Introduction to Using the Dataverse Supported Repository & the NIH DMP Dataverse Guidance Document

Sonia Maria Barbosa
Manager, Harvard Dataverse

Agenda

- The Harvard Dataverse Repository for Data Sharing and Use:
  - The Dataverse Project-Overview
  - The Dataverse Project and FAIR
  - Getting Started with The Dataverse Repository - Overview and Detailed
    - Collections
    - Datasets
    - Files/Sensitive Data
    - Search and Browse
    - Data Re-Use
  - NIH-GREI related improvements-forthcoming
- Overview of NIH DMP Guidance Document for Harvard Dataverse Repository
- Questions and Conversations

The Dataverse® Project

The Dataverse Project - Overview

The Dataverse Project

Open source research data repository software

Researchers

Enjoy full control over your data. Receive *web visibility, academic credit,* and *increased citation counts.* A personal Dataverse collection is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal Dataverse collection?

Journals

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data.* Participate in the open data movement by using a Dataverse collection as part of your journal data policy or list of repository recommendations. Want to find out more about journal Dataverse collections?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

Developers

Participate in a vibrant and growing community that is helping to drive the norms preserving, citing, exploring, and analyzing research data. Contribute code extens documentation, testing, and/or standards. *Integrate research analysis, visualizati tools,* or other research and data archival systems with the Dataverse Project. W

DATAVERSE REPOSITORIES - A WORLD VIEW

108 Installations

The Dataverse Project

Community

- Community Meetings
- Community Calls
- Global Dataverse Community Consortium

Best Practices

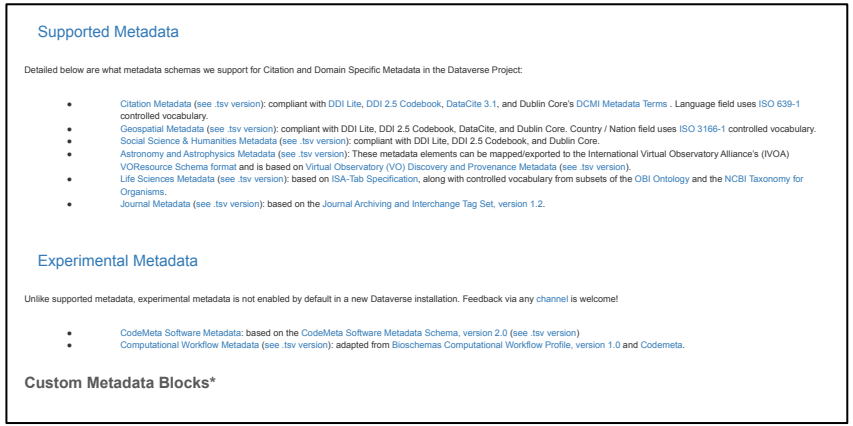
- Academic Credit
- Data Citation
- Dataverse Community Norms
- Data Management
- Replication Dataset Guidelines

Software

- Goals, Roadmap, and Releases
- Collaborations
- Integrations
- Features
- Source Code
- Guides

Dataverse for Data Management and Sharing

- Ease of data deposits
- Data citation
- Data access control
- Metadata standards
- Data sharing policies
- Data discovery



Supported Metadata

Detailed below are what metadata schemas we support for Citation and Domain Specific Metadata in the Dataverse Project:

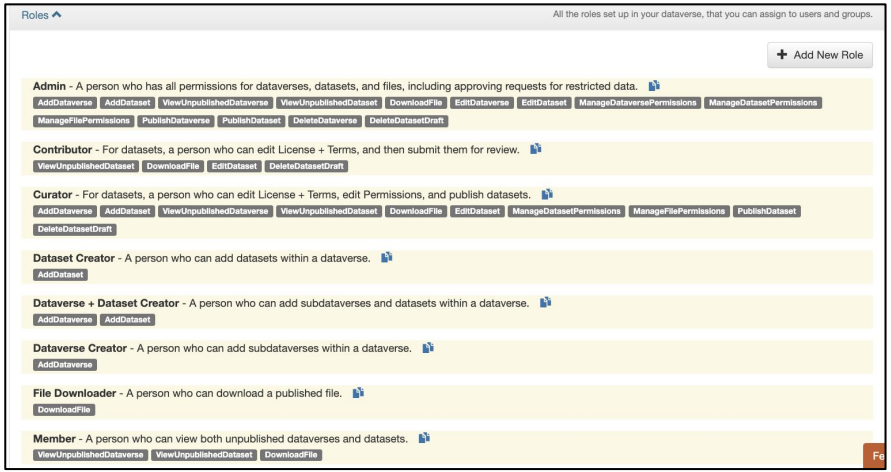
- Citation Metadata (see .tsv version): compliant with DDI Lite, DDI 2.5 Codebook, DataCite 3.1, and Dublin Core's DCMI Metadata Terms . Language field uses ISO 639-1 controlled vocabulary.
- Geospatial Metadata (see .tsv version): compliant with DDI Lite, DDI 2.5 Codebook, DataCite, and Dublin Core. Country / Nation field uses ISO 3166-1 controlled vocabulary.
- Social Science & Humanities Metadata (see .tsv version): compliant with DDI Lite, DDI 2.5 Codebook, and Dublin Core.
- Astronomy and Astrophysics Metadata (see .tsv version): These metadata elements can be mapped/exported to the International Virtual Observatory Alliance's (IVOA) VOResource Schema format and is based on Virtual Observatory (VO) Discovery and Provenance Metadata (see .tsv version).
- Life Sciences Metadata (see .tsv version): based on ISA-Tab Specification, along with controlled vocabulary from subsets of the OBI Ontology and the NCBI Taxonomy for Organisms.
- Journal Metadata (see .tsv version): based on the Journal Archiving and Interchange Tag Set, version 1.2.

Experimental Metadata

Unlike supported metadata, experimental metadata is not enabled by default in a new Dataverse installation. Feedback via any channel is welcome!

- CodeMeta Software Metadata: based on the CodeMeta Software Metadata Schema, version 2.0 (see .tsv version)
- Computational Workflow Metadata (see .tsv version): adapted from Bioschemas Computational Workflow Profile, version 1.0 and Codemeta.

Custom Metadata Blocks*



Dataverse Repository Features

The Dataverse® Project

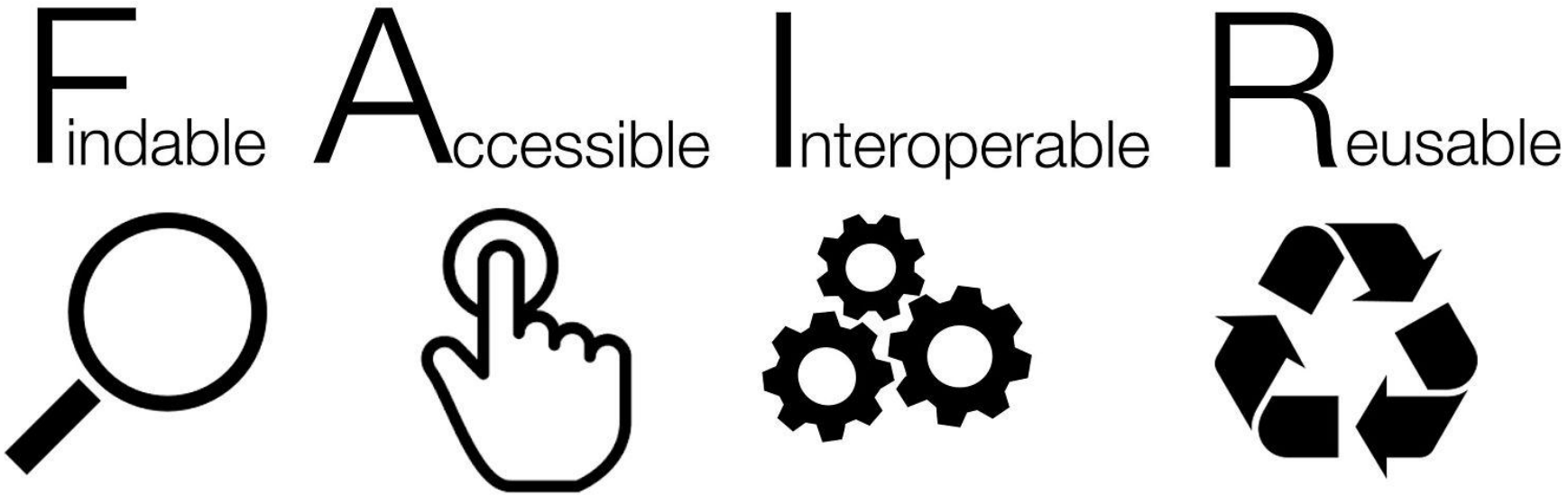
- Support for FAIR Data Principles
- Data citation for datasets and files
- OAI-PMH (Harvesting)
- APIs for interoperability and custom integrations
- API client libraries
- DataCite integration
- Login via Shibboleth
- Login via ORCID, Google, GitHub, or Microsoft
- Login via OpenID Connect (OIDC)
- Internationalization
- Versioning
- Restricted files
- Embargo
- Custom licenses
- Custom terms of use
- Publishing workflow support
- File hierarchy
- File previews
- Preview and analysis of tabular files
- Usage statistics and metrics
- Guestbook
- Fixity checks for files
- File download in R and TSV format
- Faceted search
- Customization of collections
- Private URL
- Widgets
- Notifications
- Schema.org JSON-LD
- External tools
- External vocabulary
- Dropbox integration
- GitHub integration
- Integration with Jupyter notebooks
- User management
- Curation status labels
- Branding
- Backend storage on S3 or Swift
- Direct upload and download for S3
- Export data in BagIt format
- Post-publication automation (workflows)
- Pull header metadata from Astronomy (FITS) files
- Provenance
- Support for rsync
- Auxiliary files for data files

https://dataverse.org/software-features

The Dataverse® Project

The Dataverse Project and FAIR

FAIR Principles and Dataverse



https://en.wikipedia.org/wiki/File:FAIR_data_principles.jpg

How Dataverse Supports FAIR

- FINDABLE: Persistent Identifiers (DOIs, Handles) for datasets (and files), and enabling metadata indexing by search engines
- ACCESSIBLE: Open access to (meta)data and ensures data can be downloaded in machine-readable formats
- INTEROPERABLE: Provides standardised metadata schemas and enabling the integration of of data with research tools and platforms
- REUSABLE: Licenses that clearly state how data can be used and ensuring that data are well documented and preserved for long term use

Getting started with the Dataverse Repository - Overview

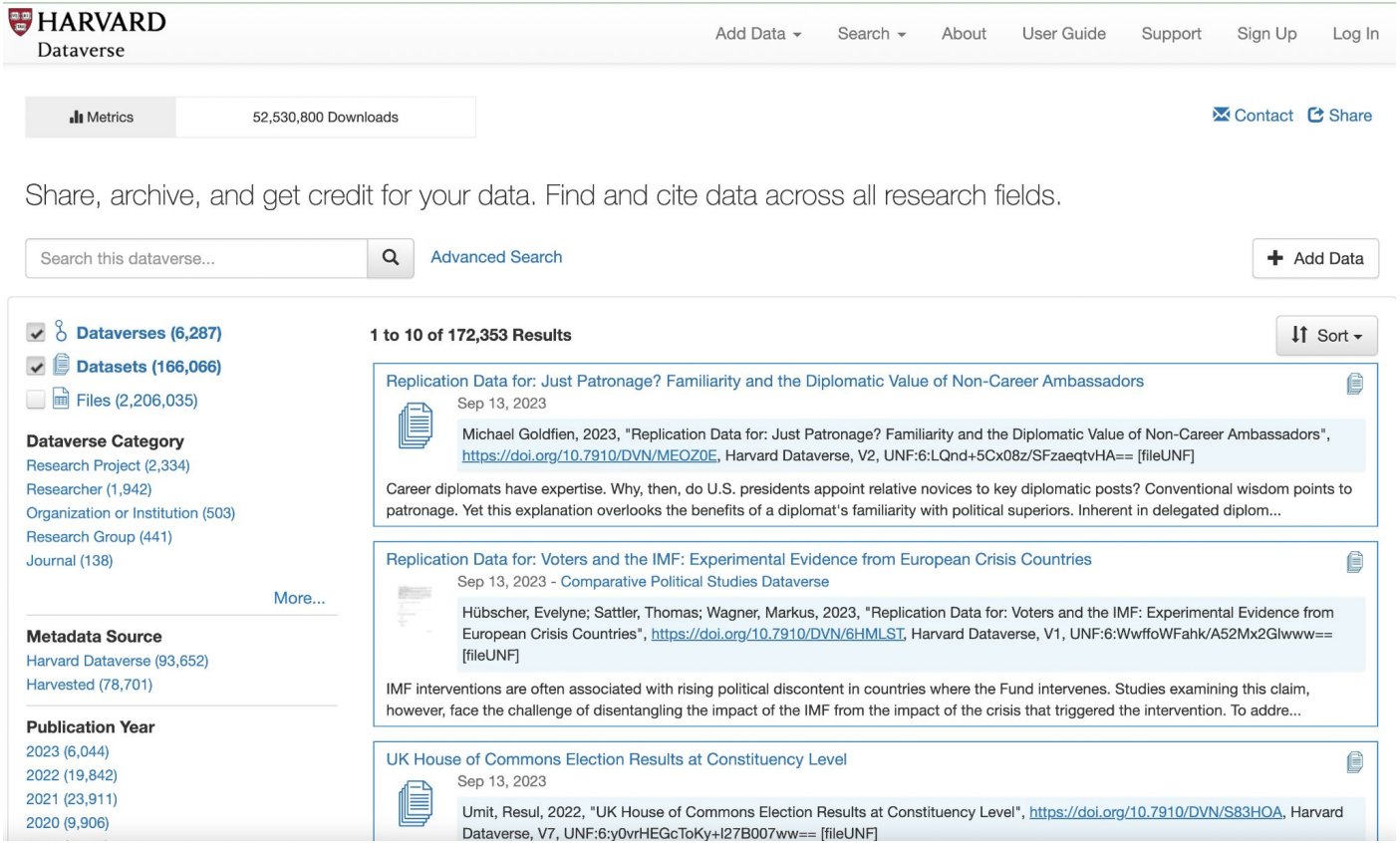
The Purpose of a Data Repository

The **purpose of a data repository** is to keep a certain population of data so that it can be mined for greater insight or business intelligence or to be used for a specific reporting need.
https://en.wikipedia.org/wiki/Information_repository

The **Registry of Research Data Repositories** (**re3data.org**) is an Open Science tool that offers researchers, funding organizations, libraries and publishers an overview of existing international repositories for research data.

Data should be submitted to discipline-specific, community-recognized repositories where possible. **Where a suitable discipline-specific resource does not exist, data should be submitted to a generalist repository.**

HARVARD Dataverse



Visit the Harvard Repository Policy page before getting started: https://support.dataverse.harvard.edu/policies

Prepare Your Data for Sharing in The Harvard Repository

HARVARD Dataverse

GATHER AND VERIFY

Determine what will be shared and ensure the metadata describes what is being shared: raw data, output data, documentation, readme, codebook, etc...

Can you share the data openly? Copyright? Deidentification? (human subjects, vulnerable/endangered species); Access restrictions needed (restrict. embargo)

What metadata standards are required to describe your data properly?

*Sensitive data support is not yet available in Harvard Dataverse Repository. All data must be deidentified!

CHOOSE AND ORGANIZE

Nonproprietary file formats for easy access, interoperability, preservation , plain text, same file but in open format version...

Organize the data by check all files for corruption, detailed file names, and usefulness. Folders should be clearly labeled and organized within zip files.

Know the file size limitations in HDV (Total project size varies 1TB to 2.5TB free, file size 2.5GB)

Tabular data get additional functionality in HDV

Know the tools available for use in HDV

SHARE YOUR DATA

Determine the repository (Domain specific should always be the first choice)

Share in one repository for one DOI (different components can be shared between domain specific and generalists but avoid sharing the same component in multiple repos)

Documentation such as README files should describe the data thoroughly and completely. Test your readme file instructions for reproducibility to ensure they are correct.

Share your citation and collection space URL

*all deposits must contain reusable data

Create a Collection, Deposit Datasets, and Upload Files

Collection-level

Current Features

Dataverse Collections

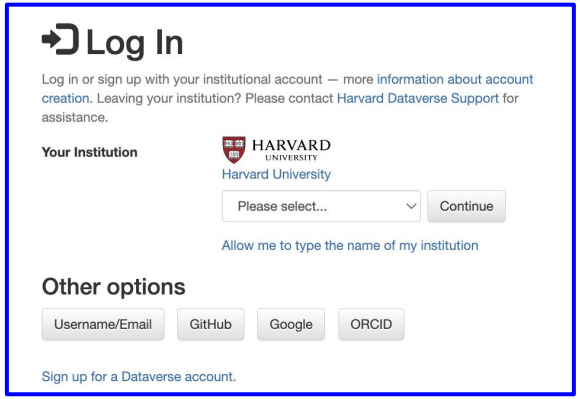
- Own administration
- Own branding (and can be embedded anywhere)

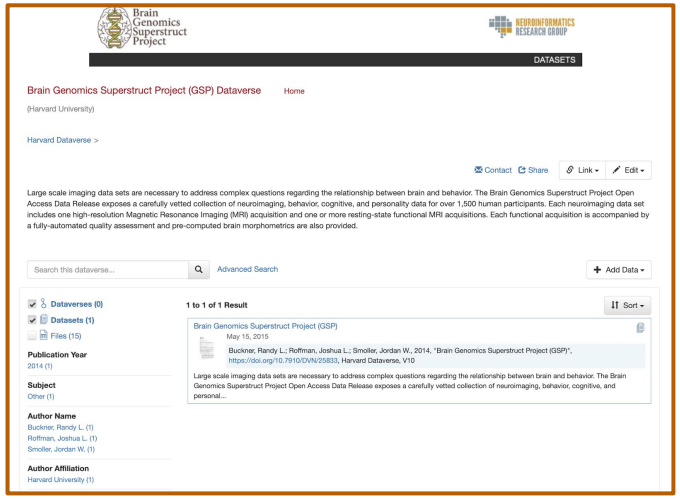
Datasets

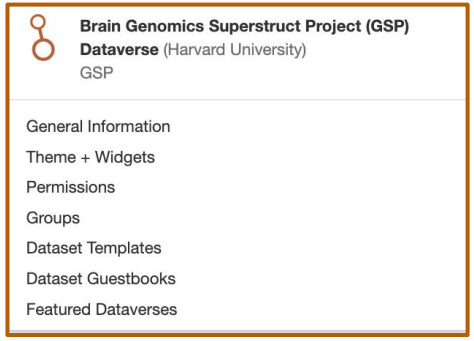
- Citation
- Metadata
- Versioning
- Private URL
- Custom Terms/Multiple License/Permissions
- Guestbooks
- Publishing Workflows

Files

- Citation
- Ingest
- Preview/Explore
- Metadata
- Versioning
- Permissions/Embargo/Restrictions







Dataset-level

Current Features

Dataverse Collections

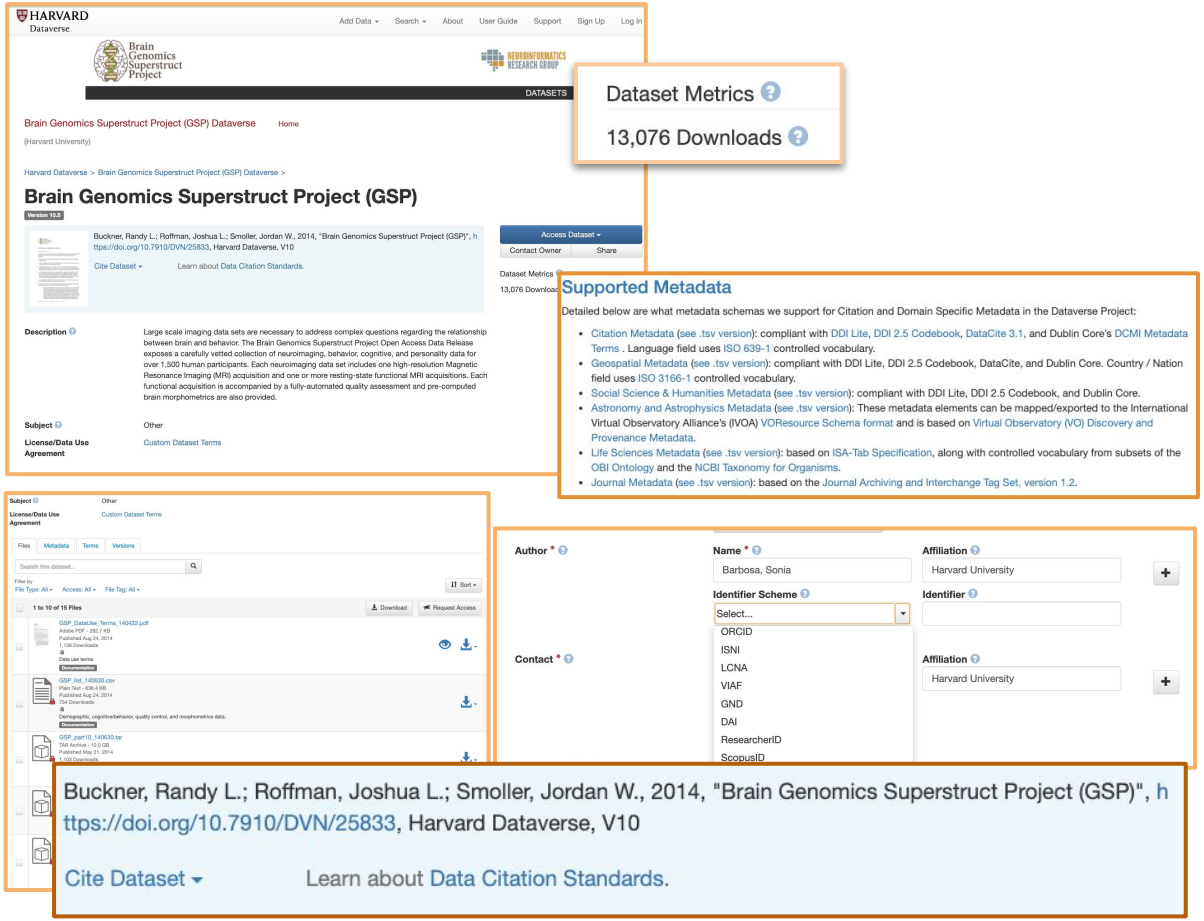
- Own administration
- Own branding (and can be embedded anywhere)

Datasets

- Citation
- Metadata
- Versioning
- Private URL
- Custom Terms/Multiple License/Permissions
- Guestbooks
- Publishing Workflows

Files

- Citation
- Ingest
- Preview/Explore
- Metadata
- Versioning
- Permissions/Embargo/Restrictions



File-level

Current Features

Dataverse Collections

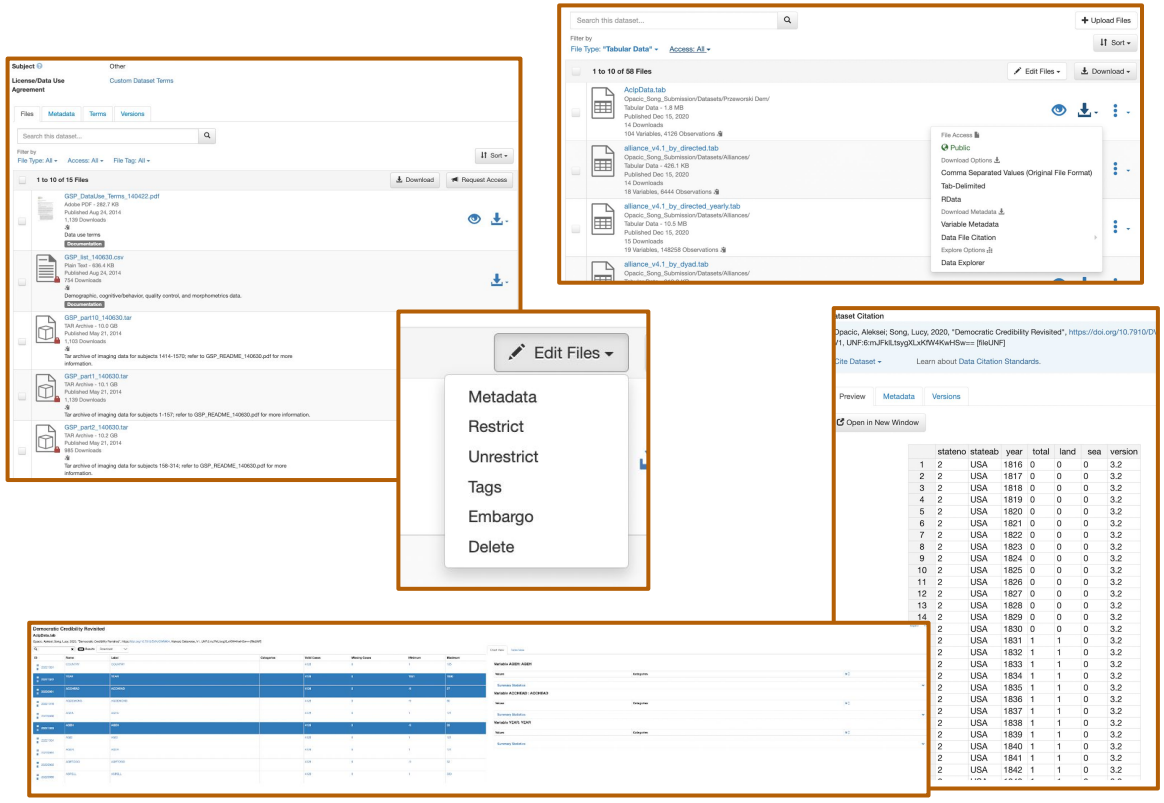
- Own administration
- Own branding (and can be embedded anywhere)

Datasets

- Citation
- Metadata
- Versioning
- Private URL/*Anonymous Peer Review
- Custom Terms/*Multiple Licenses/Permissions
- Guestbooks
- Publishing Workflows

Files

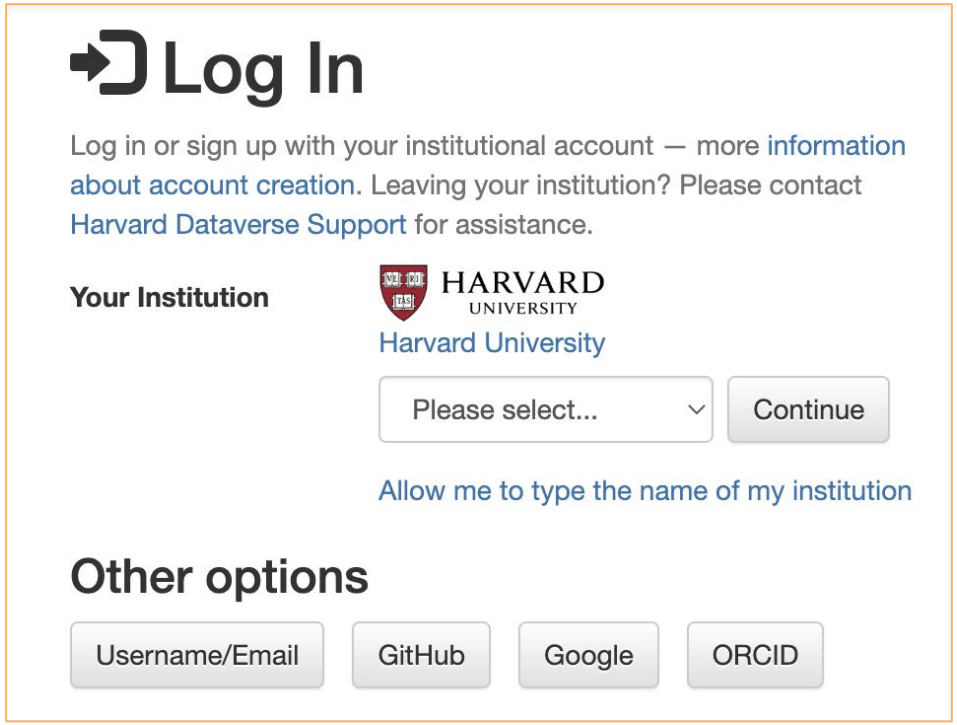
- Citation
- Ingest
- Preview/Explore
- Metadata/Provenance
- Versioning
- Permissions/Embargo/Restrictions



HARVARD
Dataverse

Get Started on a Dataverse Repository - Detailed

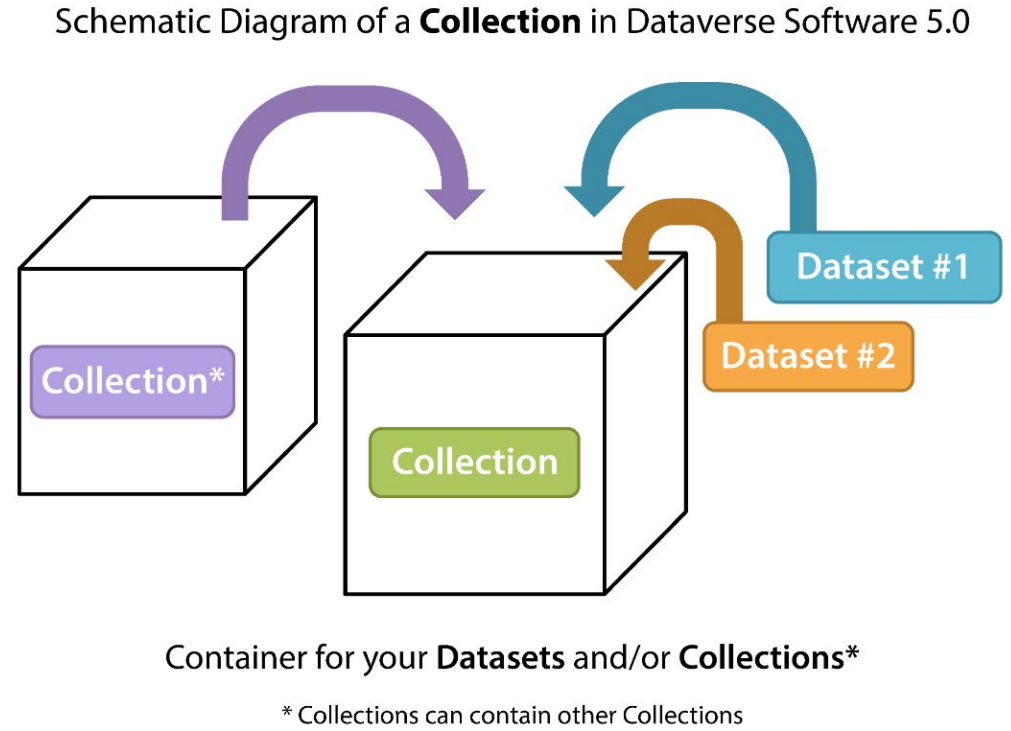
Account Creation



Create a Dataverse "Collection"

- Edit Dataverse Collection
  - General Information
  - **Theme**
  - Widgets
    - Dataverse Collection Search Box Widget
    - Dataverse Collection Listing Widget
    - Adding Widgets to an OpenScholar Website
  - **Roles & Permissions**
    - Setting Access Configurations
    - Assigning Roles to Users and Groups
  - Dataset **Templates**
  - Dataset **Guestbooks**
  - Featured Dataverse Collection
- Dataset Linking
- Dataverse Collection Linking
- **Publish Your Dataverse Collection**

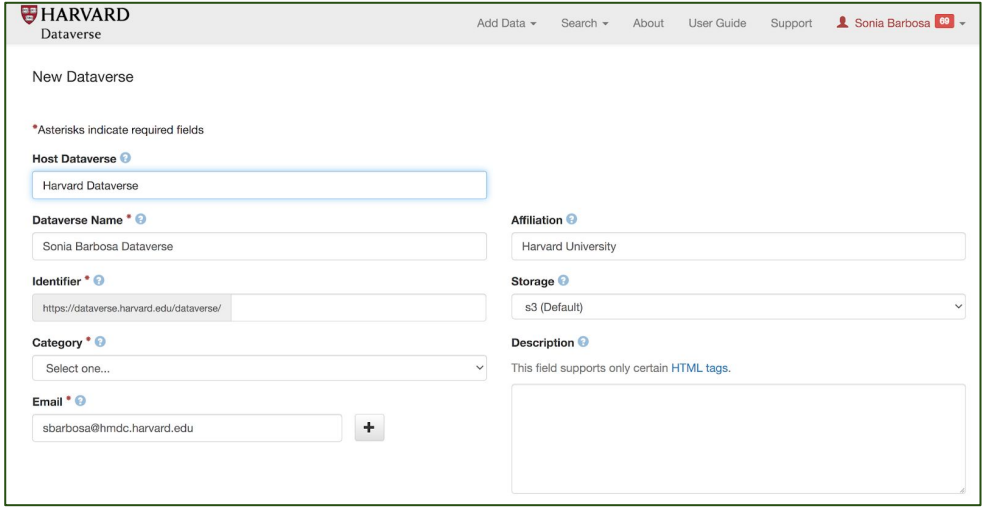
A Dataverse collection is a container for datasets (research data, code, documentation, and metadata) and other Dataverse collections, which can be setup for individual researchers, departments, journals and organizations.

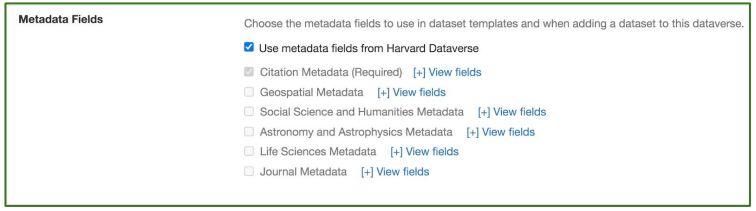


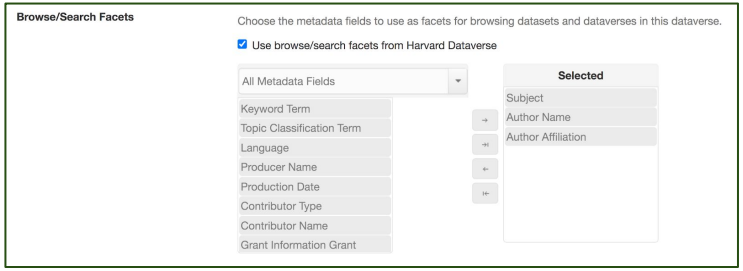
*contact support@dataverse.harvard.edu and request the creation of a collection

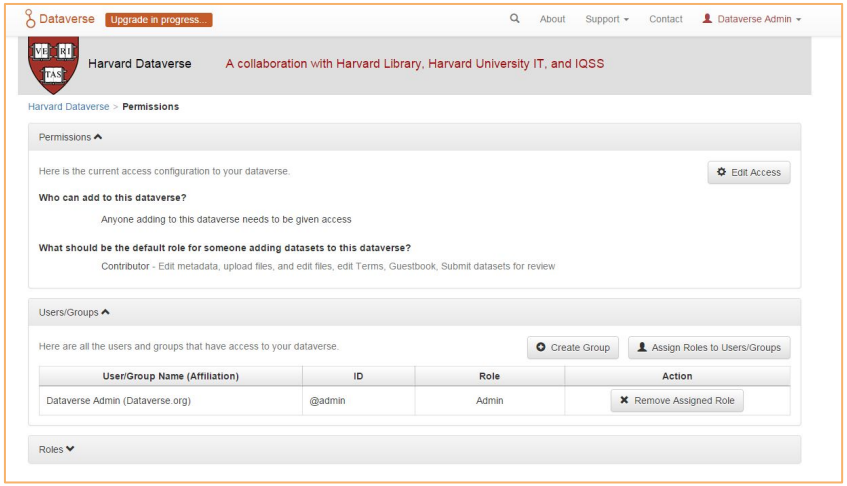
- Metadata
- Searchable facets
- Holds datasets

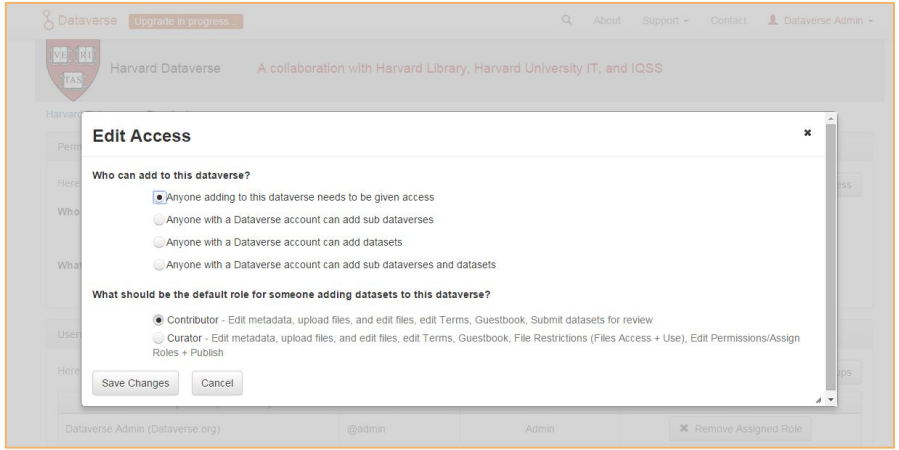
Customizable with descriptions, logos, collection space URLs, affiliation, contact emails, permissions, metadata and searchable facets. **Dataverse collections hold datasets.**

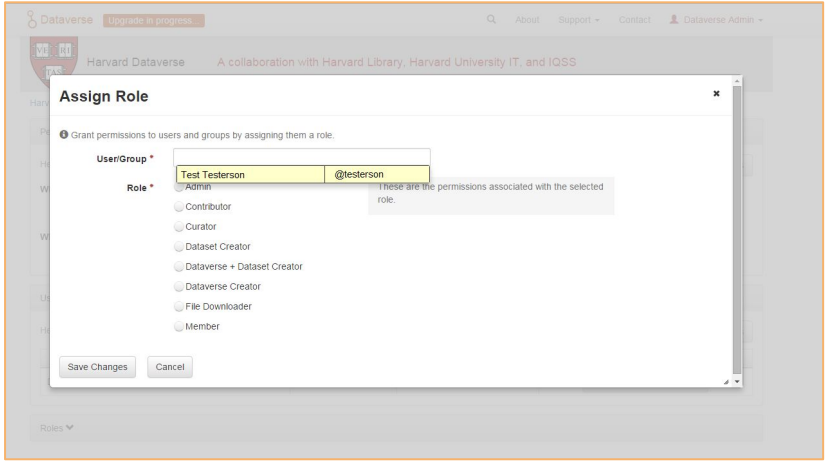


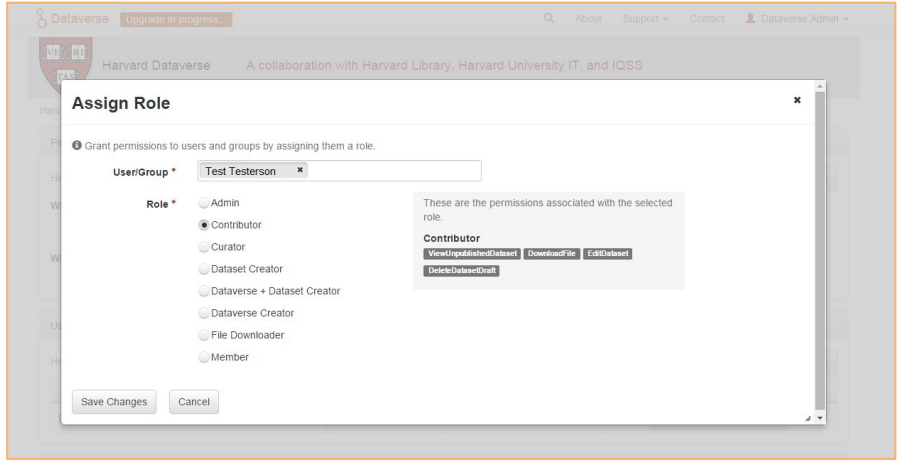


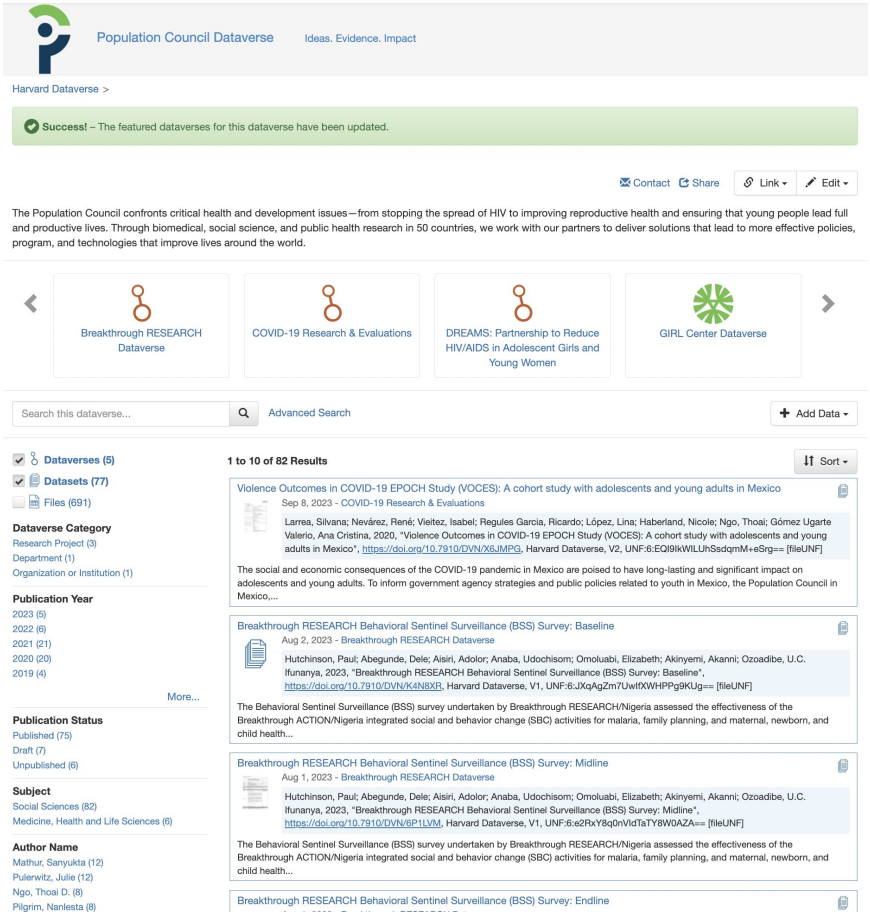



Dataverse
Upgrade in progress...
About
Support
Contact
Dataverse Admin
Harvard Dataverse
A collaboration with Harvard Library, Harvard University IT, and IQSS
Harvard Dataverse > Permissions
Permissions
Here is the current access configuration to your dataverse.
Edit Access
Who can add to this dataverse?
Anyone adding to this dataverse needs to be given access
What should be the default role for someone adding datasets to this dataverse?
Contributor - Edit metadata, upload files, and edit files, edit Terms, Guestbook, Submit datasets for review
Users/Groups
Here are all the users and groups that have access to your dataverse.
Create Group
Assign Roles to Users/Groups
User/Group Name (Affiliation)
ID
Role
Action
Dataverse Admin (Dataverse.org)
@admin
Admin
Remove Assigned Role
Roles


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
What should be the default role for someone adding datasets to this dataverse?
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


Assign Role
Grant permissions to users and groups by assigning them a role.
User/Group *
Test Testerson
@testerson
Role *
Admin
Contributor
Curator
Dataset Creator
Dataverse + Dataset Creator
Dataverse Creator
File Downloader
Member
These are the permissions associated with the selected role.
Save Changes
Cancel


Assign Role
Grant permissions to users and groups by assigning them a role.
User/Group *
Test Testerson
Role *
Admin
Contributor
Curator
Dataset Creator
Dataverse + Dataset Creator
Dataverse Creator
File Downloader
Member
These are the permissions associated with the selected role.
Contributor
ViewUnpublishedDataset
DownloadFile
EditDataset
DeleteDatasetDraft
Save Changes
Cancel



https://dataverse.harvard.edu/dataverse/popcouncil/

Time Period End Date
2018 (7)
2017 (4)
2016 (3)
2022 (3)
2019 (2)
More...

Data Type
Quantitative (36)
Qualitative (10)
Survey (6)
Quantitative data (2)
quantitative (2)
More...

Geographic Coverage State / Province
Uttar Pradesh (6)
Bihar (5)
Kilifi (3)
Kisumu (3)
Kiambu (2)
More...

Geographic Unit
Census Supervisory Area (CSA) (1)
Global South (1)
Neighborhood (1)
School (1)
Village (1)

Unit of Analysis
Individual (42)
2,156 female respondents aged 18-49 years selected one per household (2)
Individuals (2)
2067 patient participants (1)
Adolescent and adult caregiver dyads (1)
More...

Universe
Adolescent girls ages 10-19 (1)
Females and males age 15-18 receiving services from two service delivery organizations: Biruh Tesfa and ReTrak. (1)
Girls ages 11-15 and the head of their household (1)
In and out-of-school males and females ages 14-17 at baseline (1)
Married girls younger than 20 years of age (1)
More...

Publication Year
2023 (5)
2022 (6)
2021 (21)
2020 (20)
2019 (4)
More...

Publication Status
Published (75)
Draft (7)
Unpublished (6)

Subject
Social Sciences (82)
Medicine, Health and Life Sciences (6)

Author Name
Mathur, Sanyukta (12)
Pulerwitz, Julie (12)
Ngo, Thoai D. (8)
Pilgrim, Nanlesta (8)
Gottert, Ann (7)
More...

Keyword Term
COVID-19 (13)
HIV (10)
Implementation science (7)
family planning (7)
Gender Inequitable Attitudes/Beliefs (6)
More...

Geographic Coverage Country / Nation
Kenya (14)
India (9)
Malawi (7)
Nigeria (7)
Zambia (7)
More...

Geographic Coverage City
Nairobi (3)
Addis Ababa (1)
Assiut (1)
Bar Elias (1)
Bogura (1)
More...

Author Affiliation
Population Council (65)
Centre for Research, Evaluation Resources and Development (CRERD) (3)
Tulane University (3)
CSK Research Solutions (2)
Columbia University, Mailman School of Public Health (2)
More...

Topic Classification Term
Asian American (1)
HIV and AIDS (1)
concept testing (1)

Language
English (20)
Arabic (2)
Somali (1)
Swahili (1)

Producer Name
Population Council (9)
Arthur Davison Children's Hospital (1)
Center for Tuberculosis Research, Johns Hopkins University School of Medicine (1)
Department of Internal Medicine, Klerksdorp/Tshepong Hospital Complex, North West Province Department of Health (1)
Elizabeth Glaser Pediatric AIDS Foundation (1)
More...

Production Date
2021 (1)

Funding Information Agency
Bill and Melinda Gates Foundation (17)
Bill & Melinda Gates Foundation (11)
United States Agency for International Development (10)
U.S. Agency for International Development (3)
3ie: International Initiative for Impact Evaluation (2)
More...

Funding Information Identifier
AID-OAA-140060 (2)
AID-OAA-A-14-00060 (2)
AID-OAA-A-17-00018 (1)
OPP1136778 (1)
OPP1150068 (1)
More...

Deposit Date
2023 (4)
2022 (7)
2021 (16)
2020 (25)
2019 (4)
More...

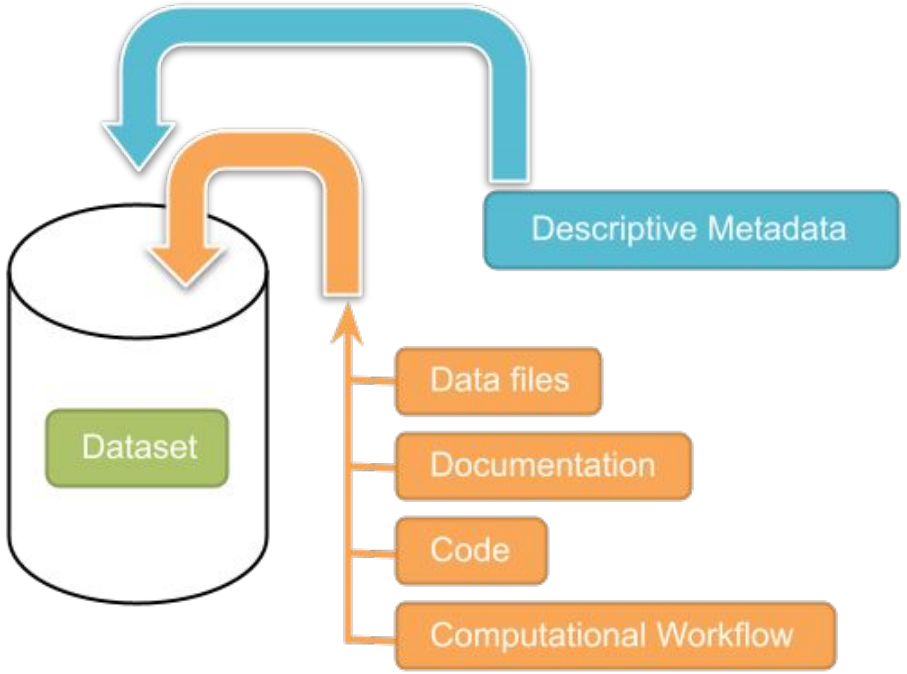
Time Period Start Date
2017 (8)
2018 (4)
2016 (3)
2020 (3)
2012 (2)
More...

Time Period End Date
2018 (7)
2017 (4)
2016 (3)
2022 (3)
2019 (2)
More...

Deposit a Dataset

- Supported Metadata
- Adding a New Dataset
- File Upload
- File Handling
- Restricted Files
- Edit Files
- Replace Files
- Terms
- Roles & Permissions
- Data Provenance
- Thumbnails + Widgets
- Publish Dataset
- Submit for Review
- Private URL to Review Unpublished Dataset
- Embargoes
- Dataset Versions
    - Version Details
- Dataset Metrics and Make Data Count
- Cloud Storage + Computing
    - Cloud Computing
    - Cloud Storage Access
- Dataset Deaccession

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



Container for your data, documentation, code, and computational workflow.

Datasets - Metadata

Detailed below are what metadata schemas we support for Citation and Domain Specific Metadata in the Dataverse Project:

- Citation Metadata: compliant with DDI Lite, DDI 2.5 Codebook, DataCite 3.1, and Dublin Core's DCMI Metadata Terms (see .tsv version). Language field uses ISO 639-1 controlled vocabulary.
- Geospatial Metadata: compliant with DDI Lite, DDI 2.5 Codebook, DataCite, and Dublin Core (see .tsv version). Country / Nation field uses ISO 3166-1 controlled vocabulary.
- Social Science & Humanities Metadata: compliant with DDI Lite, DDI 2.5 Codebook, and Dublin Core (see .tsv version).
- Astronomy and Astrophysics Metadata : These metadata elements can be mapped/exported to the International Virtual Observatory Alliance's (IVOA) VOResource Schema format and is based on Virtual Observatory (VO) Discovery and Provenance Metadata (see .tsv version).
- Life Sciences Metadata: based on ISA-Tab Specification, along with controlled vocabulary from subsets of the OBI Ontology and the NCBI Taxonomy for Organisms (see .tsv version).
- Journal Metadata: based on the Journal Archiving and Interchange Tag Set, version 1.2 (see .tsv version).

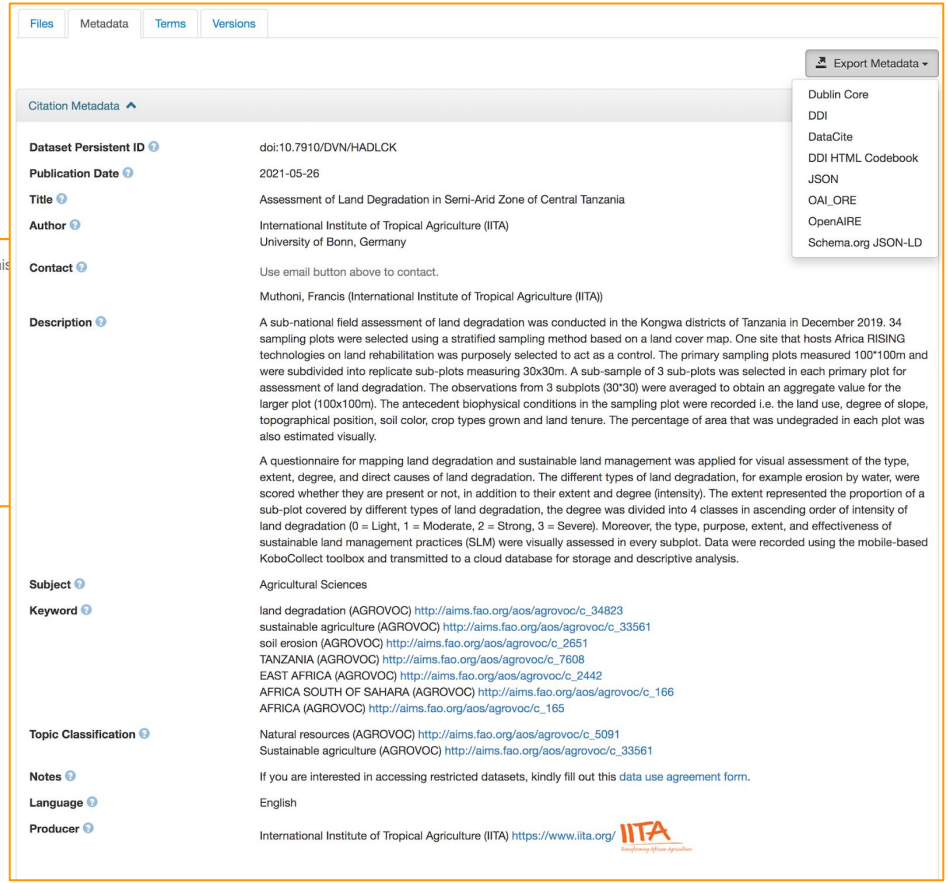
See also the Dataverse Software 4.0 Metadata Crosswalk: DDI, DataCite, DC, DCTerms, VO, ISA-Tab document and the Metadata Customization section of the Admin Guide.

Datasets - Metadata

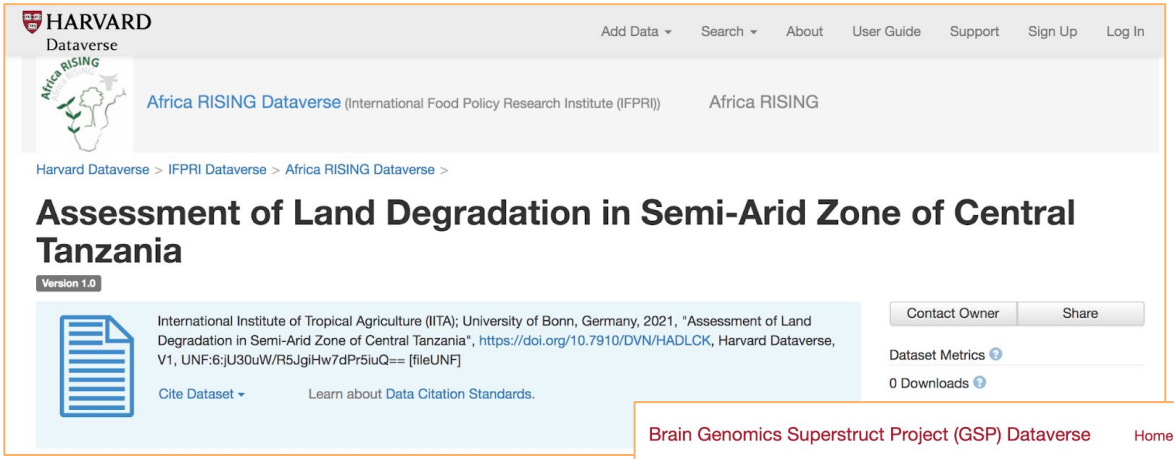
Metadata Fields

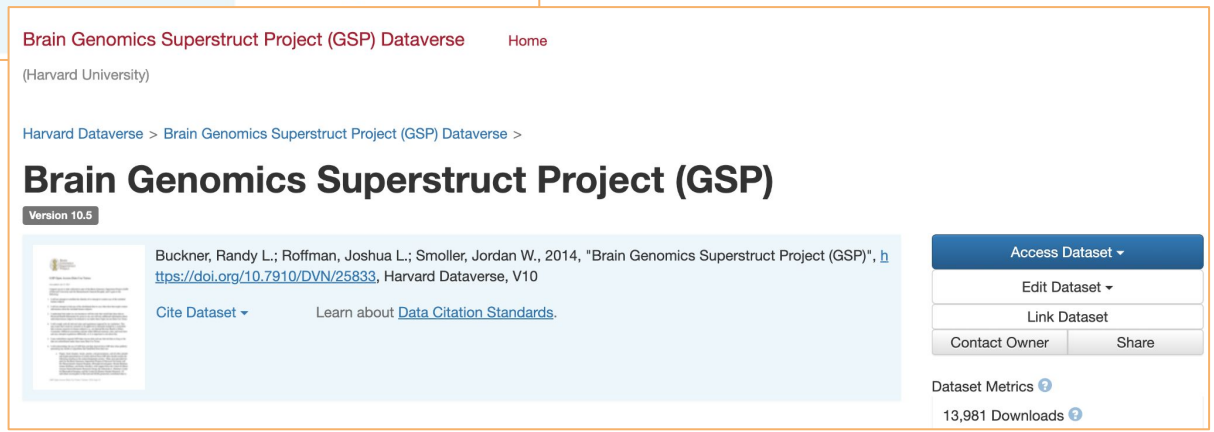
Choose the metadata fields to use in dataset templates and when adding a dataset to this

- [x] Use metadata fields from Harvard Dataverse
  - [x] Citation Metadata (Required) [+] View fields
  - [ ] 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



Datasets - Citations



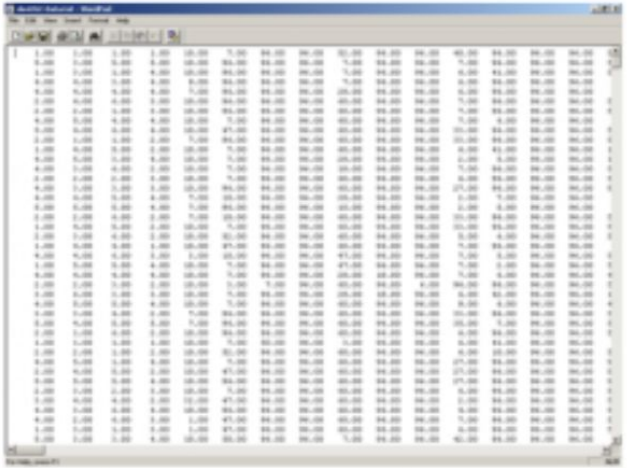


Data - Citations

Your Publication

Formal Data Citation

Your Data



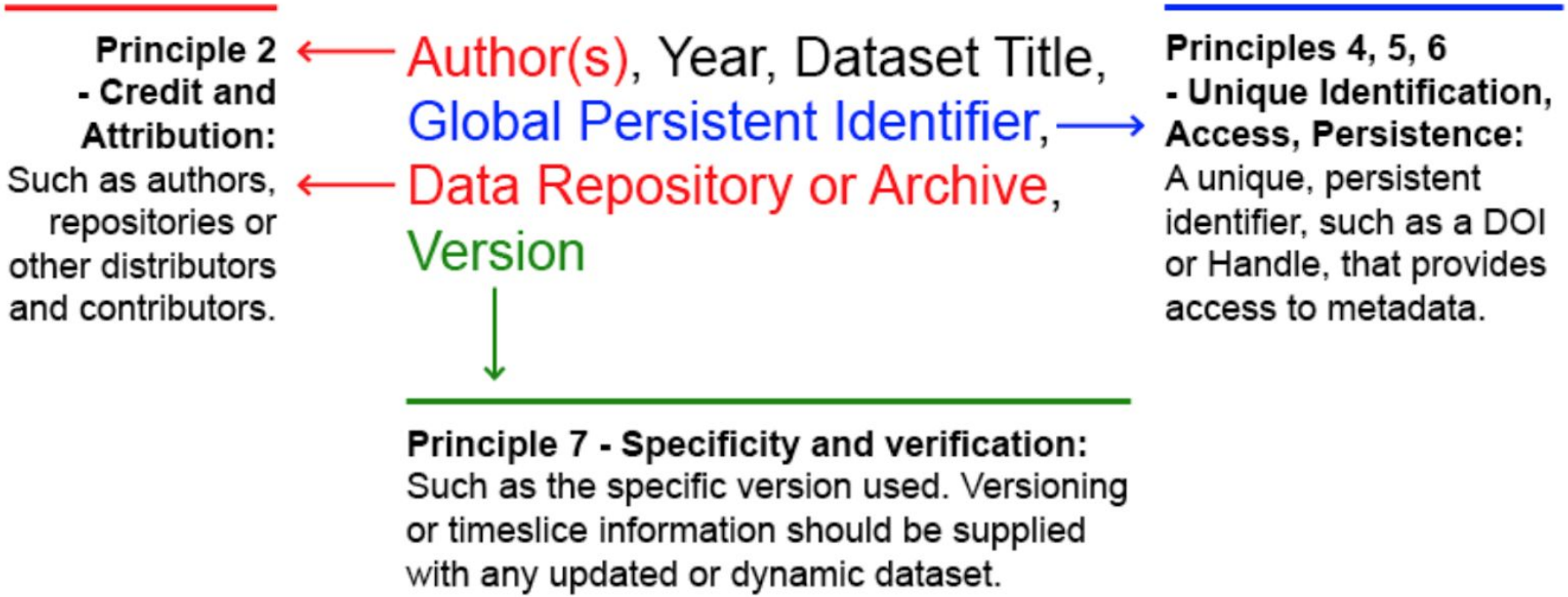


Fig. 1 Example of a data citation based on the Joint Declaration of Data Citation Principles (2014).

Data - Citations

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

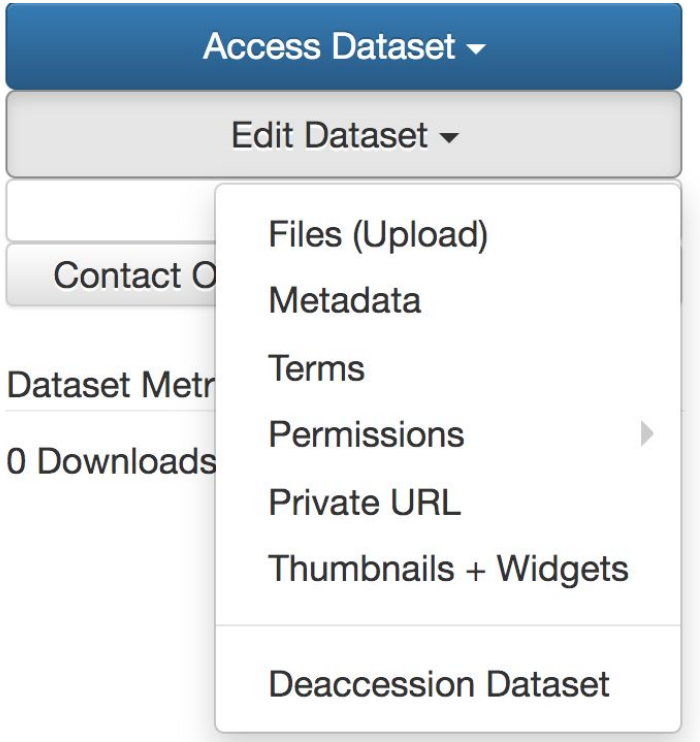
Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L. (2012) The influence of head motion on intrinsic connectivity MRI. NeuroImage , 59(1): 431-438: Link to article

Holmes, A.J., Lee, P.H., Hollinshead, M., Bakst, L., Roffman, J.L., Smoller, J.W., Buckner, R.L. (2012) Individual differences in amygdala-prefrontal anatomy link negative affect, impaired social functioning, and polygenetic depression risk. Journal of Neuroscience, 32(50): 18087-18100: Link to article

Related Publication

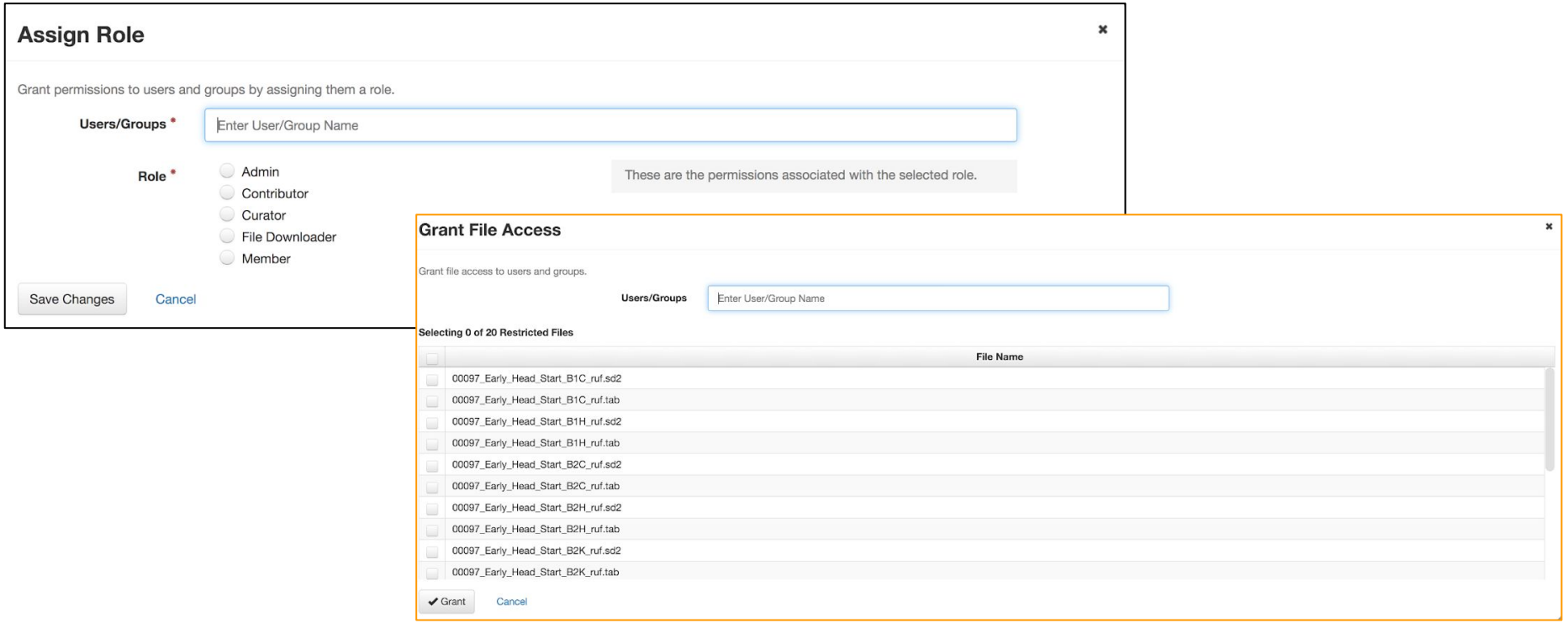
Furnas, Alexander, and Timothy LaPira. [date]. "The People Think What I Think: False Consensus and Unelected Elite Misperception of Public Opinion." *American Journal of Political Science* Forthcoming. http://ajps.org/

Datasets - Editing Options

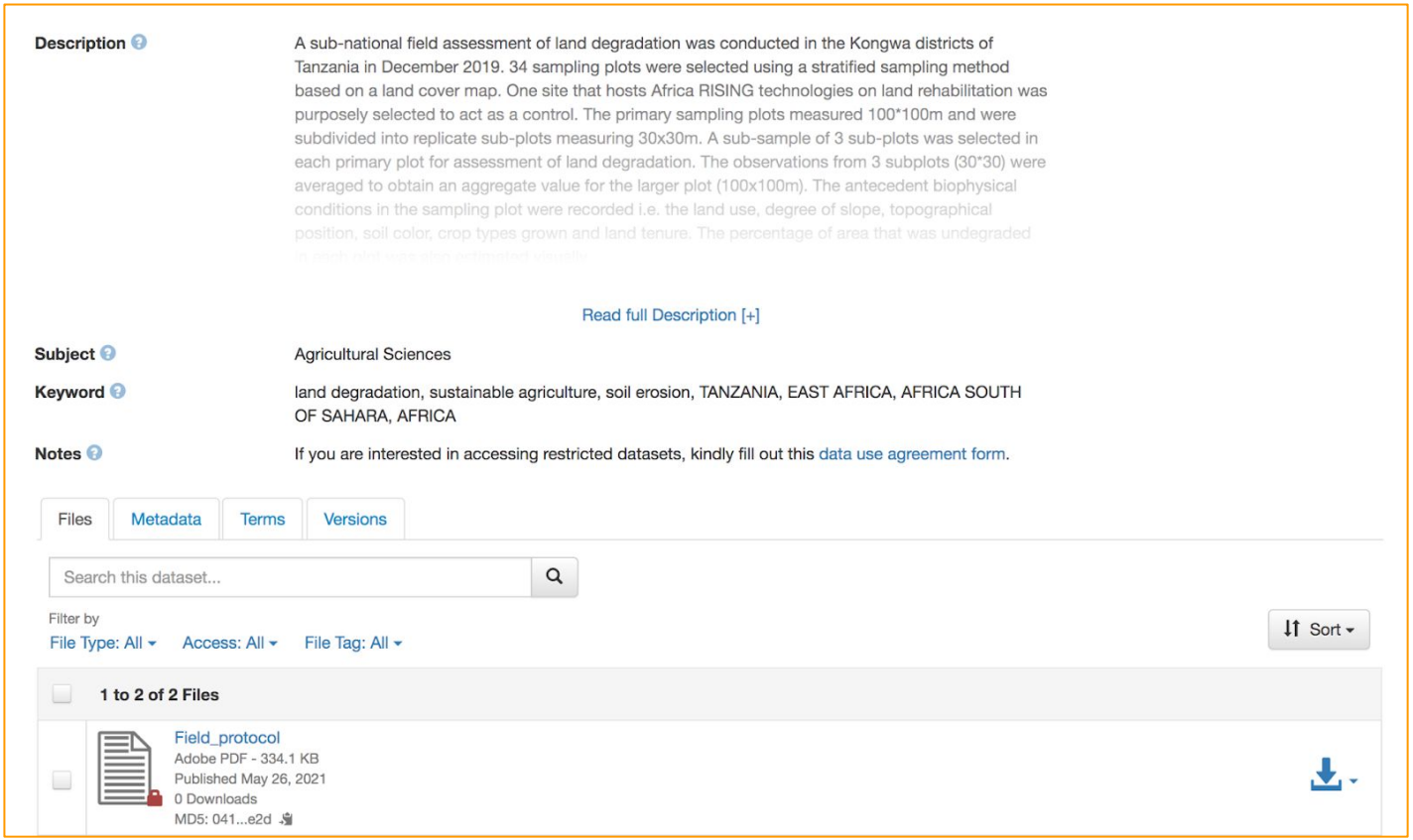


- Datasets can be edited with the options seen here, and versioning is available for dataset and files.
- A **Private URL** can be generated for DRAFT datasets to share a dataset with a coauthor or journal, before publishing.
- **Deaccession** is the only option to "remove' a dataset once it has been published and this leave the TOMBSTONE page of the citation.

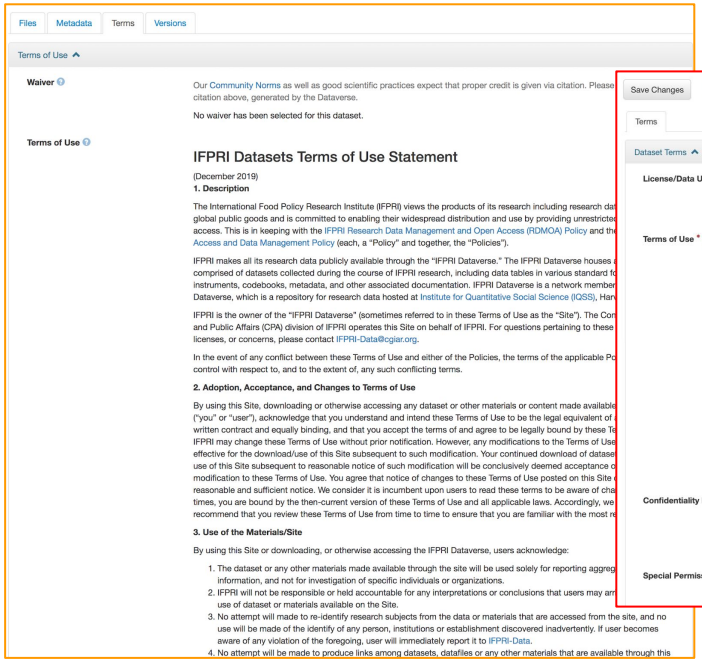
Datasets: Permissions, Workflows and Roles

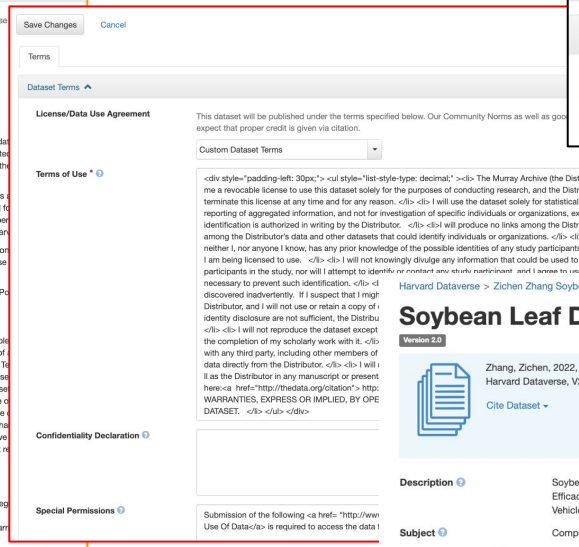


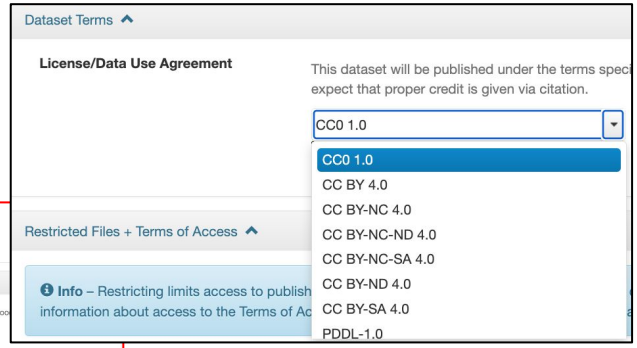
Datasets


Description
A sub-national field assessment of land degradation was conducted in the Kongwa districts of Tanzania in December 2019. 34 sampling plots were selected using a stratified sampling method based on a land cover map. One site that hosts Africa RISING technologies on land rehabilitation was purposely selected to act as a control. The primary sampling plots measured 100*100m and were subdivided into replicate sub-plots measuring 30x30m. A sub-sample of 3 sub-plots was selected in each primary plot for assessment of land degradation. The observations from 3 subplots (30*30) were averaged to obtain an aggregate value for the larger plot (100x100m). The antecedent biophysical conditions in the sampling plot were recorded i.e. the land use, degree of slope, topographical position, soil color, crop types grown and land tenure. The percentage of area that was undegraded
Read full Description [+]
Subject
Agricultural Sciences
Keyword
land degradation, sustainable agriculture, soil erosion, TANZANIA, EAST AFRICA, AFRICA SOUTH OF SAHARA, AFRICA
Notes
If you are interested in accessing restricted datasets, kindly fill out this data use agreement form.
Files
Metadata
Terms
Versions
Search this dataset...
Filter by
File Type: All
Access: All
File Tag: All
Sort
1 to 2 of 2 Files
Field_protocol
Adobe PDF - 334.1 KB
Published May 26, 2021
0 Downloads
MD5: 041...e2d

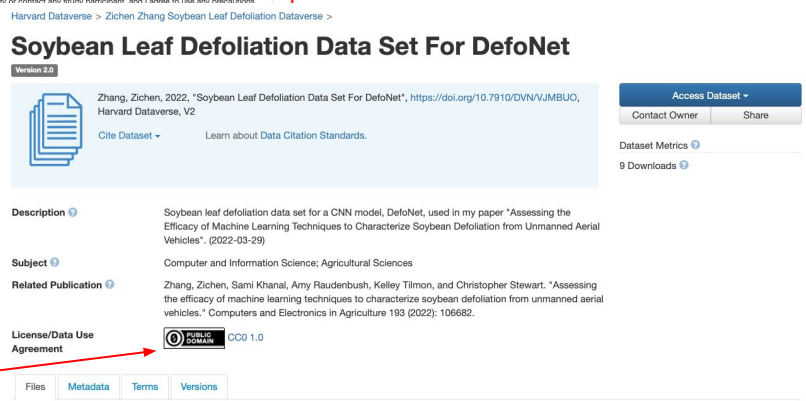
Datasets - Terms







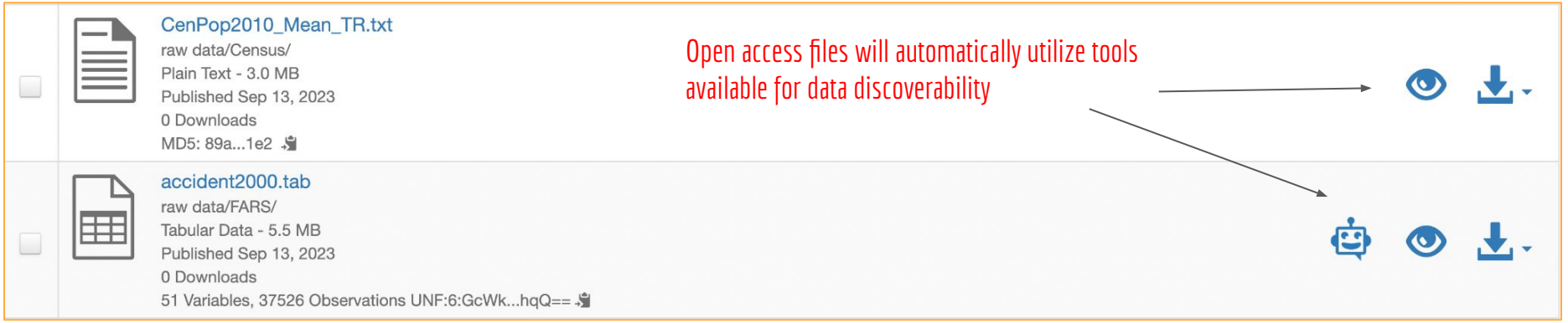
Multiple License support

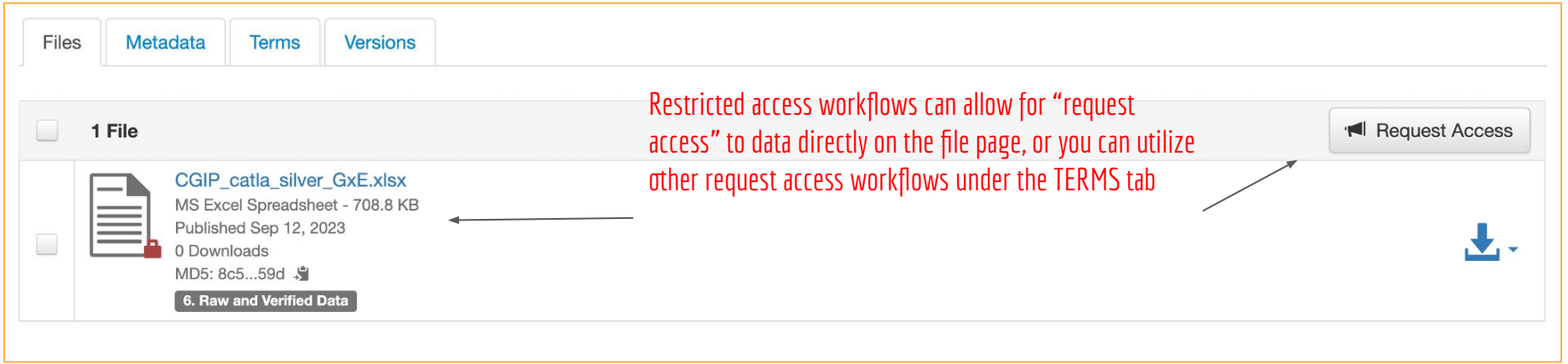


Datasets - Versions

https://dataverse.org/files/dataverseorg/files/metadataprovenance-mercecrosas.pdf

Datasets - Files





Datasets - Tabular Files

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)

See the subsections in the left sidebar for more information on each of these supported formats.

Datasets - Tabular Files

https://guides.dataverse.org/en/latest/user/tabulardataingest/index.html

Datasets - Ask the Data

Tell me what you want to know

What is this data about?

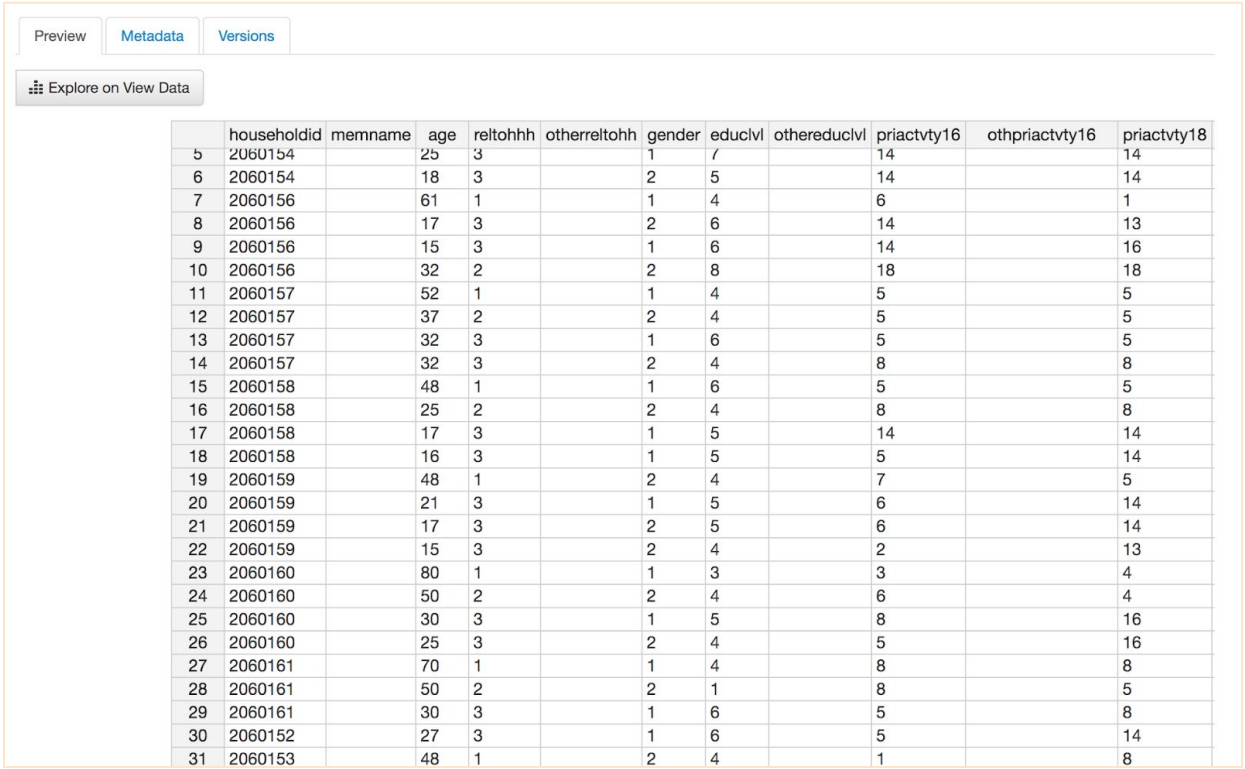
Answer please

This data is about traffic accidents in the US.

STATE	COUNTY	MONTH	DAY	HOUR	MINUTE	VE_FORMS	PERSONS	PEDS	NHS	ROAD_FNC	ROUTE	SP_JUR	HARM_EV	MAN_COLL	REL_JUNC
1	97	1	26	15	30	2	2	0	0	6	4	0	12	2	1
1	101	2	1	23	55	1	1	0	1	2	2	0	42	0	1
1	81	2	2	5	10	1	1	0	0	6	4	0	42	0	1
1	101	1	18	22	14	1	2	1	1	12	2	0	8	0	1
1	101	1	20	13	5	1	1	0	1	11	1	0	34	0	1
1	73	1	31	2	40	2	3	0	1	11	1	0	25	0	1
1	73	1	14	19	0	1	2	0	1	11	1	0	34	0	1
1	73	1	24	19	45	2	3	1	0	6	6	0	8	0	2
1	97	1	23	19	34	2	3	0	0	13	2	0	12	5	2
1	97	1	31	23	33	1	2	0	0	4	3	0	30	0	1

Viewing rows 1 through 10 of 38444

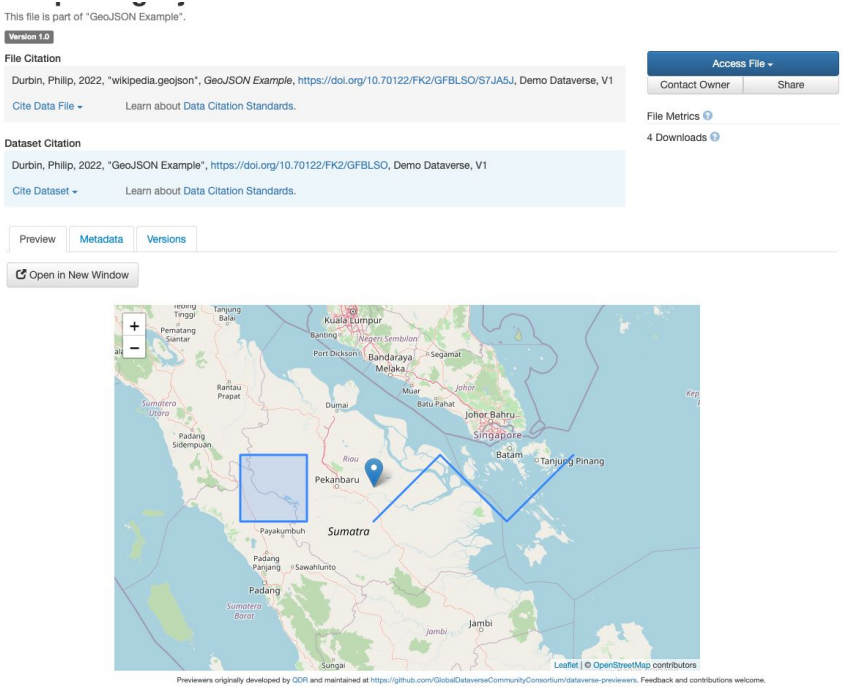
Datasets - File Previewer



	householdid	memname	age	reltohhh	otherreltohh	gender	educlvl	othereduclvl	priactvty16	othpriactvty16	priactvty18
5	2060154		25	3		1	7		14		14
6	2060154		18	3		2	5		14		14
7	2060156		61	1		1	4		6		1
8	2060156		17	3		2	6		14		13
9	2060156		15	3		1	6		14		16
10	2060156		32	2		2	8		18		18
11	2060157		52	1		1	4		5		5
12	2060157		37	2		2	4		5		5
13	2060157		32	3		1	6		5		5
14	2060157		32	3		2	4		8		8
15	2060158		48	1		1	6		5		5
16	2060158		25	2		2	4		8		8
17	2060158		17	3		1	5		14		14
18	2060158		16	3		1	5		5		14
19	2060159		48	1		2	4		7		5
20	2060159		21	3		1	5		6		14
21	2060159		17	3		2	5		6		14
22	2060159		15	3		2	4		2		13
23	2060160		80	1		1	3		3		4
24	2060160		50	2		2	4		6		4
25	2060160		30	3		1	5		8		16
26	2060160		25	3		2	4		5		16
27	2060161		70	1		1	4		8		8
28	2060161		50	2		2	1		8		5
29	2060161		30	3		1	6		5		8
30	2060152		27	3		1	6		5		14
31	2060153		48	1		2	4		1		8

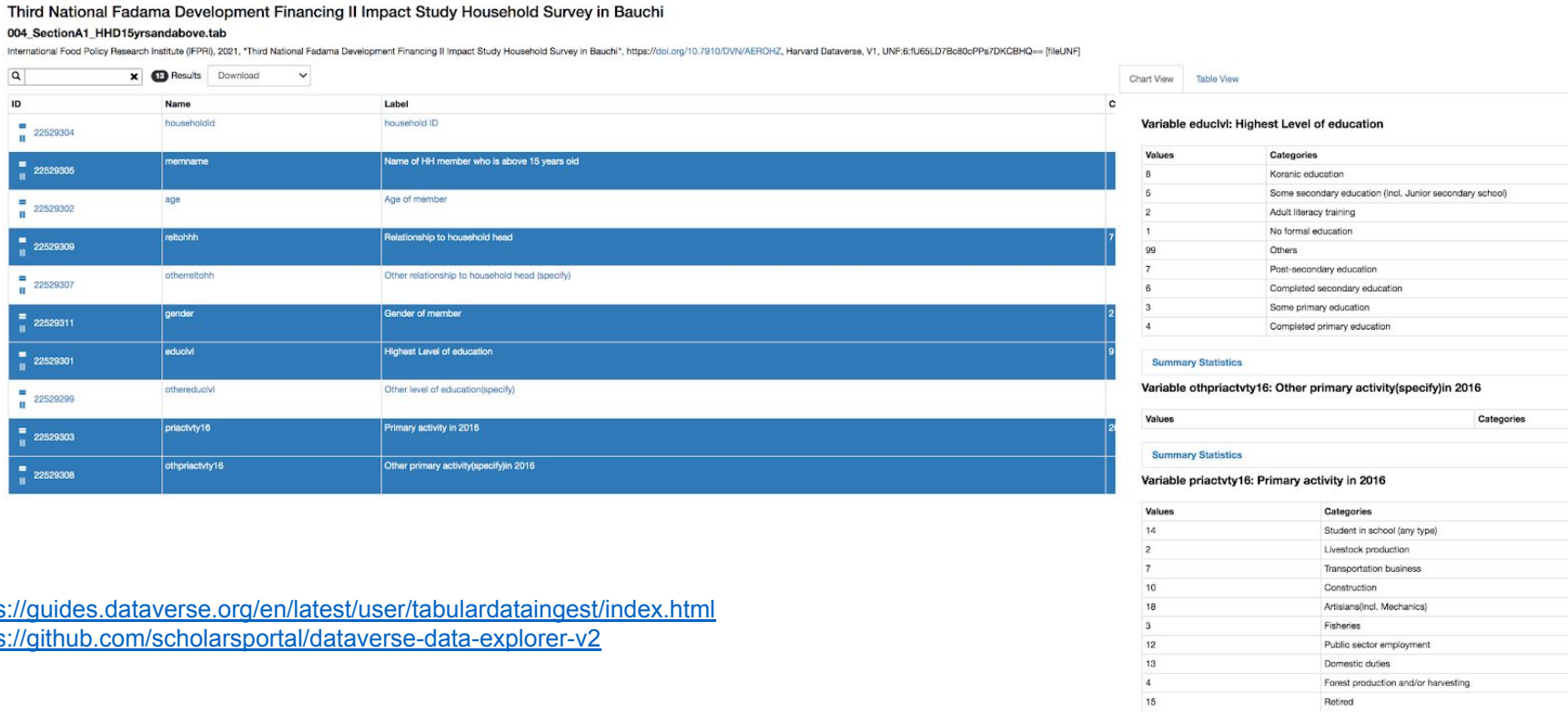
https://github.com/gdcc/dataverse-previewers

Datasets - GeoJSON Previewer



https://github.com/GlobalDataverseCommunityConsortium/dataverse-previewers

Datasets - Tabular Files/Data Explorer



Third National Fadama Development Financing II Impact Study Household Survey in Bauchi

004_SectionA1_HHD15yrsandabove.tab

International Food Policy Research Institute (IFPRI), 2021, "Third National Fadama Development Financing II Impact Study Household Survey in Bauchi", https://doi.org/10.7910/DVN/AEROHZ, Harvard Dataverse, V1, UNF:6:fU65LD7Bc80cPPs7DKCBHQ== [fileUNF]

13 Results | Download | Chart View | Table View

ID	Name	Label	C
22529304	householdid	household ID	
22529305	memname	Name of HH member who is above 15 years old	
22529302	age	Age of member	
22529309	reltohhh	Relationship to household head	7
22529307	otherreltohh	Other relationship to household head (specify)	
22529311	gender	Gender of member	2
22529301	educlvl	Highest Level of education	9
22529299	othereduclvl	Other level of education(specify)	
22529303	priactvty16	Primary activity in 2016	2
22529308	othpriactvty16	Other primary activity(specify)in 2016	

Variable educlvl: Highest Level of education

Values	Categories
8	Koranic education
5	Some secondary education (incl. Junior secondary school)
2	Adult literacy training
1	No formal education
99	Others
7	Post-secondary education
6	Completed secondary education
3	Some primary education
4	Completed primary education

Summary Statistics

Variable othpriactvty16: Other primary activity(specify)in 2016

Values	Categories

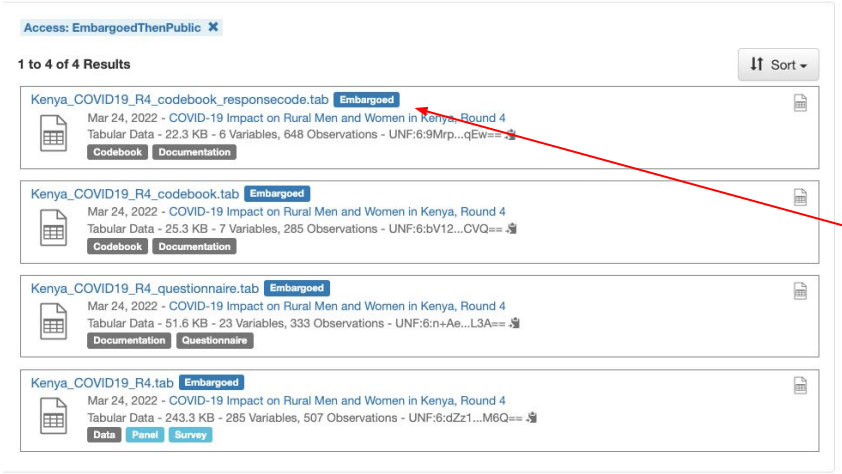
Summary Statistics

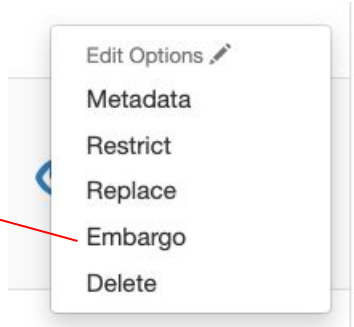
Variable priactvty16: Primary activity in 2016

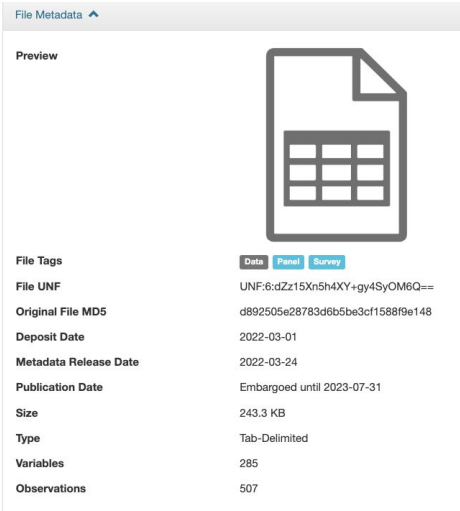
Values	Categories
14	Student in school (any type)
2	Livestock production
7	Transportation business
10	Construction
18	Artisians(Incl. Mechanics)
3	Fisheries
12	Public sector employment
13	Domestic duties
4	Forest production and/or harvesting
15	Retired

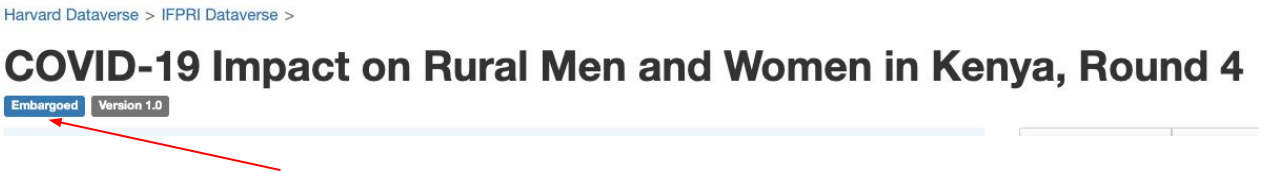
https://guides.dataverse.org/en/latest/user/tabulardataingest/index.html
https://github.com/scholarsportal/dataverse-data-explorer-v2

Datasets - File Level Embargo









Datasets - Deaccession

Replication Data for: Democratic Subversion: Elite Cooptation and Opposition Fragmentation **Deaccessioned**

Jun 8, 2021

Arriola, Leonardo R.; DeVaro, Jed; Meng, Anne, 2021, "Replication Data for: Democratic Subversion: Elite Cooptation and Opposition Fragmentation", https://doi.org/10.7910/DVN/GHDKSY, Harvard Dataverse, V1, DEACCESSIONED VERSION, UNF:6:Nk/xhGPVreurGI55KFcKrQ== [fileUNF]

The dataset has been transferred to another repository

Considerations in Sharing Sensitive Data in Harvard Dataverse Repository

Example of de-identified human subjects data shared in Harvard Dataverse

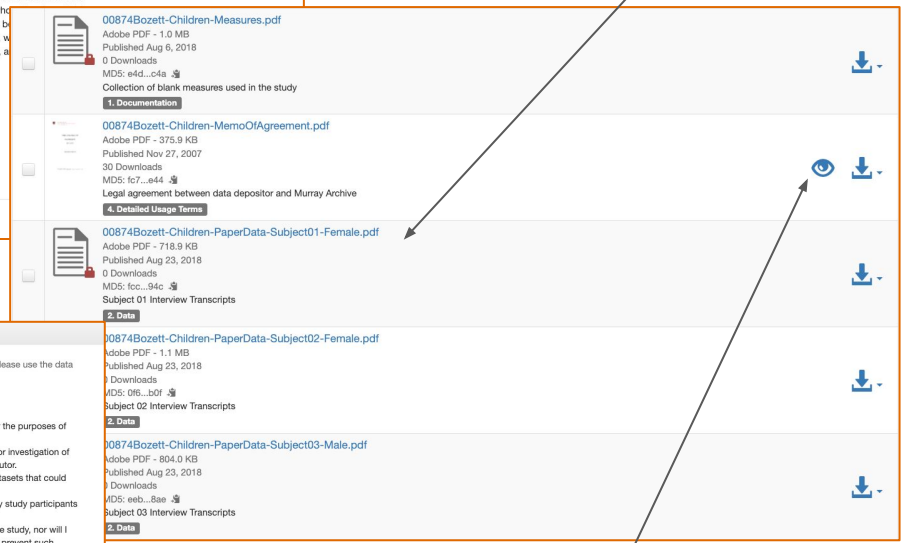
Terms of use and application form for data requests

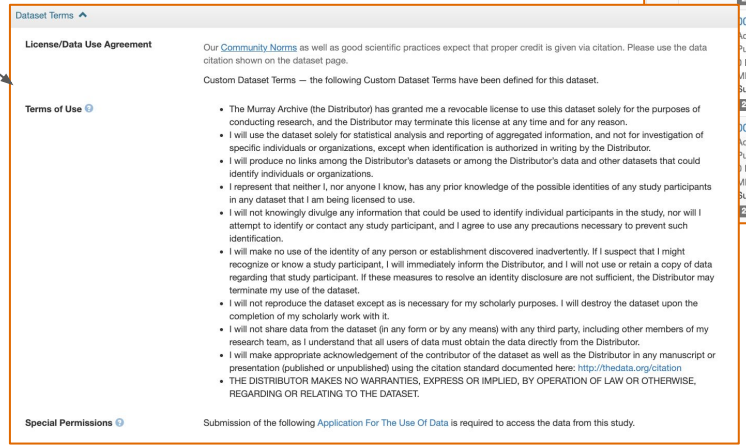
*Sensitive data support is not yet available in Harvard Dataverse Repository. All data must be deidentified!

Utilize the "guestbook" feature

Restricted, de-identified data

Documentation unrestricted for access, to give users information on the project



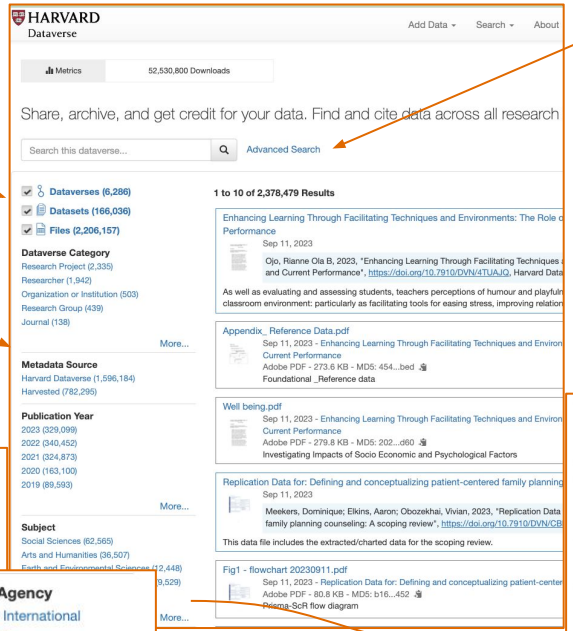


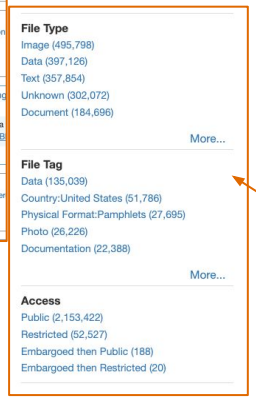
Search and Browse Functionality

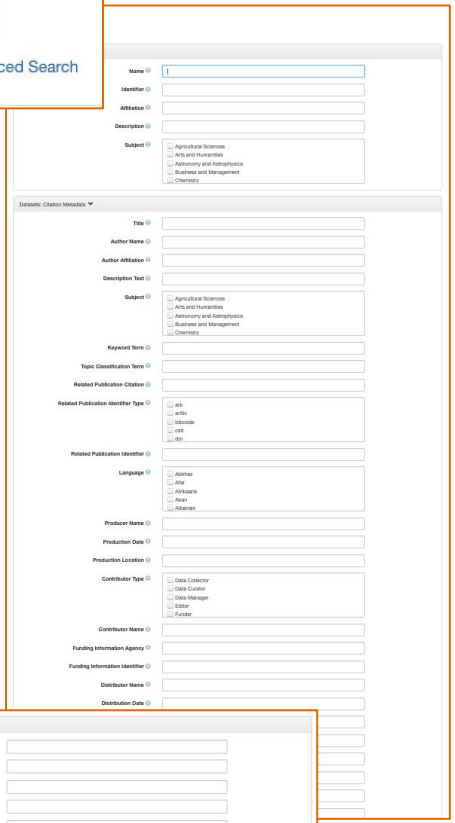
Search & browse

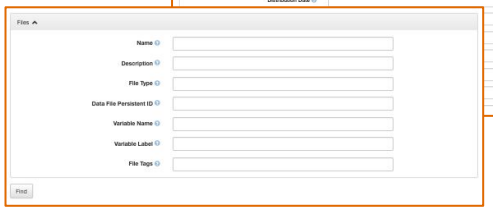
Search:

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









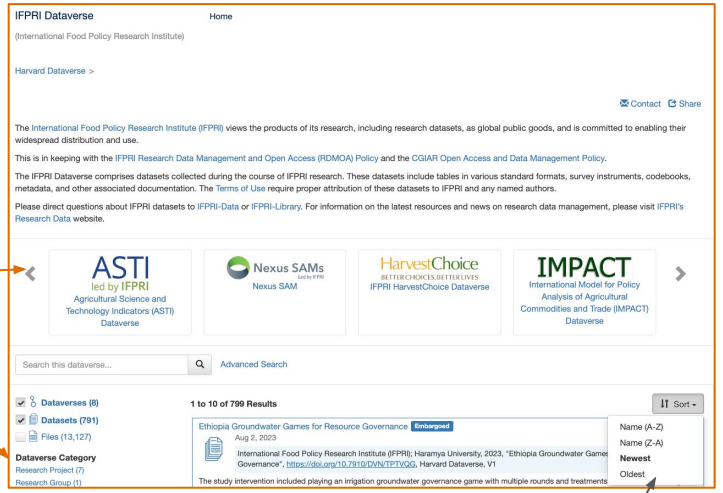
Search & browse

Dataset details:

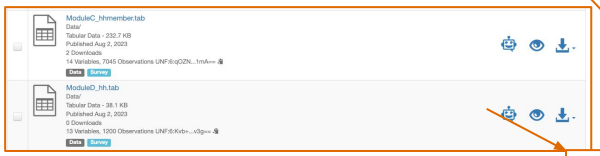
- Detailed dataset page
- Metadata
- Description
- Authors
- Citation
- Download options
- Documentation, readme, code…
- Multiple format download options

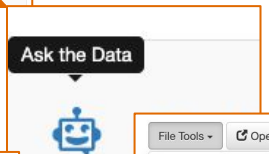
Browse:

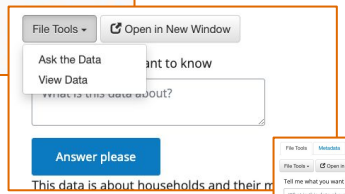
- Categories/Subjects
- Featured Datasets/collections

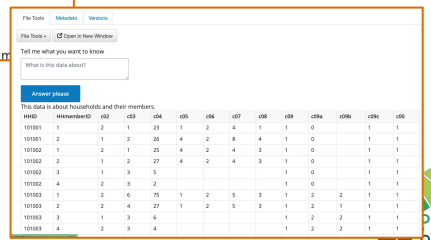


Browse: Recently added



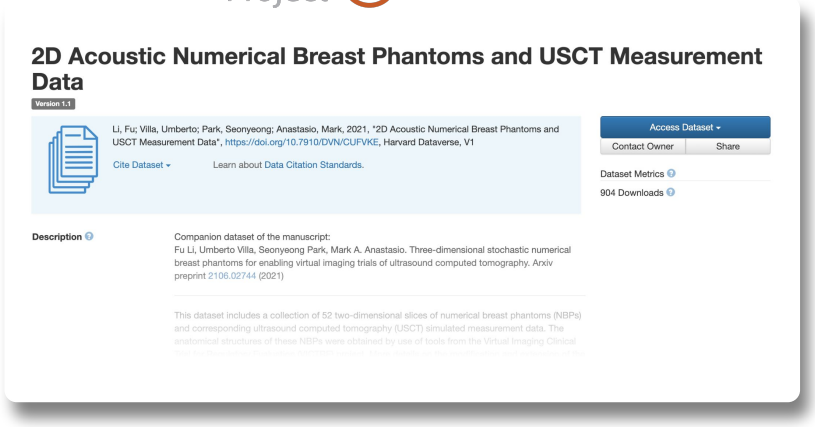




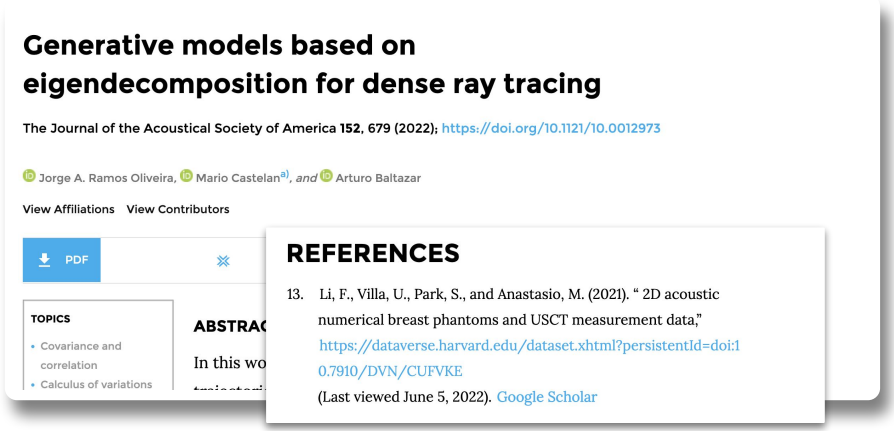


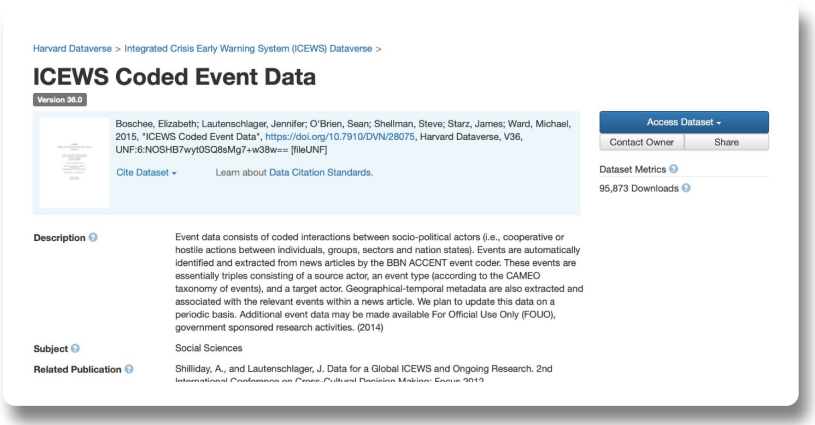
Data Re-Use

Re-use of Datasets

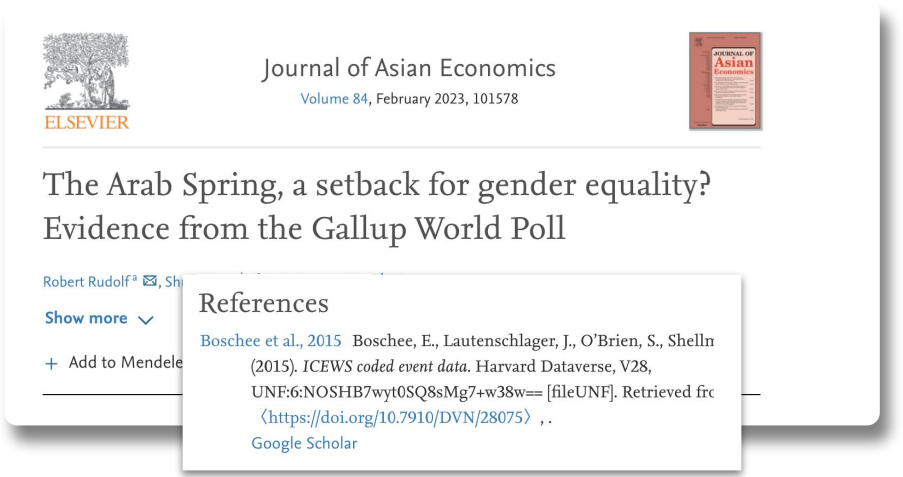


cited by



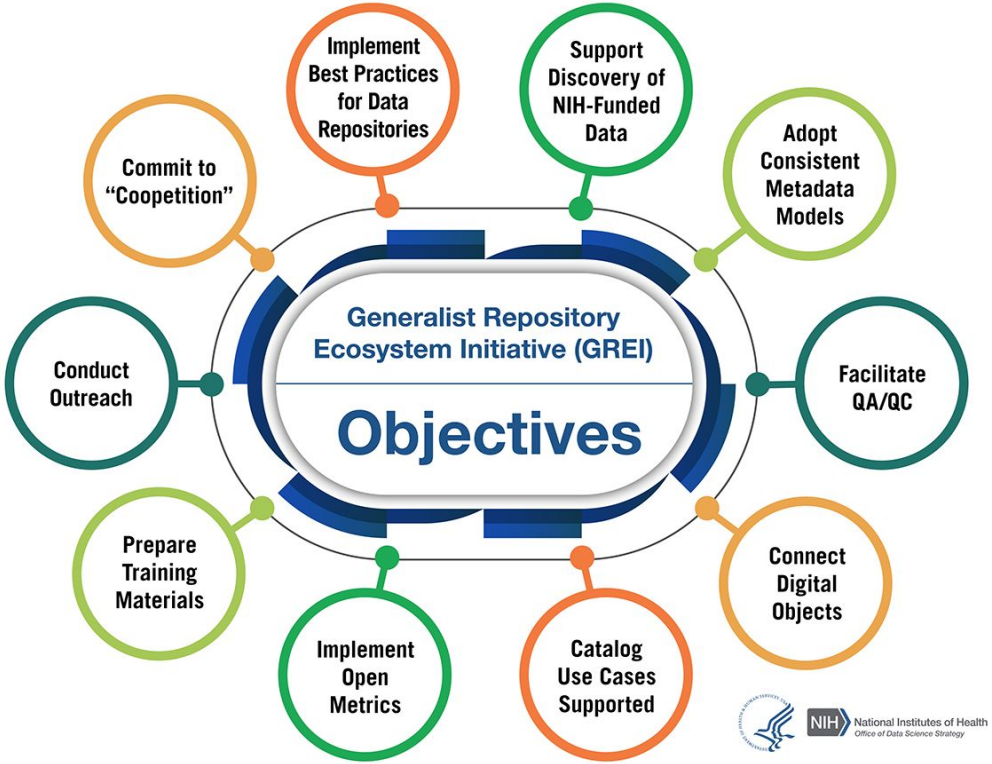


cited by



NIH-GREI

NIH GREI



NIH · The Dataverse Project · DRYAD · figshare · OSF · MENDELEY DATA · zenodo · Vivli

https://datascience.nih.gov/data-ecosystem/generalist-repository-ecosystem-initiative

HARVARD Dataverse

NIH GREI

IQSS

Current Features

Installations

Collections

Datasets

Files

NIH OTA Adds…

Datasets

- New Workflows for **Large Data Support** (Globus, demo)
- **Remote** Data Support (TRSA)
- Flexible **Biomedical Metadata** Support through External Vocabularies and Data Dictionaries (UMLS, CEDAR, MeSH)

Files

- Additional Metadata for **Code files** (Codemeta)
- Support for **Replication Packages** through Workflows and Containers
- **Sensitive Data Support** through Differential Privacy (OpenDP, DataTags, PSIprivacy)
- Encryption

- Usage Metrics
- UX/UI Enhancements
- New APIs
- Interoperability
- Metadata Harvesting
- Curation Services
- Training and Outreach

Summary

- **The Dataverse Project** facilitates the creation of digital repositories to ensure that the data are high quality, properly documented, organized, and accessible.
- **The Harvard Dataverse Repository** is open to the world wide research community for data sharing (free up to 1 terabyte of data, and 2.5GB per file limitation). See our policies page
- Metadata is an essential component of the Dataverse Project
- **Best Practices** in a Dataverse repository: Collection creation and customization, dataset creation and file uploads, publishing workflows, and QA to ensure FAIR

NIH Data Management and Sharing Policy: Harvard Dataverse

2023

Data Sharing in Context

"NIH is aware that not just technological advancement, but behavioral change is also necessary to advance data science goals"

Science, Digital; Goodey, Gregory; Hahnel, Mark; Zhou, Yuanchun; Jiang, Lulu; Chandramouliswaran, Ishwar; et al. (2022): The State of Open Data 2022. Digital Science. Report. https://doi.org/10.6084/m9.figshare.21276984.v5

NIH Data Management and Sharing Policy

Submission of a Data Management and Sharing Plan (DMSP)

Proposed repository to be used consistent with NIH guidance

Compliance with the awardee's plan as approved by NIH ICO

Elements of a Data Management & Sharing Plan

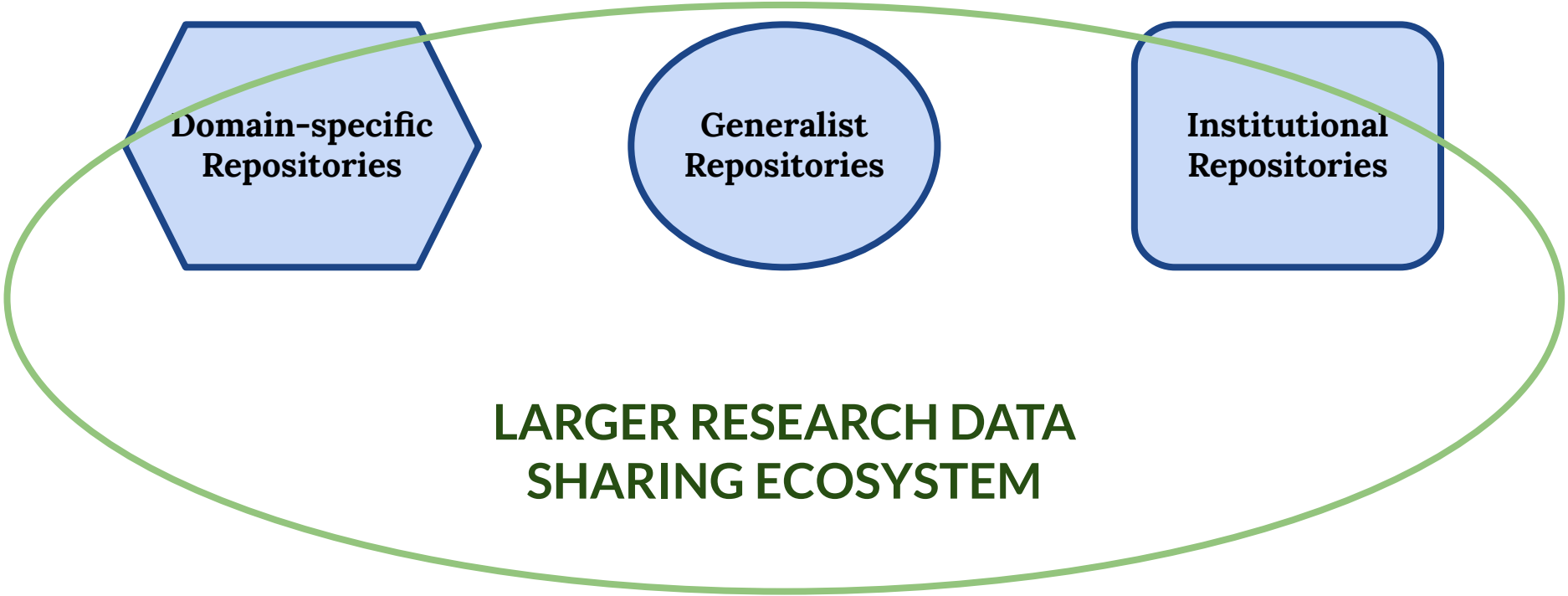
1. **Data Types**: Data to be preserved and shared
2. **Related Tools, Software, Code**: Tools and software needed to access and manipulate data
3. **Common Data Standards**: Standards to be applied to scientific data & metadata
4. **Data Preservation, Access, Timelines**: Repository to be used, persistent unique identifiers, and when/how long data will be available
5. **Access, Distribution, Reuse Considerations**: Factors affecting data access, distribution, or reuse related to informed consent or privacy & confidentiality protections
6. **Oversight of Data Management**: How Plan compliance will be monitored & managed and by whom

NIH Data Repository Desired Characteristics

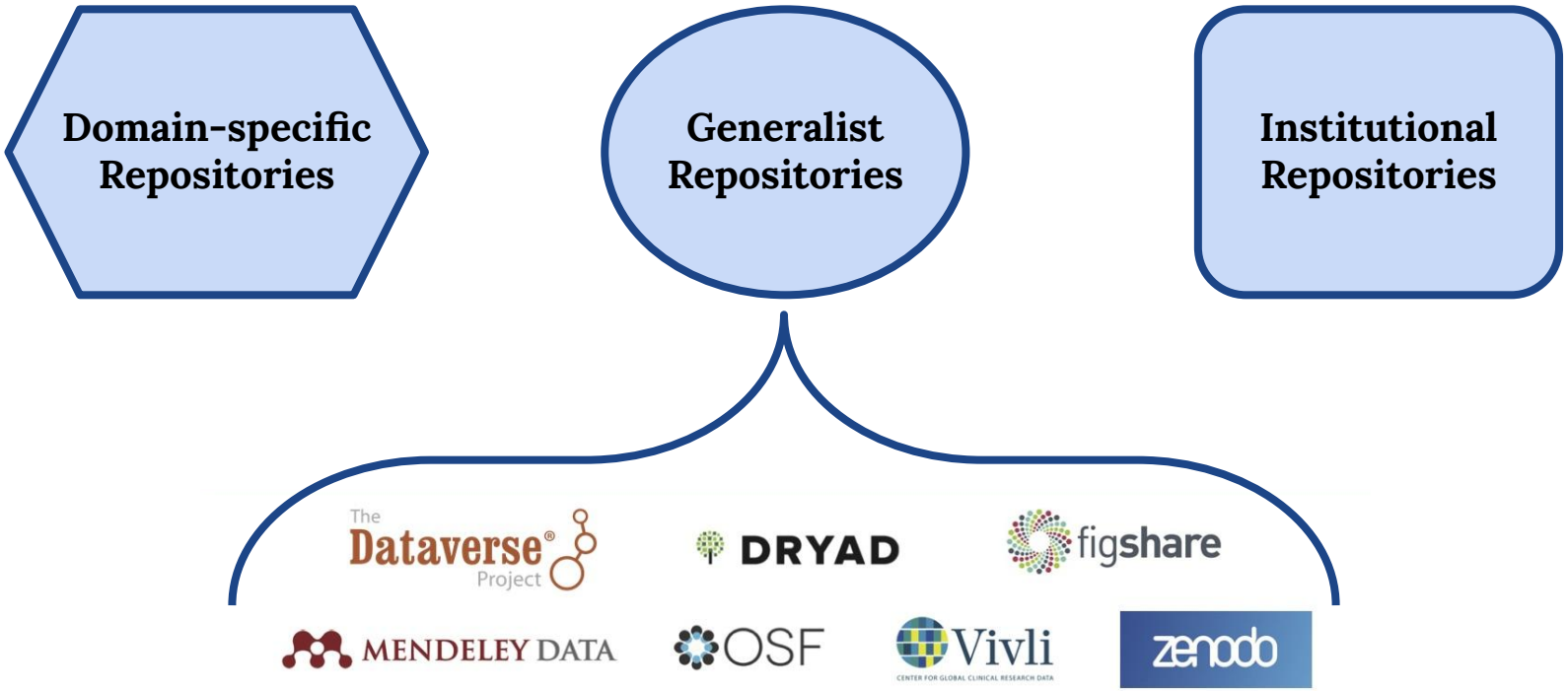
NIH promotes the use of established data repositories because deposit in a quality data repository generally improves the FAIRness (Findable, Accessible, Interoperable, and Re-usable) of the data.

- Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html
- Open Domain-Specific Data Sharing Repositories: https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
- Generalist Repositories: https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html

NIH Generalist Repository Ecosystem Initiative



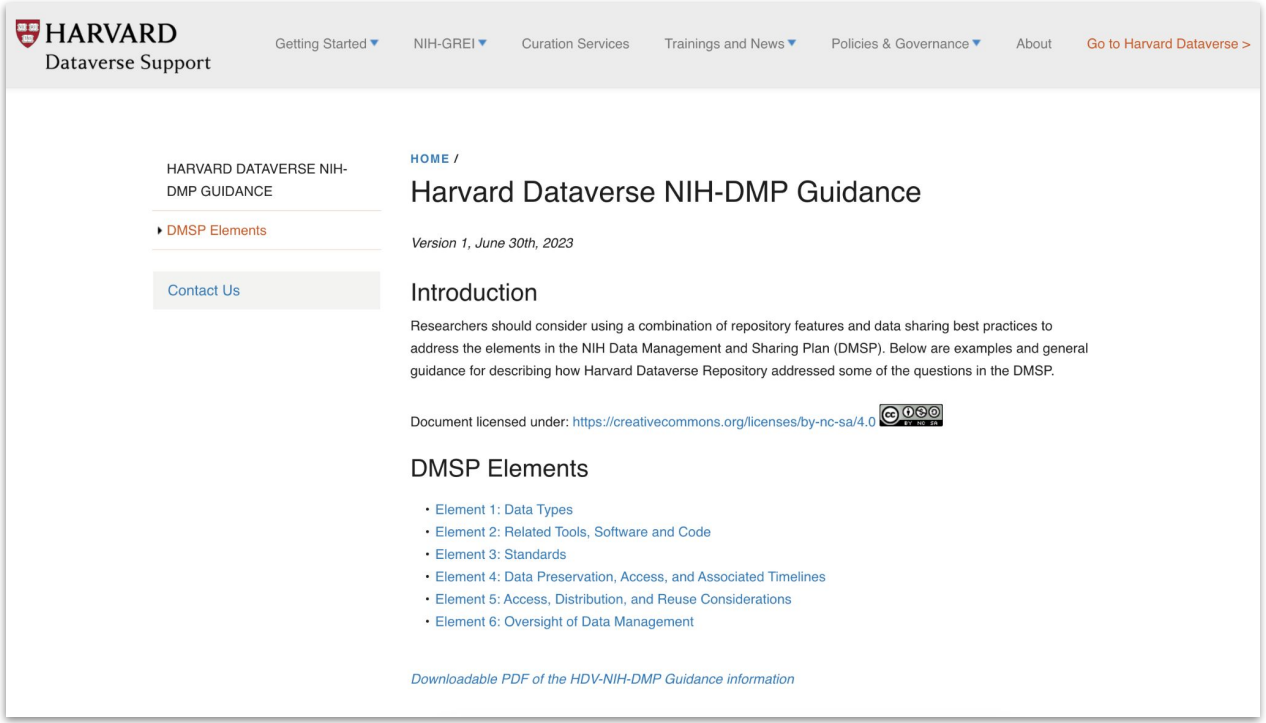
NIH Generalist Repository Ecosystem Initiative




The
Dataverse
Project

Harvard Dataverse NIH-DMP Guidance

https://support.dataverse.harvard.edu/harvard-dataverse-nih-dmp-guidance



Element 1: Data Types

Harvard Dataverse Supports:

- All data types
  - Any data format regardless of file extension
  - Data formats for preservation: https://www.openaire.eu/data-formats-preservation-guide
- Tabular data
  - For SPSS, Stata, RData, and CSV files, the ingest process extracts the data content from the user's files and archives it in an application-neutral, easily-readable format.
  - Note: enhanced ingest for Excel files is not available.

Sample DMP Text: For tabular data: PI will ensure that tabular data files are ingested properly by the Dataverse software. To best support archival preservation, Harvard Dataverse stores the raw data content extracted from successfully ingested tabular data files in plain text, TAB-delimited files. The metadata information that describes this content is stored separately, in a relational database, so that it can be accessed efficiently by the application. For the purposes of archival preservation it can be exported, in plain text XML files, using a standardized, open DDI Codebook format.

Element 2: Related Tools, Software and Code

Harvard Dataverse Recommends:

- Publish analysis code
  - Best practice guidelines encourage the publication of analysis code, configuration, and execution files to be published with data and other input files used to generate or compute research results. This is especially important for replication datasets.
- Use Harvard Dataverse tools
  - Github connector; GeoJSON; Data Explorer; Supported File Previewers; Binder
- Follow best practices for research code
  - Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to the research data deposited in Dataverse repositories.
- Publish research & scientific software
  - Generalized or multi-purpose software that can be used across different datasets may be published as a separate research object

Sample Text: To support data access and reuse all code files generated as a product of this research will be published in Harvard Dataverse as part of the replication dataset that contains data and documentation files.

Original software developed to analyze imaging data (e.g.) will be published in Harvard Dataverse.

Element 3: Standards

Harvard Dataverse Recommends:

- Apply relevant metadata
    - Choose metadata fields from Harvard Dataverse Supported Metadata Blocks
- Apply additional NIH project metadata
    - Life Sciences Metadata
    - CodeMeta (forthcoming)
- Establish bidirectional link between dataset and related publications
    - Cite related publication/s using related publication field
    - Cite dataset DOI in your publications and reference lists

Sample Text: Harvard Dataverse is committed to using standard-compliant metadata to ensure that metadata can be mapped easily to standard metadata schemas and be exported into JSON format (XML for tabular file metadata) for preservation and interoperability.

The Life Sciences metadata block will be applied to all datasets resulting from this research.
Metadata, configurations, and parameters related specifically to software applications that cannot be described in Dataverse will be recorded in documentation files and will be published along with data.

Element 4: Data Preservation, Access, and Associated Timelines

Harvard Dataverse Policies:

- Repository hosting, backups, and preservation
  - Harvard Library Technical Services (LTS), in collaboration with Harvard University Information Technology (HUIT) and the Institute for Quantitative Social Science (IQSS), hosts the Harvard's Dataverse repository using Amazon Web Services and S3, and maintains a full backup of all data and directories using Amazon Glacier.
  - FAS Research Computing at Harvard University keeps a backup of all Harvard Dataverse data and directories. This means that there are always full, recent copies of the Harvard Dataverse repository at multiple locations.
- Data files access
  - Access to data files is open by default but can be mediated through a variety of features including: Embargos; Restricted file access; Guestbooks; and User permissions.
- Metadata access
  - HDV policy is that published dataset and file level metadata are always openly accessible.

Sample Text: Scientific data from the research/study/project will be published in Harvard Dataverse, a nonprofit, generalist data repository built with the open source Dataverse Project software and affiliated with Harvard University's Institute for Quantitative Social Science and Harvard Library.
The Harvard Dataverse Repository is a free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, to share, archive, cite, access, and explore research data.
Data published in Harvard Dataverse are described according to standard metadata schemas, exposed to the web via OAI-PMH and schema.org, assigned DOIs automatically, and backed up regularly to Amazon Glacier and Harvard's Faculty of Arts and Sciences Research Computing servers.
Data published in Harvard Dataverse will be accessible to the research community for the long-term.

Element 5: Access, Distribution, and Reuse Considerations

Harvard Dataverse Policies:

- Harvard Dataverse does not support sensitive and personal identifiable data deposits
  - The depositor is responsible for removing all sensitive and personal identifiers for all files uploaded to Harvard repository.
  - Harvard Policy on sensitive data: https://support.dataverse.harvard.edu/harvard-dataverse-general-terms-use
- Access and embargo features, custom terms of access and use available
  - Support includes: file restricted access control via the "request access" feature; file level embargo; depositor-defined terms of use and access
  - Depositor defined access control (example): Early Head Start Research and Evaluation Project, 1996 - 2001: Childcare/Teacher Data

Sample Text: Data will be published openly under a CC-0 public domain waiver.

Restricted files will be accessible via the "request access" workflow

Additional restricted content in Harvard Dataverse may be accessible via custom terms of access and reuse.

Additional, sensitive data, not published in Harvard Dataverse may be accessible via a formal Data Use Agreement negotiated between research institutions.

Data will be embargoed for 50 years to protect survey participants from potential reidentification.

Element 6: Oversight of Data Management

Harvard Dataverse Policies:

- Role of Harvard Dataverse
  - Harvard Dataverse is not a responsible party that may be named in a Data Management and Sharing Plan. Users retain ownership and control of all deposited data.
- Apply relevant metadata
  - Include depositor information in dataset metadata in "Depositor" field.
- Curation Services can be included in proposed budget
  - The Harvard Dataverse Repository provides fee-based curation services to researchers around the world who are depositing data into The Harvard Dataverse Repository. Research data replication datasets, data for related publications, and all file types and domains are welcomed in the Harvard Dataverse Repository. Our curation services will ensure that your datasets are discoverable, accessible, interoperable, and reusable (FAIR).

Sample Text: FirstName LastName, Job Title, will be responsible for depositing all scientific data resulting from this project in a Harvard Dataverse collection.

We will utilize custom curation services, outlined in our budget justification, to expertly curate datasets to ensure that they are discoverable, accessible, interoperable, and reusable (FAIR).

Budgeting for Harvard Dataverse

- **Curating data**, such as engaging a generalist repository for curation support services
- **Developing supporting documentation**, such as research personnel effort needed to create and retain appropriate documentation beyond the effort needed to simply generate or collect the data
- **Formatting data according to accepted community standards**, or for transmission to and storage at a selected repository for long-term preservation and access, such as support services that may be offered by the repository or engaging a service center for support with data deposits
- **De-identifying data**, such as research personnel effort to systematically remove all identifiers from a data set prior to sharing or engaging a third-party vendor to certify that a data set has been fully de-identified
- **Preparing metadata** to foster discoverability, interpretation, and reuse, such as research personnel effort needed to document appropriate metadata beyond the effort needed to simply generate or collect the data
- **Unique, project-specific information resources** necessary to provide local management and preservation, such as charges from an external vendor for cloud storage prior to deposit into an established repository or costs to use a third-party electronic lab notebook
- **Preserving and sharing data through established repositories**, such as data deposit fees
- **Large data** support requires additional charge for storage/downloads.

Harvard Dataverse Curation Services

https://support.dataverse.harvard.edu/curation-services

Service Name	Description	Service Components	Costs	
Data Curation Services			Harvard	Non-Harvard
Free consultation and assessment	Triage level + Office hours Follow-up with recipients of support letters and consultation clients	Demonstrations of software Consult on projects size and scope and fit for Harvard Dataverse	Up to 3 hours	Up to 2 hours
Extended consultation services		• Demonstrations of software • Deidentification consult • Dataverse organization consult • File organization consult • Replication verification consult	$100/hour	$200/hour
Dataverse set-up services and dataset and data file ingest	Establish data curation infrastructure and organization enabling data owners to self-curate their data.	• Extended Consultation • Dataverse creation and customization • Dataset creation • Data file ingest • Metadata enrichment of dataset and data files • Documentation for use of final product	$2000 per "collection" plus $100 minimum per dataset	$4000 per "collection" plus $200 minimum per dataset
Ongoing dataverse administration and curation services	Ongoing maintenance of Dataverse and datasets Continuous updates of dataverse and datasets as needed to reflect software upgrades and sharing standards	• Extended consultation services • Dataverse set-up services and dataset and data file ingest • Maintenance of dataverses and datasets within Harvard Dataverse	$10,000/year plus $100 minimum per dataset	$20,000/year plus $200 minimum per datasets
Custom services for existing dataverse	One time curation services This includes: • file-upload • metadata enhancement (bi-directional linking), • linking affiliated datasets • Documentation • Digitization of audio/paper/content for sharing purposes		$100/hour	$200/hour

Thank You!

The Dataverse Project
https://dataverse.org/

The Harvard Dataverse repository
https://dataverse.harvard.edu/

Integrating Dataverse and DSpace
https://osf.io/72w4m

The Dataverse Guide
https://guides.dataverse.org/en/latest/

Dataverse on Github
https://github.com/IQSS/dataverse

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

The Dataverse® Project

IQSS
The Institute for Quantitative Social Science

HARVARD LIBRARY

HARVARD UNIVERSITY
Information Technology

Sonia Maria Barbosa
Manager, Harvard Dataverse