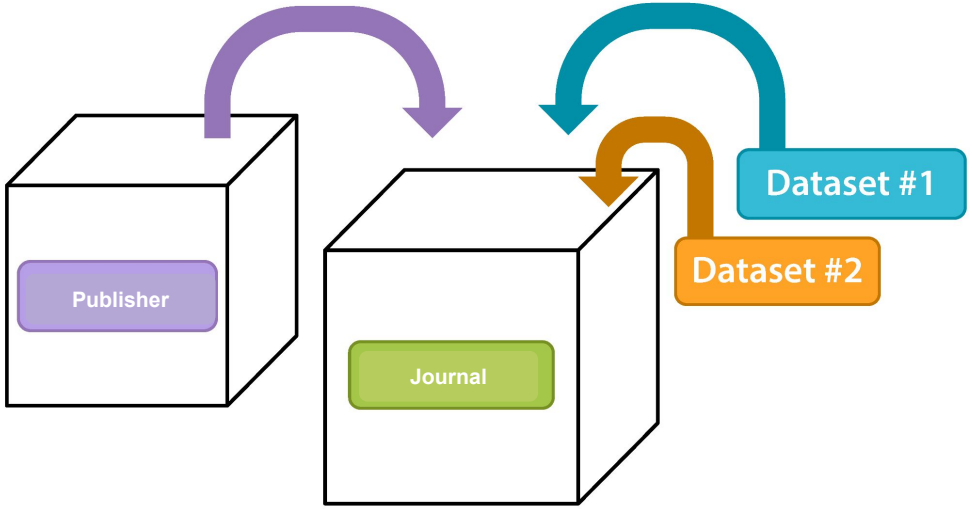
Container for your **Datasets** and/or **Collections***

* Collections can contain other Collections

Dataverse Collections

- Ability to create Dataverse **collections** to organize datasets according to your needs
- Dataverses collections can also contain other collections, enabling any **hierarchical** structure
- **Different rules** can be applied for different Dataverse collections, e.g. for Metadata, Permissions, etc.

Schematic Diagram of a **Collection** in Dataverse Software 5.0



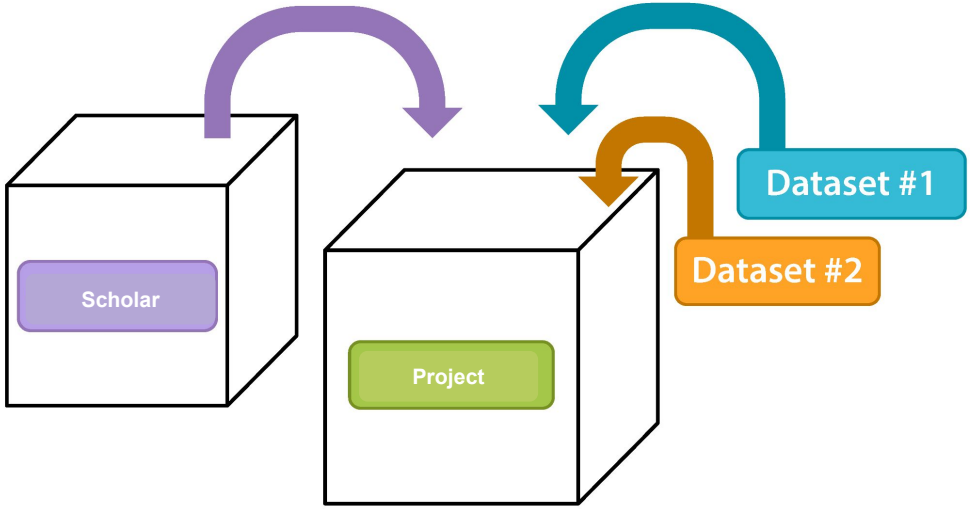
Container for your **Datasets** and/or **Collections***

* Collections can contain other Collections

Dataverse Collections

- Ability to create Dataverse **collections** to organize datasets according to your needs
- Dataverses collections can also contain other collections, enabling any **hierarchical** structure
- **Different rules** can be applied for different Dataverse collections, e.g. for Metadata, Permissions, etc.

Schematic Diagram of a **Collection** in Dataverse Software 5.0



Container for your **Datasets** and/or **Collections***

* Collections can contain other Collections

Dynamic Metadata

- **Metadata** is **defined dynamically** at the database level, allowing for modularly adding new Metadata blocks
- Supports:
  - single or multiple values
  - simple or compound values
  - controlled vocabularies
  - external vocabularies

Choose the metadata fields to use in dataset templates and when adding a dataset to this dataverse.

- [x] Citation Metadata (Required) [+] View fields + set as hidden, required, or optional
- [ ] Geospatial Metadata [+] View fields
- [ ] Social Science and Humanities Metadata [+] View fields
- [ ] Astronomy and Astrophysics Metadata [+] View fields
- [ ] Life Sciences Metadata [+] View fields
- [ ] Journal Metadata [+] View fields

Citation Metadata

Title *
Enter title...
Add "Replication Data for" to Title

Author *
Name *: Admin, Dataverse
Affiliation: Dataverse.org
Identifier Scheme: Select...
Identifier:

Contact *
Name: Admin, Dataverse
Affiliation: Dataverse.org
E-mail *: dataverseadmin@iq.harvard.edu

Description *
This field supports only certain HTML tags.
Text *
Date: YYYY-MM-DD

Metadata Standards

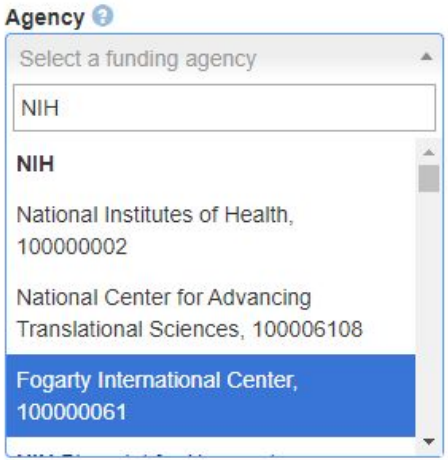
1. **Citation Metadata:** any metadata that would be needed for generating a data citation and other general metadata that could be applied to any dataset;

2. **Domain Specific Metadata:** with specific support currently for Social Science, Life Science, Geospatial, and Astronomy datasets;

3. **File-level Metadata:** varies depending on the type of data file and include options like file tags, descriptions, variable names, and hierarchy preservation.

Controlled vocabularies support

- Built-in support for smaller, static vocabularies, e.g. country, language lists
- Plugin mechanism for larger and/or dynamic vocabularies
  - Dataverse stores the persistent identifier for the term, users see the textual name with details, icon (e.g. , link to definition, etc.
  - Associates JavaScript edit/view widgets with any dataset metadata field(s)
  - JavaScript can query external services to support type-ahead look-up and provide details, internationalization, etc.
  - JavaScripts can present drop-down, hierarchical, map-based, or other appropriate input options
  - JavaScripts are not Dataverse-specific - they rely on HTML data-* attributes to find the correct fields and can potentially be reused in other repositories



Input ⇧
View (with popup) ⇩

Citation Metadata

Persistent Identifier	doi:10.7910/DVN/25833
Publication Date	2014-05-22
Title	Brain Genomics Superstruct Project (GSP)
Author	Buckner, Randy L. (Harvard University) Roffman, Joshua L. Smoller, Jordan W.
Point of Contact	Use email button above to contact. GSP Data Release (Harvard University)
Description	Large scale imaging data sets are necessary to address complex questions regarding the relationship between brain and behavior. The Brain Genomics Superstruct Project Open Access Data Release exposes a carefully vetted collection of neuroimaging, behavior, cognitive, and personality data for over 1,500 human participants. Each neuroimaging data set includes one high-resolution Magnetic Resonance Imaging (MRI) acquisition and one or more resting-state functional MRI acquisitions. Each functional acquisition is accompanied by a fully-automated quality assessment and pre-computed brain morphometrics are also provided.
Subject	Other
Producer	Neuroinformatics Research Group (Harvard University) (NRG) http://neuroinformatics.harvard.edu/ Logo URL
Production Date	2014
Distributor	Harvard Dataverse Network (Harvard University) http://thedata.harvard.edu/dvn/ Logo URL
Distribution Date	2014-05
Deposit Date	2014-05
Related Material	Examples of prior publications of GSP data with partial data description: Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R, Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L. (2011) The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology, 106(3): 1125-1165: Link to article Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Ye o, B.T. (2011) The organization of the human cerebellum estimated by intrinsic functional connectivity. Journal of Neurophysiology, 106(5): 2322-2345: Link to article Choi, E.Y., Yeo, B.T.T., Buckner, R.L. (2012) The organization of the human striatum estimated by intrinsic functional connectivity. Journal of Neurophysiology, 108(8): 2242-2263: Link text Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L. (2012) The influence of head motion on intrinsic connectivity MRI. NeuroImage , 59(1): 431-438: Link to article

Note: **extensive metadata (depositor provided)** and related materials (**depositor provided**) to improve understanding of the dataset, **bidirectional linking** to related articles, as well as **file level metadata** with files in **interoperable tabular formats** that allow **visualization** and **online data analysis** and the **download of files in multiple formats (**open format file by depositor)

Files | Metadata | Terms | Versions

Search this dataset...

Filter by
File Type: All Access: All Sort

1 to 10 of 12 Files Download

EthnonationalistGenderNorms_Analysis_FINAL.do
Stata Syntax - 13.8 KB
Published Feb 15, 2023
0 Downloads
MD5: c46...372
STATA do-file that generates all the tables and figures in the main manuscript and the appendix othe
Figures 2 and 3 in the main manuscript.

EthnonationalistGenderNorms_Codebook.pdf
Adobe PDF - 107.0 KB
Published Feb 15, 2023
0 Downloads
MD5: 3ef...0cc
Codebook

EthnonationalistGenderNorms_DataCleaning_FINAL.do
Stata Syntax - 26.7 KB
Published Feb 15, 2023
0 Downloads
MD5: ce7...2cd
STATA .do file that transforms the two original data source files into four analysis data sets.

EthnonationalistGenderNorms_Electoral.tab
Tabular Data - 8.9 MB
Published Feb 15, 2023
0 Downloads
258 Variables, 16682 Observations UNF:6:mNpn...UsA==
The original data source file containing electoral data from Bihar state assembly elections (Ananay et. al. 2021) and data on candidate caste backgrounds. Both datasets are from the Trivedi Centre for Political Data at Ashoka University, New Delhi, India. See Codebook for further details.

EthnonationalistGenderNorms_Electoral_Cleaned.tab
Tabular Data - 1.1 MB
Published Feb 15, 2023
0 Downloads
16 Variables, 16196 Observations UNF:6:MThl...USQ==
This is the cleaned dataset based on the electoral data and data on candidate caste backgrour

EthnonationalistGenderNorms_FactorVariables_ConjointPlots.R
R Syntax - 1.7 KB
Published Feb 15, 2023
0 Downloads
MD5: 660...cb0
R source file that produces Figures 2 and 3 in the main manuscript

EthnonationalistGenderNorms_ForConjoint_lowercaste.tab
Tabular Data - 748.0 KB
Published Feb 15, 2023
0 Downloads
6 Variables, 14584 Observations UNF:6:fUmY...xAw==
This is the dataset based on our original survey dataset with factor variables used to generate our conjoint plots Figures 2 and 3.

descriptive file names, proper file extension (depositor)

visualization (repository)

file level metadata (depositor and repository)

download of files in multiple formats (repository, depositor)

File Access
Public
Download Options
Stata 14 Binary (Original File Format)
Tab-Delimited
RData
Download Metadata
Variable Metadata
Data File Citation
Explore Options
Data Explorer

online data analysis (repository)

Flexible Permission System

- Supports multiple workflows by controlling who can add to your Dataverse collection, what they can, and what role they have on and created Datasets
- Roles are defined as a set of permissions to grant to users or to groups
- Groups can be defined statically or dynamically (e.g. users logging in from the same institution, via Shibboleth)

Edit Access

Who can add to this dataverse?

- ○ Anyone adding to this dataverse needs to be given access
- ○ Anyone with a Dataverse account can add sub dataverses
- ○ Anyone with a Dataverse account can add datasets
- ● Anyone with a Dataverse account can add sub dataverses and datasets

When a user adds a new dataset to this dataverse, which role should be automatically assigned to them on that dataset?

- ● Contributor - Edit metadata, upload files, and edit files, edit Terms, Guestbook, Submit datasets for review
- ○ Curator - Edit metadata, upload files, and edit files, edit Terms, Guestbook, File Restrictions (Files Access + Use), Edit Permissions/Assign Roles + Publish

Save Changes Cancel

2 Users/Groups

User/Group Name (Affiliation)	ID	Role
Dataverse Admin (Dataverse.org)	@dataverseAdmin	Admin
Anyone with a Dataverse account	:authenticated-users	Dataverse + Dataset Creator

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Sign Up | Log In

Brain Genomics Superstruct Project

Brain Genomics Superstruct Project (GSP) Dataverse Home

(Harvard University)

Harvard Dataverse > Brain Genomics Superstruct Project (GSP) Dataverse >

Brain Genomics Superstruct Project (GSP)

Version 10.5

Buckner, Randy L.; Roffman, Joshua L.; Smoller, Jordan W., 2014, "Brain Genomics Superstruct Project (GSP)", https://doi.org/10.7910/DVN/25833, Harvard Dataverse, V10

Cite Dataset Learn about Data Citation Standards.

Contact Owner

Dataset Metrics

13,287 Downloads

Description

Large scale imaging data sets are necessary to address complex questions regarding the relationship between brain and behavior. The Brain Genomics Superstruct Project Open Access Data Release exposes a carefully vetted collection of neuroimaging, behavior, cognitive, and personality data for over 1,500 human participants. Each neuroimaging data set includes one high-resolution Magnetic Resonance Imaging (MRI) acquisition and one or more resting-state functional MRI acquisitions. Each functional acquisition is accompanied by a fully-automated quality assessment and pre-computed brain morphometrics are also provided.

Citation Metadata

Persistent Identifier	doi:10.7910/DVN/25833
Publication Date	2014-05-22
Title	Brain Genomics Superstruct Project (GSP)
Author	Buckner, Randy L. (Harvard University) Roffman, Joshua L. Smoller, Jordan W.
Point of Contact	Use email button above to contact. GSP Data Release (Harvard University)
Description	Large scale imaging data sets are necessary to address complex questions regarding the relationship between brain and behavior. The Brain Genomics Superstruct Project Open Access Data Release exposes a carefully vetted collection of neuroimaging, behavior, cognitive, and personality data for over 1,500 human participants. Each neuroimaging data set includes one high-resolution Magnetic Resonance Imaging (MRI) acquisition and one or more resting-state functional MRI acquisitions. Each functional acquisition is accompanied by a fully-automated quality assessment and pre-computed brain morphometrics are also provided.

Files | Metadata | Terms | Versions

Search this dataset...

Filter by
File Type: All | Access: All | File Tag: All

1 to 10 of 15 Files

- GSP_DataUse_Terms_140422.pdf — Adobe PDF - 282.7 KB; Published Aug 24, 2014; 1,199 Downloads; MD5: e64...059; Data use terms; Documentation
- GSP_list_140630.csv — Plain Text - 636.4 KB; Published Aug 24, 2014; 773 Downloads; MD5: 27e...4a8; Demographic, cognitive/behavior, quality control, and morphometrics data.; Documentation
- GSP_part10_140630.tar — TAR Archive - 10.0 GB; Published May 21, 2014; 1,114 Downloads; MD5: 302...48e; Tar archive of imaging data for subjects 1414-1570; refer to GSP_README_140630.pdf for more information.
- GSP_part1_140630.tar — TAR Archive - 10.1 GB; Published May 21, 2014; 1,151 Downloads; MD5: a33...f55; Tar archive of imaging data for subjects 1-157; refer to GSP_README_140630.pdf for more info
- GSP_part2_140630.tar — TAR Archive - 10.2 GB; Published May 21, 2014; 992 Downloads; MD5: f99...028; Tar archive of imaging data for subjects 158-314; refer to GSP_README_140630.pdf for more information.
- GSP_part3_140630.tar — TAR Archive - 10.1 GB; Published May 21, 2014; 1,089 Downloads; MD5: 57e...45e; Tar archive of imaging data for subjects 315-471; refer to GSP_README_140630.pdf for more information.
- GSP_part4_140630.tar — TAR Archive - 9.7 GB; Published May 21, 2014; 668 Downloads

Dataset Terms

License/Data Use Agreement

Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation shown on the dataset page.

Custom Dataset Terms — the following Custom Dataset Terms have been defined for this dataset.

Terms of Use

These restrictive terms of use take precedence over any less restrictive use terms that apply generally to Dataverse Network Terms of Use

I request access to data collected as part of the Brain Genomics Superstruct Project (GSP) of Harvard University and the Massachusetts General Hospital, and I agree to the following:
1. I will not attempt to establish the identity of or attempt to contact any of the included human subjects.
2. I will not attempt to link any of the distributed data to any other data that might contain information about the included human subjects.
3. I understand that under no circumstances will the code that would link these data to Protected Health Information be given to me, nor will any additional information about individual human subjects be released to me under these Open Access Data Use Terms.
4. I will comply with all relevant rules and regulations imposed by my institution. This may mean that I need my research to be approved or declared exempt by a committee that oversees research on human subjects e.g., my Inter nal Review Board or Ethics Committee. Different committees operate under different national, state, and local laws and may interpret regulations differently, so it is important to ask about this.
5. I may redistribute original GSP Open Access data and any derived data as long as the data are redistributed under these

Guestbook

Guestbook

The following guestbook will prompt a user to provide additional information when downloading a file.

Brain Genomics Superstruct Project (GSP) Preview Guestbook

Files | Metadata | Terms | Versions

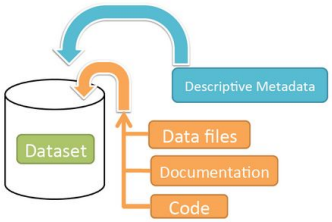
View Differences

Dataset Version	Summary	Contributors	Published on
10.5	**Citation Metadata:** Description (1 Changed); View Details	Timothy O'Keefe	2020-03-09
10.4	View Details	Sonia Barbosa	2017-02-28
10.3	**Citation Metadata:** Description (1 Changed); View Details	Timothy O'Keefe	2015-07-15
10.2	**Terms of Use/Access Changed** View Details	Timothy O'Keefe	2015-05-22
10.1	**Terms of Use/Access Changed** View Details	Timothy O'Keefe	2015-05-22
10.0	**Additional Citation Metadata:** (1 Added); **Files (Added: 1; Removed: 1);** View Details	Timothy O'Keefe	2015-05-15
9.0	**Citation Metadata:** Description (1 Changed); **Terms of Use/Access Changed** View Details	Timothy O'Keefe	2014-10-06

The Dataverse® Project

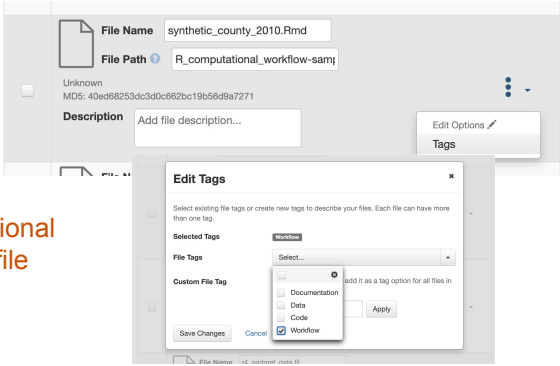
File Types, Format, Documentation

Schematic Diagram of a **Dataset** in Dataverse 4.0

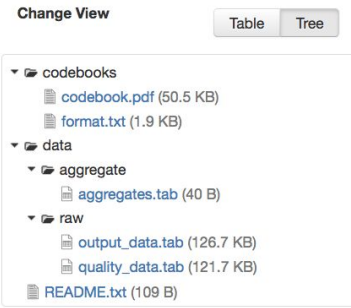


Container for your data, documentation, and code.

Computational workflow file support



Folder hierarchy support



Tabular data support

Supported File Formats

Tabular Data ingest supports the following file formats:

File format	Versions supported
SPSS (POR and SAV formats)	7 to 22
STATA	4 to 15
R	up to 3
Excel	XLSX only (XLS is NOT supported)
CSV (comma-separated values)	(limited support)

File level access control

Access

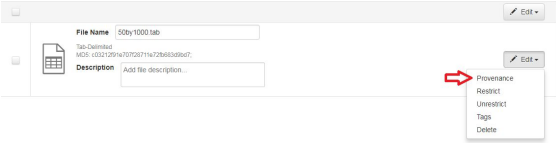
Public (1,784,339)

Restricted (50,591)

Embargoed then Public (154)

Embargoed then Restricted (14)

File level Provenance support



The Dataverse® Project

Persistent Identifiers (PIDs)

DataCite

A persistent identifier (PID) is a unique, long-lasting reference to an entity.

https://doi.org/10.5061/dryad.708gr

https://datadryad.org/stash/dataset/doi:10.5061/dryad.708gr

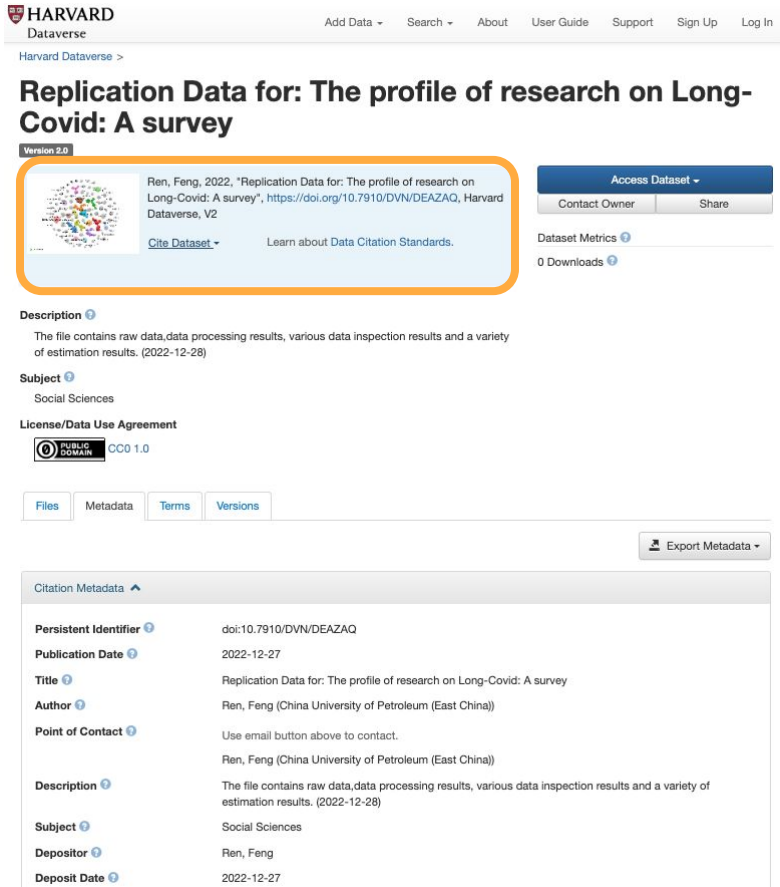
Special URL that is registered in a known system, like DOI, ORCID or ROR

Always points to the same resource (or a metadata representation)

Example: Dataset DOI - Harvard Dataverse

https://doi.org/10.7910/DVN/DEAZAQ

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DEAZAQ



Persistent Identifiers (PIDs)

- PIDs are assigned for every dataset
- PIDs can also be assigned per file, configurable per installation
  - Can also be configured for only specific Dataverse collections within the installation
  - Can be defined to be “dependent” on the dataset PID or “independent”

File Citation

Admin, Dataverse, 2023, "argentina.jpeg", *GPD Previewers*, https://doi.org/10.70122/FK2/BLXVAF/WKNVG3, Demo Dataverse, V3

Cite Data File

Learn about Data Citation Standards.

Dataset Citation

Admin, Dataverse, 2023, "GPD Previewers", https://doi.org/10.70122/FK2/BLXVAF, Demo Dataverse, V3

Cite Dataset

Learn about Data Citation Standards.

Persistent Identifiers (PIDs)

GREI chose the DataCite metadata schema because:

- All GREI repositories already use it to register DOIs
- It's domain agnostic
- DataCite already collaborates closely with GREI
- Other services rely on metadata expressed in DataCite's schema, including metadata aggregators and DataCite's own Event Data service

- **The GREI Metadata Recommendations highlight** specific properties from the DataCite Metadata Schema (v4.4), beyond the minimum required fields.
- **Repositories are encouraged to** incorporate these properties in their metadata or identify a local equivalent field.
    - For example, an "Author Identifier" field may be mapped to the DataCite "nameIdentifier" sub-property of "Creator".
- **When registering a DOI with DataCite, recommended properties should** be included in the DataCite DOI metadata.

Persistent Identifiers (PIDs) provide value and insight across research stakeholders

There are PIDs for people, places, and things

PIDs for people (researchers) include ISNIs and ORCID iDs

https://orcid.org/0000-0002-5989-8244

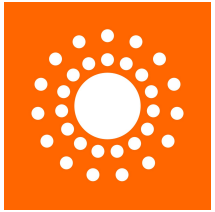
PIDs for places (research organizations) include ROR and Funder IDs

https://ror.org/05d5mza29

PIDs for things (research outputs/inputs like grants, papers, projects, etc.) include Crossref and DataCite DOIs, IGSNs, and more

https://doi.org/10.17605/osf.io/jzu37

Dataverse supports multiple metadata export schemas

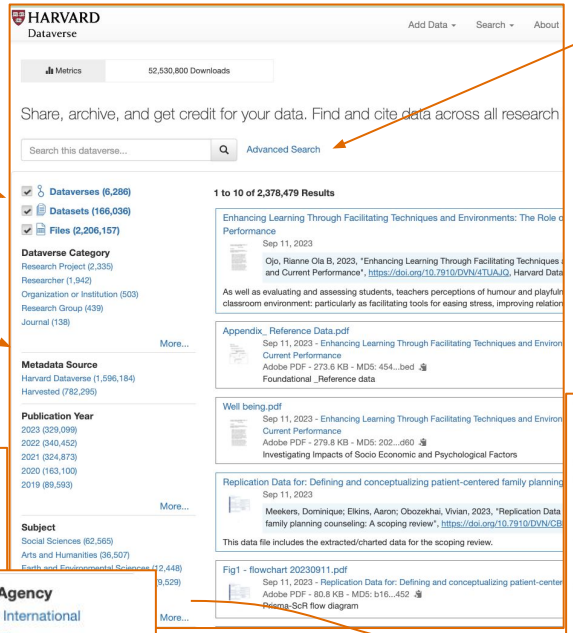


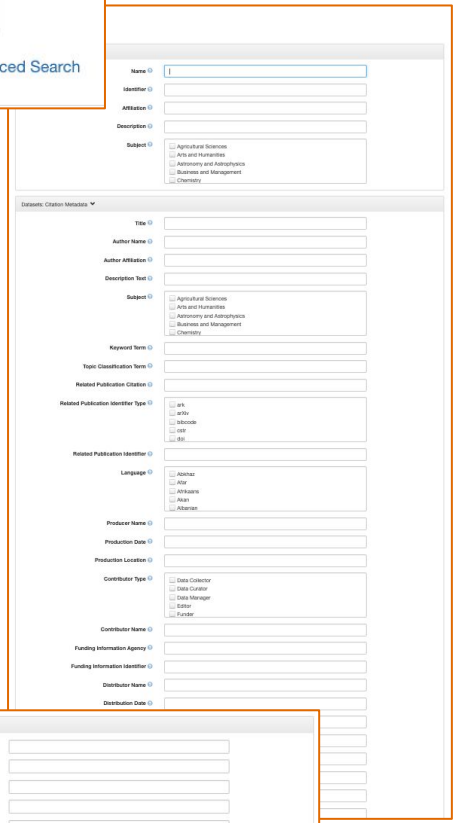
The Dataverse® Project

Search & browse

Search:

- Keyword search
- Advanced search
- Funding agency
- Faceted search
- Sorting
- Cross repository integration





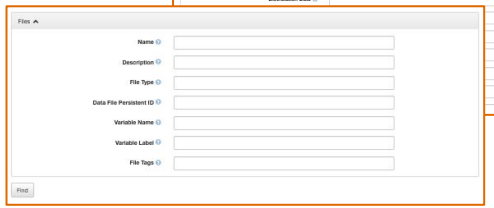
Keyword Term
AFRICA (328)
AFRICA SOUTH OF SAHARA (309)
EAST AFRICA (162)
ASIA (148)
health (145)

Funding Information Agency
United States Agency for International Development (USAID) (247)
Bill and Melinda Gates Foundation (BMGF) (77)
Bill and Melinda Gates Foundation (31)
World Bank (24)
Bill & Melinda Gates Foundation (BMGF) (23)
More...

File Type
Image (495,798)
Data (397,126)
Text (357,854)
Unknown (302,072)
Document (184,696)
More...

File Tag
Data (135,039)
Country:United States (51,786)
Physical Format:Pamphlets (27,695)
Photo (26,226)
Documentation (22,388)
More...

Access
Public (2,153,422)
Restricted (52,527)
Embargoed then Public (188)
Embargoed then Restricted (20)



The Dataverse® Project

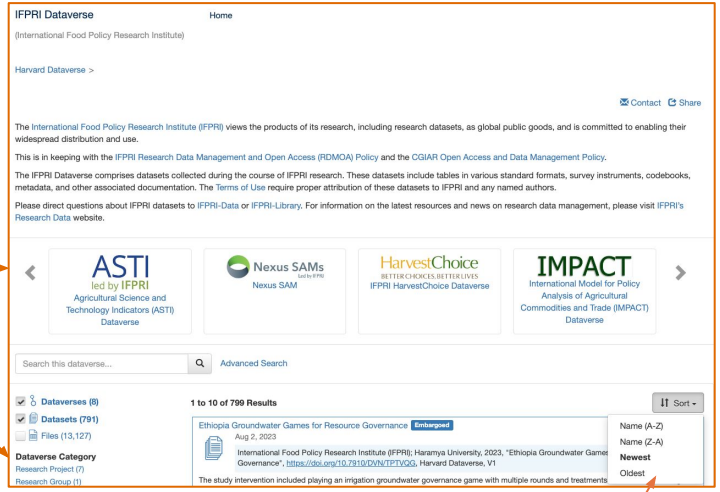
Search & browse

Dataset details:

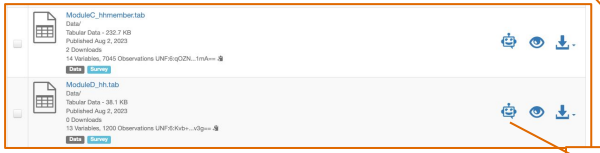
- Detailed dataset page
- Metadata
- Description
- Authors
- Citation
- Download options
- Documentation, README, code, etc.
- Multiple format download options

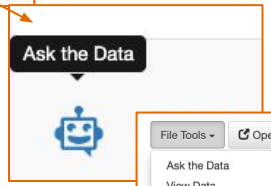
Browse:

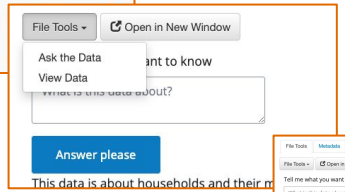
- Categories/Subjects
- Featured Datasets/collections

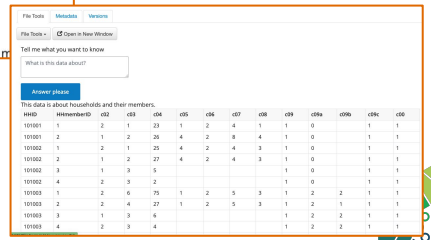


Browse: Recently added

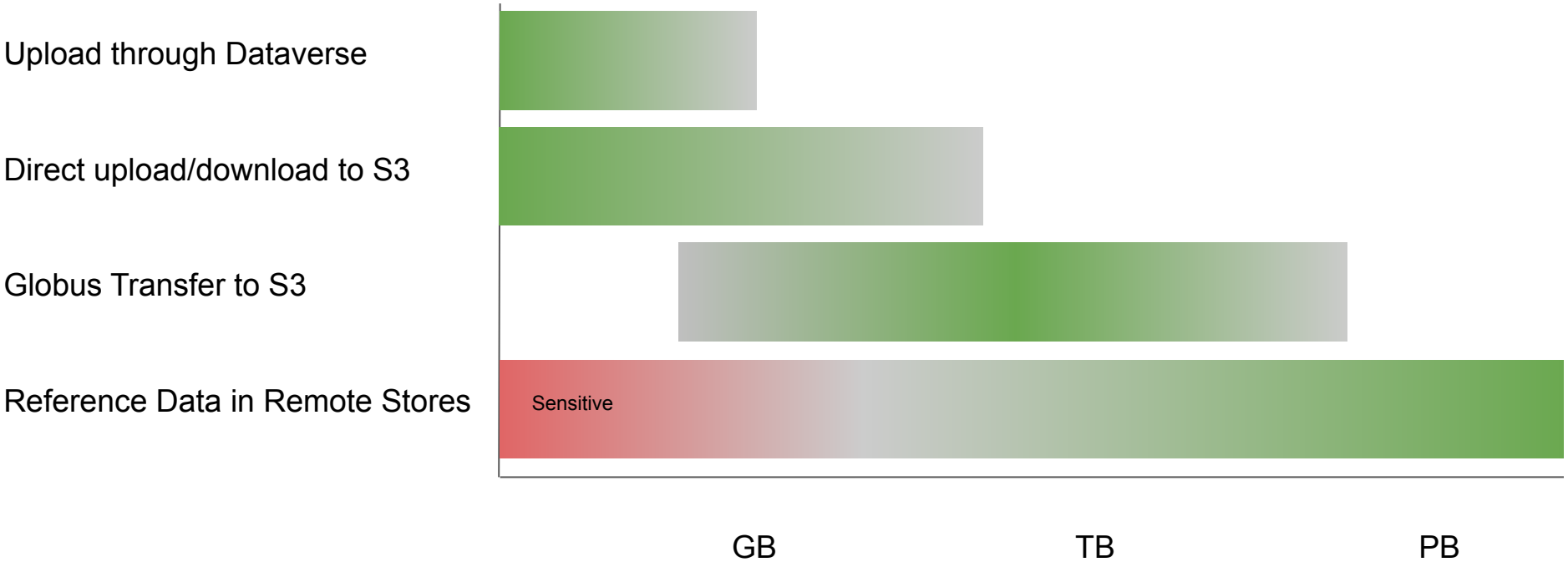








Large Data support in Dataverse



Through native Globus API (rather than S3) for user friendly interaction with remote storage

https://dataverse.org/software-features

Login via ORCID, Google, GitHub, or Microsoft

Log in using popular OAuth2 providers. More information.

Login via OpenID Connect (OIDC)

Log in using your institution's identity provider or a third party. More information.

Internationalization

The Dataverse software has been translated into multiple languages. More information.

Versioning

History of changes to datasets and files are preserved. More information.

Restricted files

Control who can download files and choose whether or not to enable a "Request Access" button. More information.

Embargo

Make content inaccessible until an embargo end date. More information.

External vocabulary

Let users pick from external vocabularies (provided via API/SKOSMOS) when filling in metadata. More information.

Dropbox integration

Upload files stored on Dropbox. More information.

GitHub integration

A GitHub Action is available to upload files from GitHub to a dataset. More information.

Integration with Jupyter notebooks

Datasets can be opened in Binder to run code in Jupyter notebooks, RStudio, and other computation environments. More information.

User management

Dashboard for common user-related tasks. More information.

Curation status labels

Let curators mark datasets with a status label customized to your needs. More information.

https://dataverse.org/software-features

Custom licenses
CC0 by default but add as many standard licenses as you like or create your own. More information.

Custom terms of use
Custom terms of use can be used in place of a license or disabled by an administrator. More information.

Publishing workflow support
Datasets start as drafts and can be submitted for review before publication. More information.

File hierarchy
Users are able to control dataset file hierarchy and directory structure. More information.

File previews
A preview is available for text, tabular, image, audio, video, and geospatial files. More information.

Preview and analysis of tabular files
Data Explorer allows for searching, charting and cross tabulation analysis More information.

Usage statistics and metrics
Download counters, support for Make Data Count. More information.

Branding
Your installation can be branded with a custom homepage, header, footer, CSS, etc. More information.

Backend storage on S3 or Swift
Choose between filesystem or object storage, configurable per collection and per dataset. More information.

Direct upload and download for S3
After a permission check, files can pass freely and directly between a client computer and S3. More information.

Export data in BagIt format
For preservation, bags can be sent to the local filesystem, Duracloud, and Google Cloud. More information.

Post-publication automation (workflows)
Allow publication of a dataset to kick off external processes and integrations. More information.

Pull header metadata from Astronomy (FITS) files
Dataset metadata prepopulated from FITS file metadata. More information.

Provenance
Upload standard W3C provenance files or enter free text instead. More information.

https://dataverse.org/software-features

Support for FAIR Data Principles
Findable, Accessible, Interoperable, Reusable. More information.

Data citation for datasets and files
EndNote XML, RIS, or BibTeX format at the dataset or file level. More information.

OAI-PMH (Harvesting)
Gather and expose metadata from and to other systems using standardized metadata formats: Dublin Core, Data Document Initiative (DDI), OpenAIRE, etc. More information.

APIs for interoperability and custom integrations
Search API, Data Deposit (SWORD) API, Data Access API, Metrics API, Migration API, etc. More information.

API client libraries
Interact with Dataverse APIs from Python, R, Javascript, Java, and Ruby More information.

DataCite integration
DOIs are reserved, and when datasets are published, their metadata is published to DataCite. More information.

Faceted search
Facets are data driven and customizable per collection. More information.

Customization of collections
Each personal or organizational collection can be customized and branded. More information.

Private URL
Create a URL for reviewers to view an unpublished (and optionally anonymized) dataset. More information.

Widgets
Embed listings of data in external websites. More information.

Notifications
In app and email notifications for access requests, requests for review, etc. More information.

Schema.org JSON-LD
Used by Google Dataset Search and other services for discoverability. More information.

https://dataverse.org/software-features

Support for FAIR Data Principles
Findable, Accessible, Interoperable, Reusable. More information.

Data citation for datasets and files
EndNote XML, RIS, or BibTeX format at the dataset or file level. More information.

OAI-PMH (Harvesting)
Gather and expose metadata from and to other systems using standardized metadata formats: Dublin Core, Data Document Initiative (DDI), OpenAIRE, etc. More information.

APIs for interoperability and custom integrations
Search API, Data Deposit (SWORD) API, Data Access API, Metrics API, Migration API, etc. More information.

API client libraries
Interact with Dataverse APIs from Python, R, Javascript, Java, and Ruby More information.

DataCite integration
DOIs are reserved, and when datasets are published, their metadata is published to DataCite. More information.

Faceted search
Facets are data driven and customizable per collection. More information.

Customization of collections
Each personal or organizational collection can be customized and branded. More information.

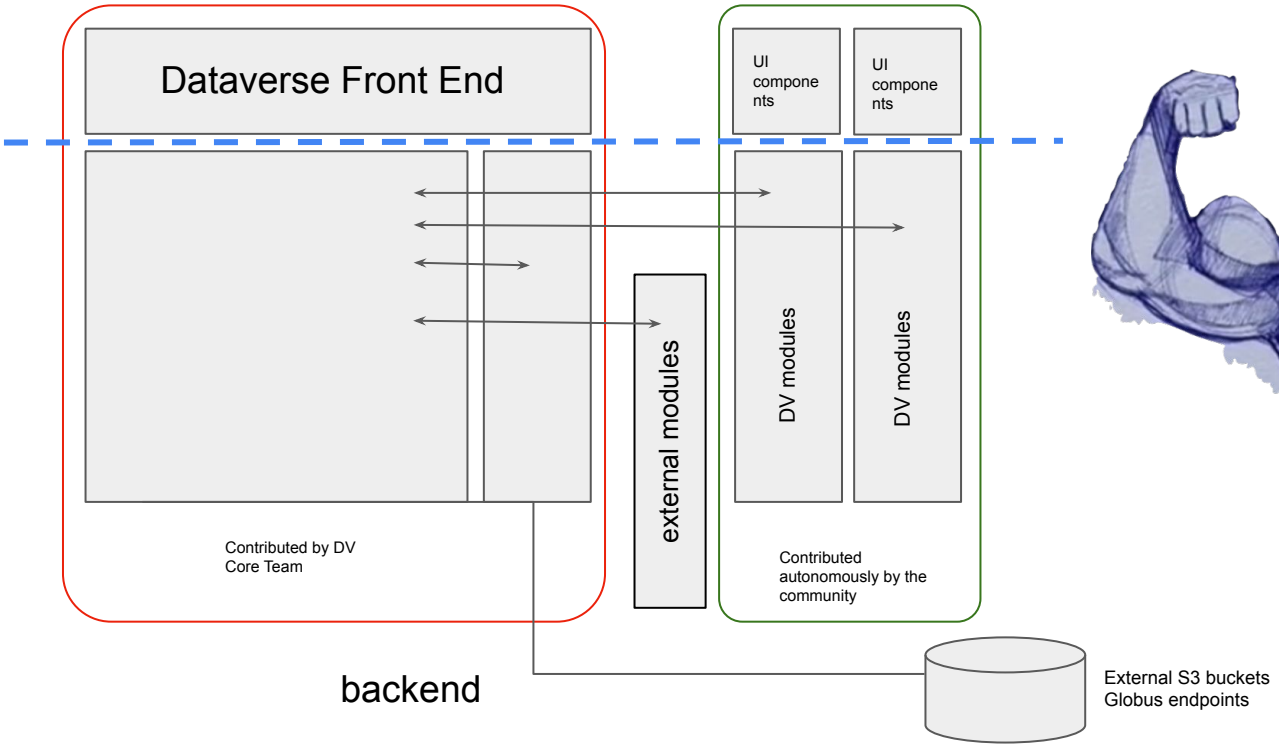
Private URL
Create a URL for reviewers to view an unpublished (and optionally anonymized) dataset. More information.

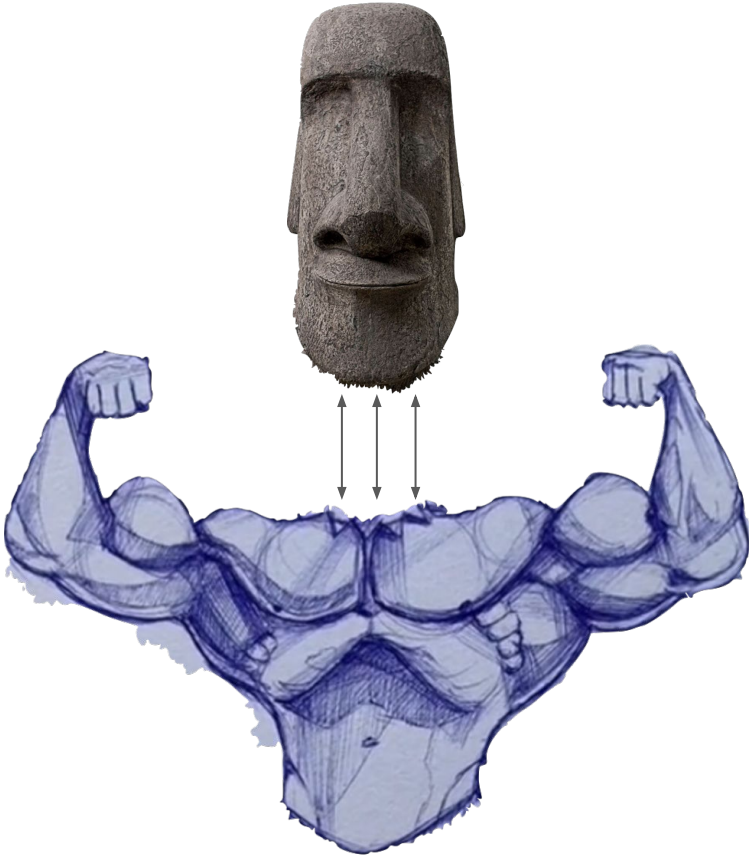
Widgets
Embed listings of data in external websites. More information.

Notifications
In app and email notifications for access requests, requests for review, etc. More information.

Schema.org JSON-LD
Used by Google Dataset Search and other services for discoverability. More information.

DV as an headless API-first application

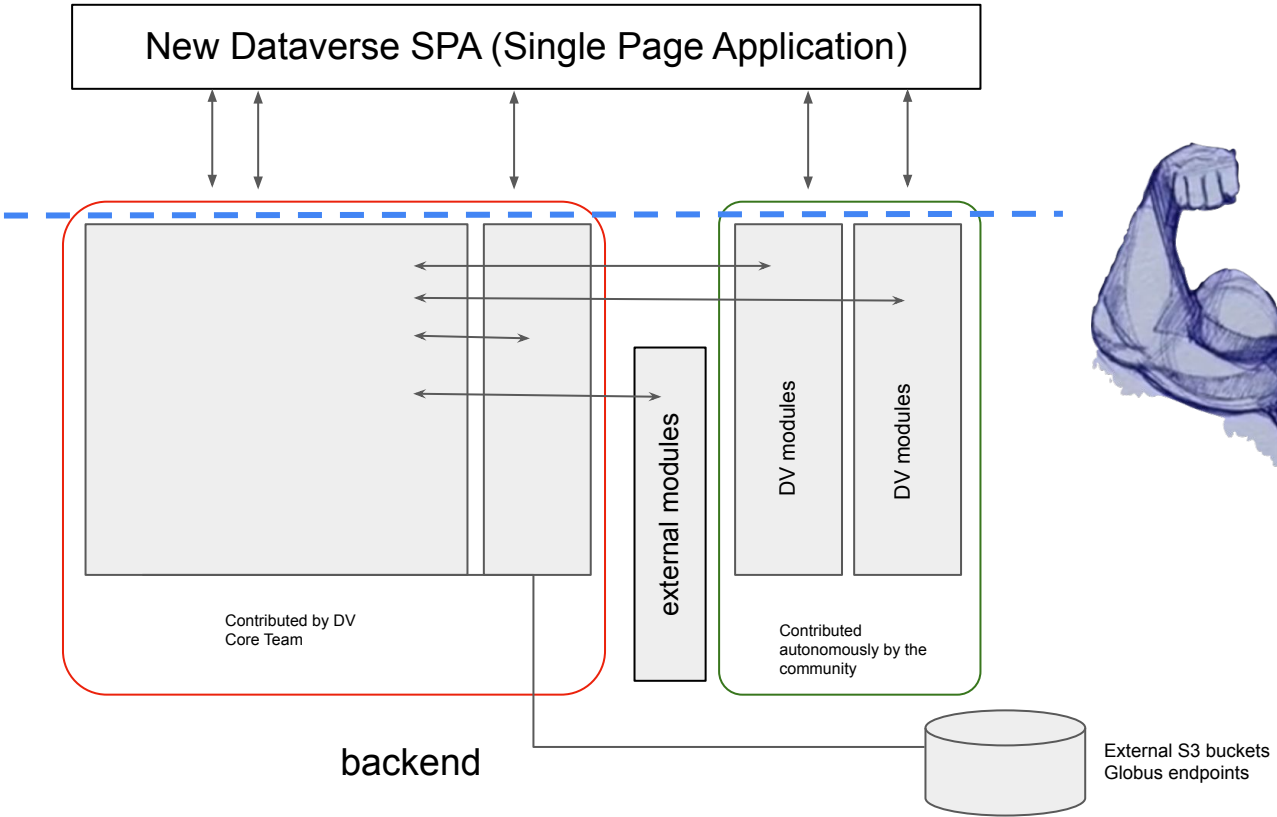


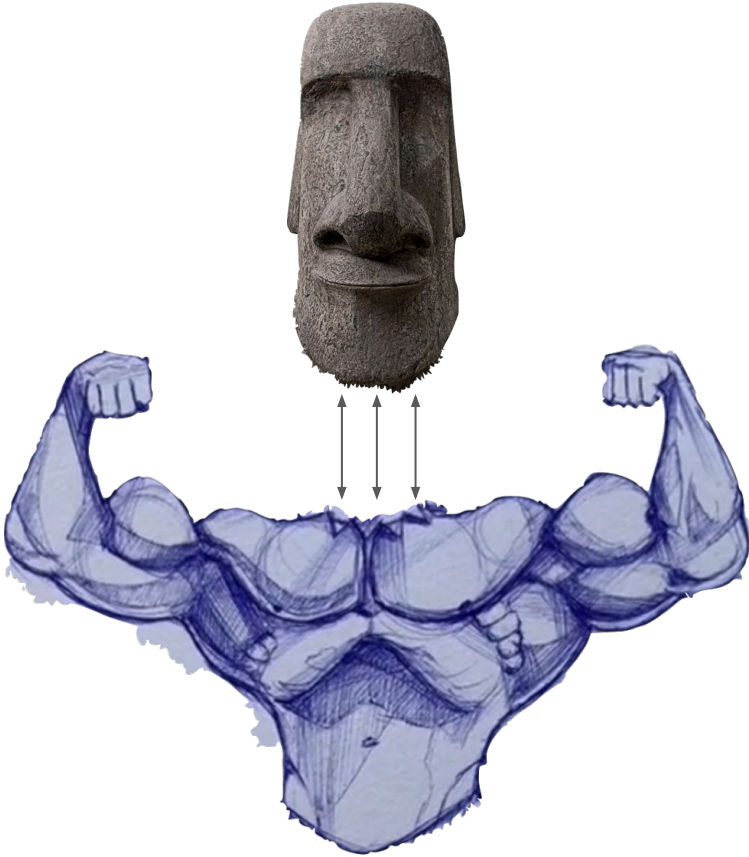


Re-architecture

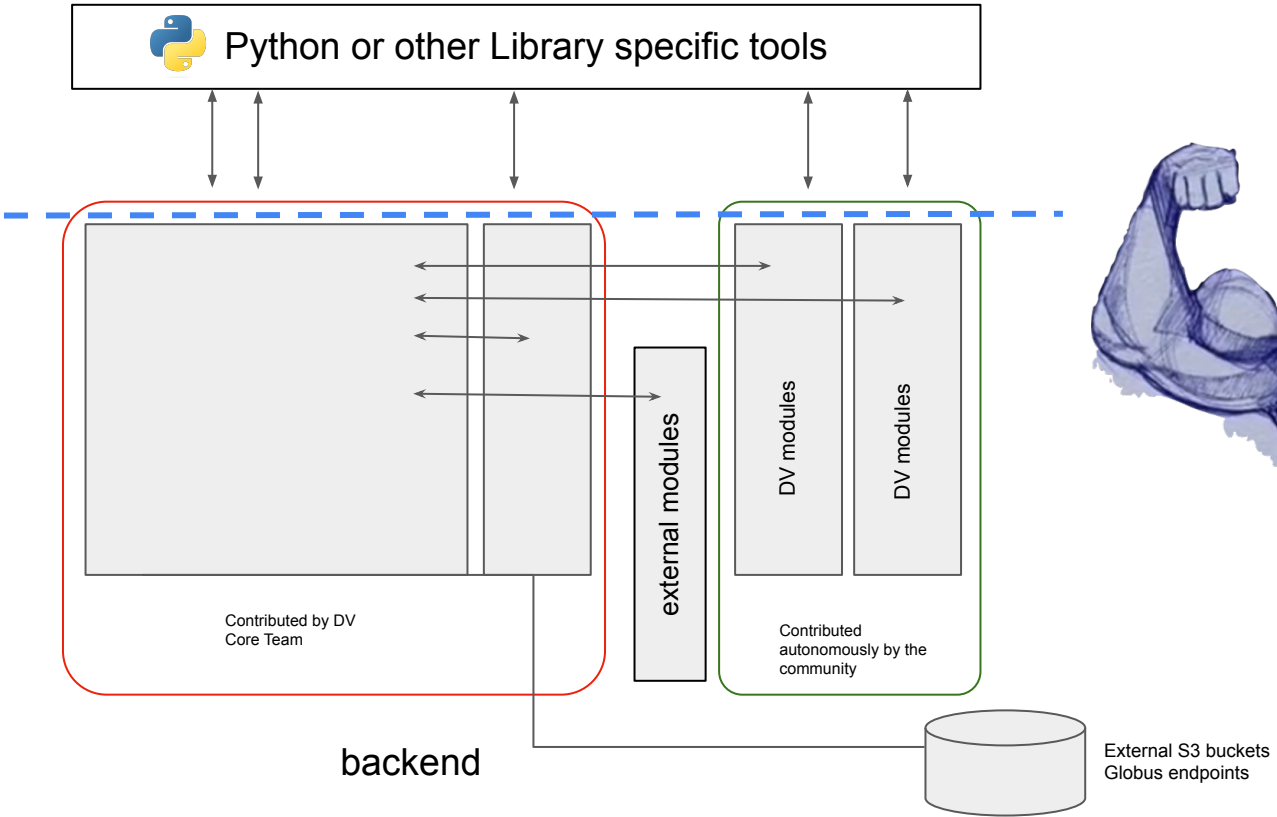
The Dataverse® Project

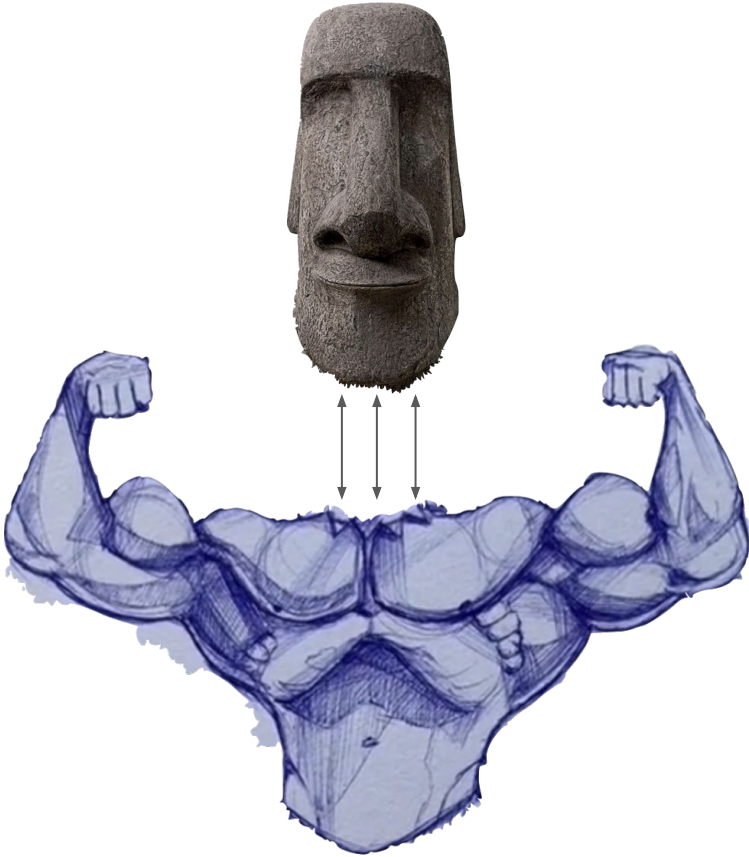
DV as an headless API-first application





DV as an headless API-first application





Join the Dataverse Community!

Harvard Dataverse: dataverse.harvard.edu

The Dataverse Project: dataverse.org

Test out features at demo.dataverse.org

Preservation policy and other governance information: https://support.dataverse.harvard.edu/harvard-dataverse-preservation-policy

Get help: support@dataverse.harvard.edu

Leave an issue about a bug or a feature: https://github.com/IQSS/dataverse/issues

Dataverse-Users Google Group: https://groups.google.com/g/dataverse-community

Bi-Weekly Community Call: https://dataverse.org/community-calls

HARVARD UNIVERSITY

IQSS

With contribution by & credits to:

Sonia Barbosa
Julian Gautier
Ceilyn Boyd
Katie Mika
the NIH GREI coopetition partners
and the Dataverse Team at large

siacus@iq.harvard.edu