The Use Experience in a Dataverse Supported Repository

Sonia Barbosa

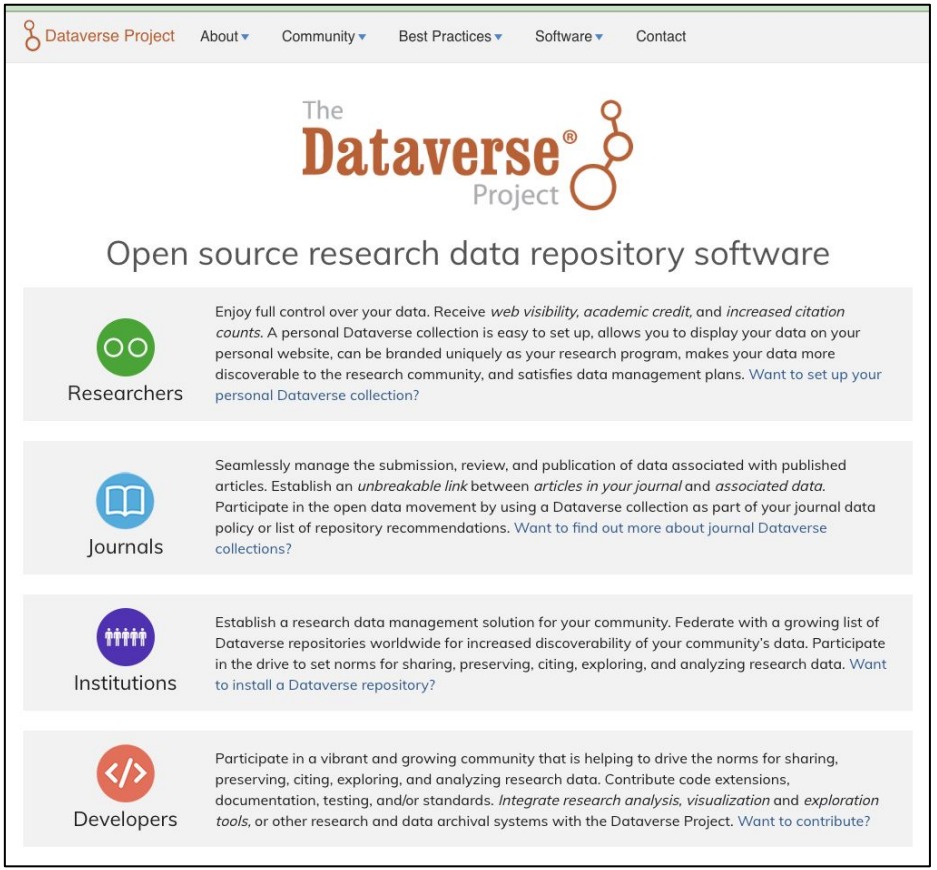
National Yang Ming Chiao Tung University, February 2024

Sonia Maria Barbosa
Manager, Harvard Dataverse

Agenda

AGENDA: The User Experience

- Community
- Collections
- Datasets
- Files
- Files and External Tools
- Permissions and Access
- Data Exploration
- Search and Browse Functionality



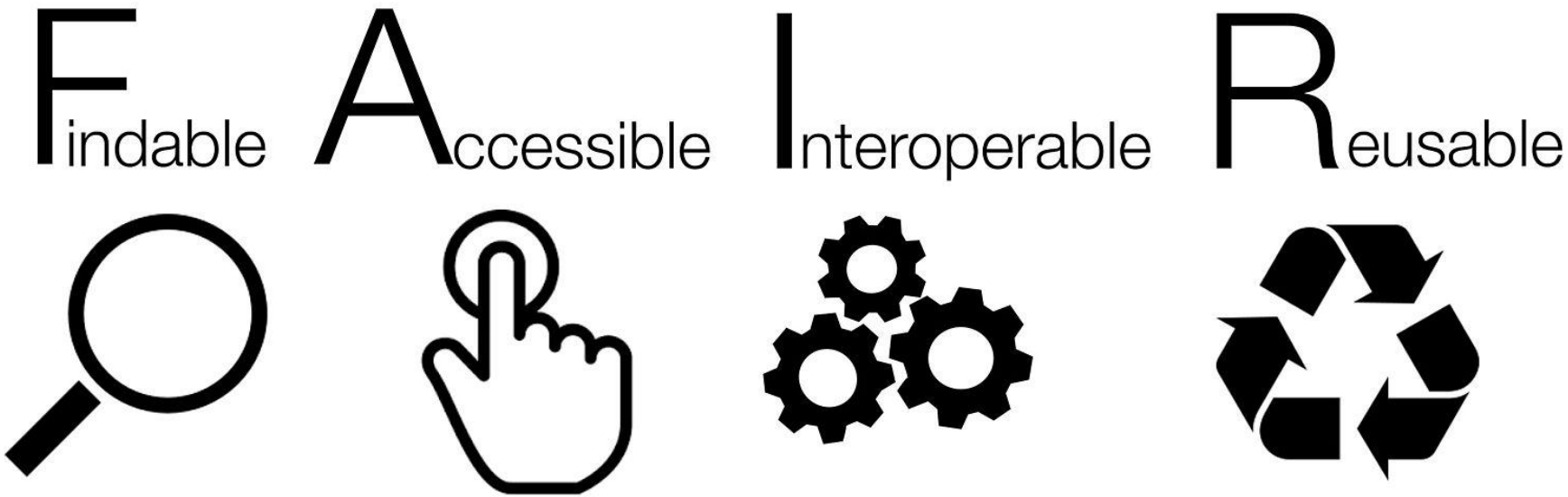
Dataverse Repository Features

- 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 and FAIR

FAIR Principles and Dataverse

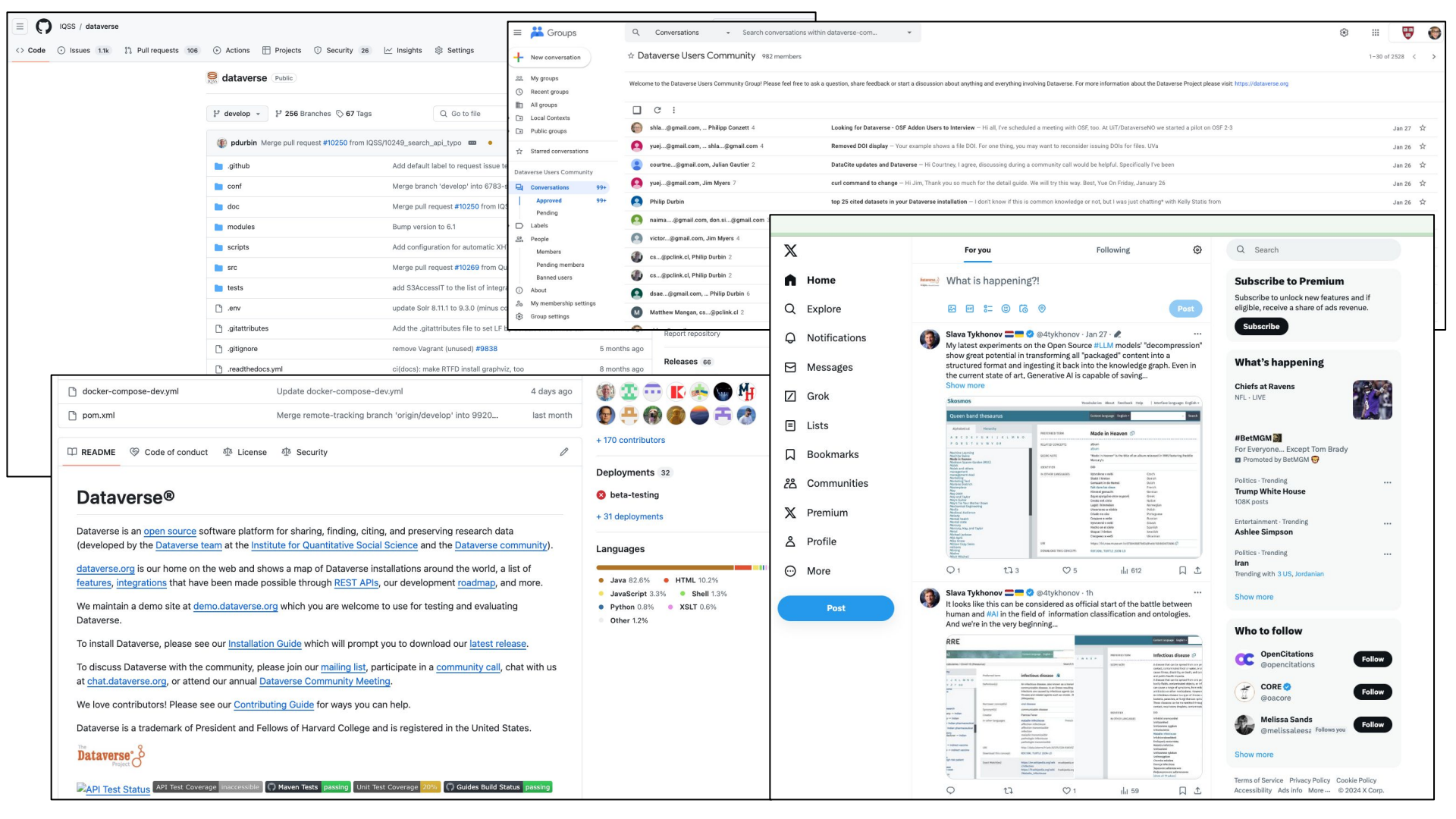


https://en.wikipedia.org/wiki/File:FAIR_data_principles.jpg

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

The User Experience - an Overview


IQSS / dataverse
Code
Issues 1.1k
Pull requests 106
Actions
Projects
Security 26
Insights
Settings
dataverse Public
develop
256 Branches
67 Tags
Go to file
pdurbin Merge pull request #10250 from IQSS/10249_search_api_typo
.github
conf
doc
modules
scripts
src
tests
.env
.gitattributes
.gitignore
.readthedocs.yml
Groups
New conversation
My groups
Recent groups
All groups
Local Contexts
Public groups
Starred conversations
Dataverse Users Community
Conversations
Approved
Pending
Labels
People
Members
Pending members
Banned users
About
My membership settings
Group settings
Conversations
Search conversations within dataverse-com...
Dataverse Users Community 982 members
Welcome to the Dataverse Users Community Group! Please feel free to ask a question, share feedback or start a discussion about anything and everything involving Dataverse. For more information about the Dataverse Project please visit: https://dataverse.org
Looking for Dataverse - OSF Addon Users to Interview
Removed DOI display
DataCite updates and Dataverse
curl command to change
top 25 cited datasets in your Dataverse installation
Report repository
Releases 66
docker-compose-dev.yml
Update docker-compose-dev.yml
4 days ago
pom.xml
Merge remote-tracking branch 'origin/develop' into 9920...
last month
README
Code of conduct
License
Security
Dataverse®
Dataverse is an open source software platform for sharing, finding, citing, and preserving research data (developed by the Dataverse team at the Institute for Quantitative Social Science and the Dataverse community).
dataverse.org is our home on the web and shows a map of Dataverse installations around the world, a list of features, integrations that have been made possible through REST APIs, our development roadmap, and more.
We maintain a demo site at demo.dataverse.org which you are welcome to use for testing and evaluating Dataverse.
To install Dataverse, please see our Installation Guide which will prompt you to download our latest release.
To discuss Dataverse with the community, please join our mailing list, participate in a community call, chat with us at chat.dataverse.org, or attend our annual Dataverse Community Meeting.
We love contributors! Please see our Contributing Guide for ways you can help.
Dataverse is a trademark of President and Fellows of Harvard College and is registered in the United States.
API Test Status
API Test Coverage inaccessible
Maven Tests passing
Unit Test Coverage 20%
Guides Build Status passing
+ 170 contributors
Deployments 32
beta-testing
+ 31 deployments
Languages
Java 82.6%
HTML 10.2%
JavaScript 3.3%
Shell 1.3%
Python 0.8%
XSLT 0.6%
Other 1.2%
Home
Explore
Notifications
Messages
Grok
Lists
Bookmarks
Communities
Premium
Profile
More
Post
For you
Following
What is happening?!
Slava Tykhonov @4tykhonov · Jan 27
My latest experiments on the Open Source #LLM models' "decompression" show great potential in transforming all "packaged" content into a structured format and ingesting it back into the knowledge graph. Even in the current state of art, Generative AI is capable of saving...
Show more
Slava Tykhonov @4tykhonov · 1h
It looks like this can be considered as official start of the battle between human and #AI in the field of information classification and ontologies. And we're in the very beginning...
Search
Subscribe to Premium
Subscribe to unlock new features and if eligible, receive a share of ads revenue.
Subscribe
What's happening
Chiefs at Ravens
NFL · LIVE
#BetMGM
For Everyone... Except Tom Brady
Promoted by BetMGM
Politics · Trending
Trump White House
108K posts
Entertainment · Trending
Ashlee Simpson
Politics · Trending
Iran
Trending with 3 US, Jordanian
Show more
Who to follow
OpenCitations
@opencitations
Follow
CORE
@oacore
Follow
Melissa Sands
@melissaleesa
Follows you
Follow
Show more
Terms of Service
Privacy Policy
Cookie Policy
Accessibility
Ads info
More ...
© 2024 X Corp.

Log In

Log in or sign up with your institutional account — more information about account creation. Leaving your institution? Please contact Harvard Dataverse Support for assistance.

Your Institution

HARVARD UNIVERSITY

Harvard University

Please select... Continue

Allow me to type the name of my institution

Other options

Username/Email GitHub Google ORCID

Sign up for a Dataverse account.

Collection-level

Current Features

Dataverse Collections

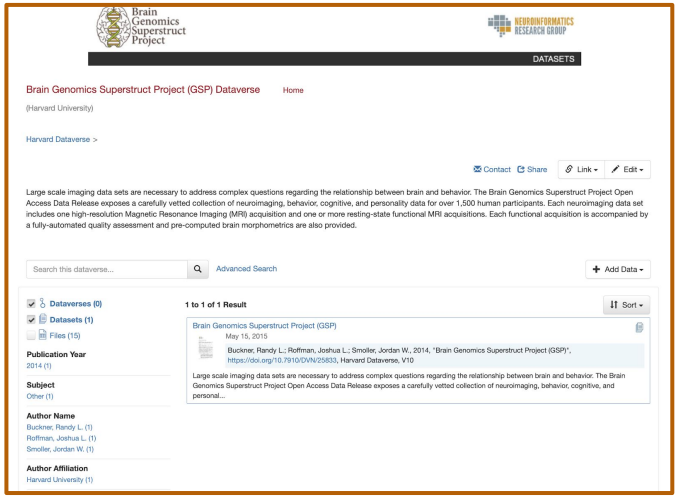
- Own administration
- Own branding (and can be embedded anywhere)

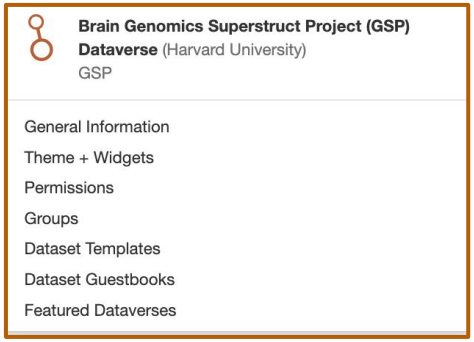
Datasets

- 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

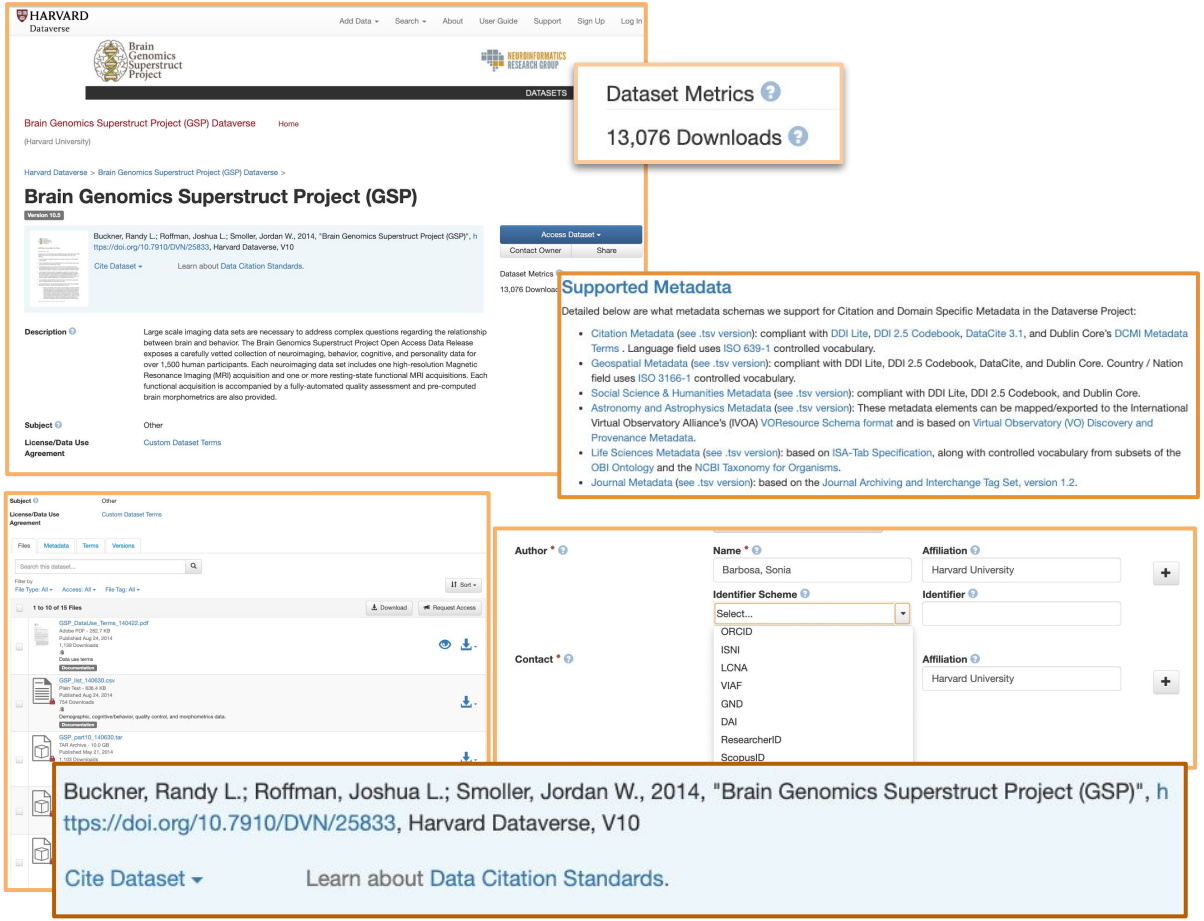
- Own administration
- Own branding (and can be embedded anywhere)

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

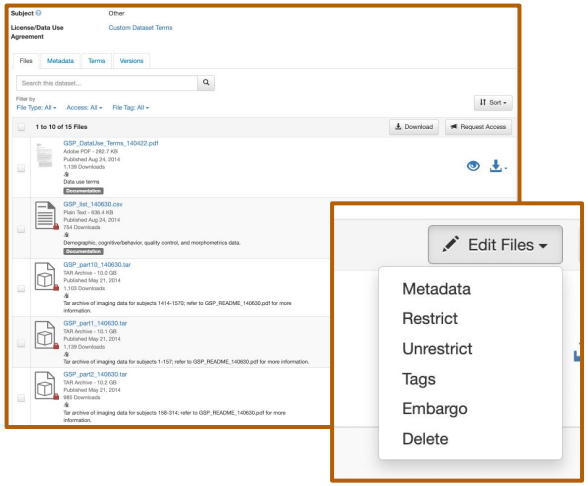
- Own administration
- Own branding (and can be embedded anywhere)

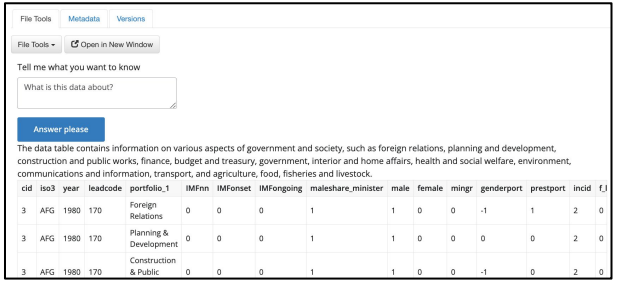
Datasets

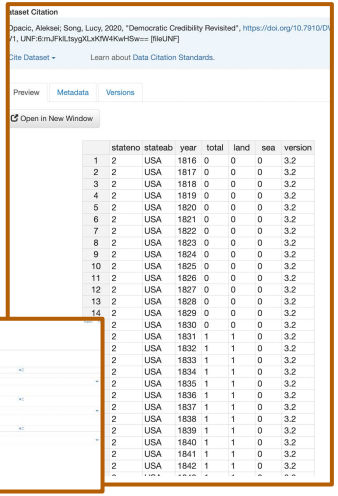
- Citation
- Metadata
- Versioning
- Private URL/*Anonymous Peer Review
- Custom Terms/*Multiple Licenses/Permissions
- Guestbooks
- Publishing Workflows

Files

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





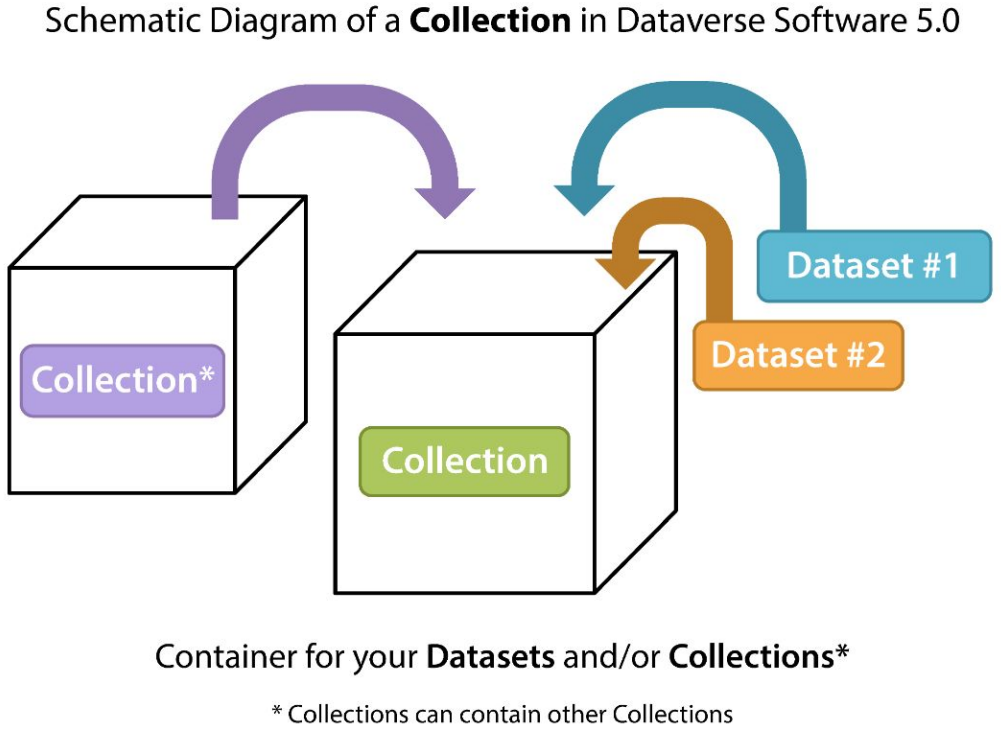


The User Experience: Collections

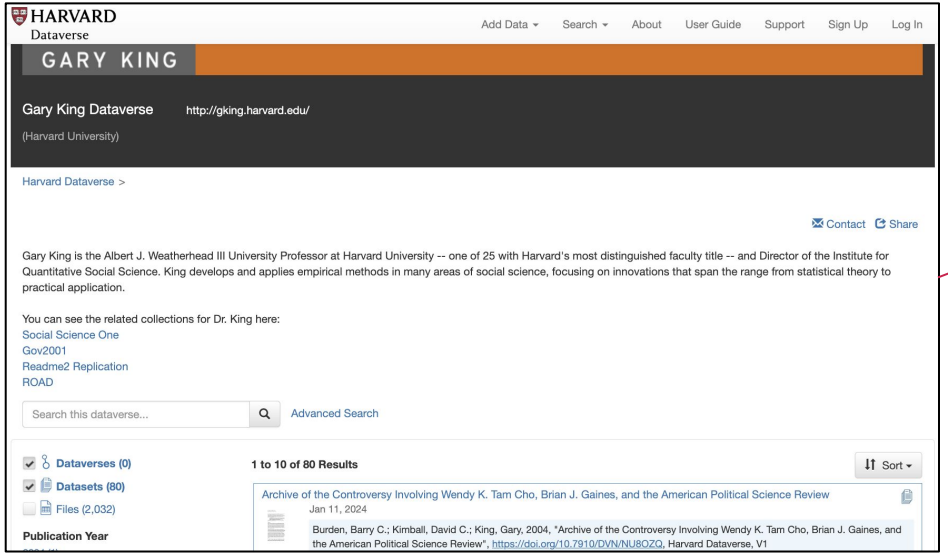
Create a Dataverse "Collection"

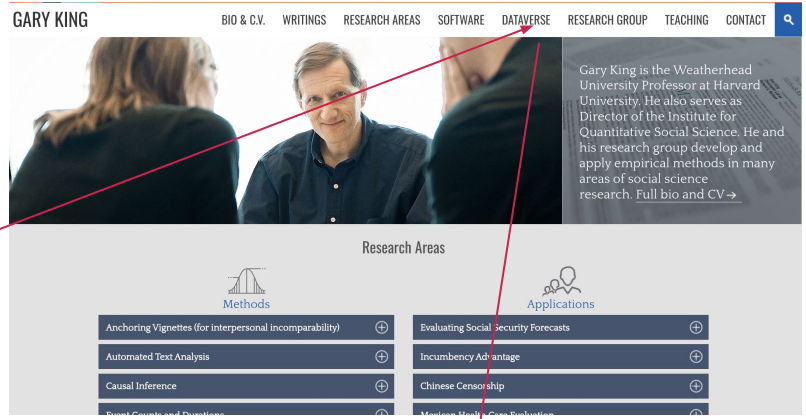
- Edit Dataverse Collection
  - General Information
  - **Theme**
  - Widgets
    - Dataverse Collection Search Box Widget
    - Dataverse Collection Listing Widget
    - Adding Widgets to an OpenScholar Website
  - **Roles & Permissions**
    - Setting Access Configurations
    - Assigning Roles to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

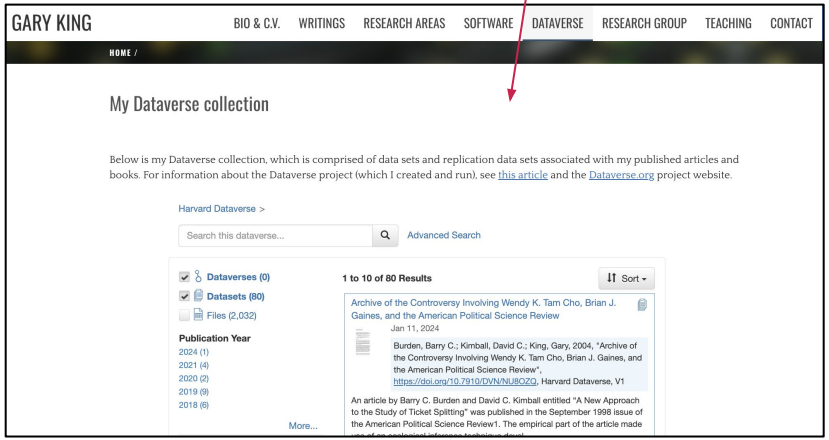
A Dataverse collection is a container for datasets (research data, code, documentation, and metadata) and other Dataverse collections, which can be setup for individual researchers, departments, journals and organizations.



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


HARVARD
Dataverse
Add Data
Search
About
User Guide
Support
Sign Up
Log In
GARY KING
Gary King Dataverse
http://gking.harvard.edu/
(Harvard University)
Harvard Dataverse >
Contact
Share
Gary King is the Albert J. Weatherhead III University Professor at Harvard University -- one of 25 with Harvard's most distinguished faculty title -- and Director of the Institute for Quantitative Social Science. King develops and applies empirical methods in many areas of social science, focusing on innovations that span the range from statistical theory to practical application.
You can see the related collections for Dr. King here:
Social Science One
Gov2001
Readme2 Replication
ROAD
Search this dataverse...
Advanced Search
Dataverses (0)
Datasets (80)
Files (2,032)
Publication Year
1 to 10 of 80 Results
Sort
Archive of the Controversy Involving Wendy K. Tam Cho, Brian J. Gaines, and the American Political Science Review
Jan 11, 2024
Burden, Barry C.; Kimball, David C.; King, Gary, 2004, "Archive of the Controversy Involving Wendy K. Tam Cho, Brian J. Gaines, and the American Political Science Review", https://doi.org/10.7910/DVN/NU8OZQ, Harvard Dataverse, V1


GARY KING
BIO & C.V.
WRITINGS
RESEARCH AREAS
SOFTWARE
DATAVERSE
RESEARCH GROUP
TEACHING
CONTACT
Gary King is the Weatherhead University Professor at Harvard University. He also serves as Director of the Institute for Quantitative Social Science. He and his research group develop and apply empirical methods in many areas of social science research. Full bio and CV →
Research Areas
Methods
Applications
Anchoring Vignettes (for interpersonal incomparability)
Automated Text Analysis
Causal Inference
Evaluating Social Security Forecasts
Incumbency Advantage
Chinese Censorship


GARY KING
BIO & C.V.
WRITINGS
RESEARCH AREAS
SOFTWARE
DATAVERSE
RESEARCH GROUP
TEACHING
CONTACT
HOME /
My Dataverse collection
Below is my Dataverse collection, which is comprised of data sets and replication data sets associated with my published articles and books. For information about the Dataverse project (which I created and run), see this article and the Dataverse.org project website.
Harvard Dataverse >
Search this dataverse...
Advanced Search
Dataverses (0)
Datasets (80)
Files (2,032)
Publication Year
2024 (1)
2021 (4)
2020 (2)
2019 (9)
2018 (6)
More...
1 to 10 of 80 Results
Sort
Archive of the Controversy Involving Wendy K. Tam Cho, Brian J. Gaines, and the American Political Science Review
Jan 11, 2024
Burden, Barry C.; Kimball, David C.; King, Gary, 2004, "Archive of the Controversy Involving Wendy K. Tam Cho, Brian J. Gaines, and the American Political Science Review", https://doi.org/10.7910/DVN/NU8OZQ, Harvard Dataverse, V1
An article by Barry C. Burden and David C. Kimball entitled "A New Approach to the Study of Ticket Splitting" was published in the September 1998 issue of the American Political Science Review1. The empirical part of the article made

DANIEL P. ALDRICH | BIO & C.V. | RESEARCH | DATA | PUBLICATIONS | MEDIA | more...

DANIEL P. ALDRICH

Daniel P. Aldrich is full professor of political science and Director of the Security and Resilience Studies Program at Northeastern University. He researches post-disaster recovery, countering violent extremism, the siting of controversial facilities and the interaction between civil society and the state. He has held posts as a Fulbright Research Fellow and an Abe Fellow at Tokyo University and as an AAAS Science and Technology Fellow with USAID. Aldrich is a contributor to the New York Times, CNN, The Conversation, and the Asahi Shimbun, among other media.

HARVARD Dataverse

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

Daniel Aldrich Dataverse

Northeastern University

(Northeastern University)

Harvard Dataverse >

Contact | Share

Search this dataverse... | Advanced Search

Dataverses (0)
Datasets (10)
Files (27)

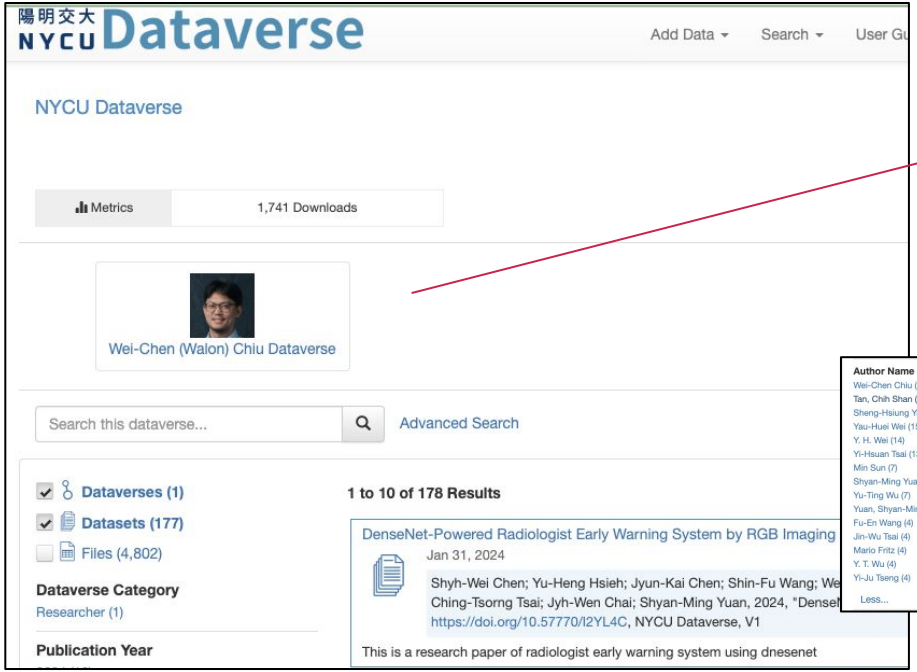
Publication Year
2022 (1)
2020 (3)
2019 (2)
2016 (1)
2012 (1)

More...

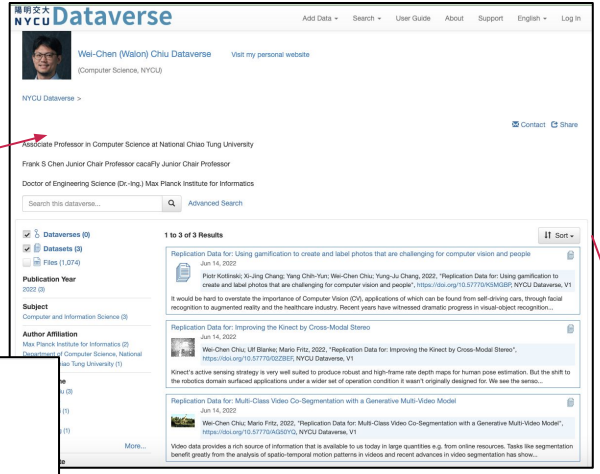
1 to 10 of 10 Results | Sort

Replication Data for: How Social Infrastructure Saves Lives: A Quantitative Analysis of Japan's 3/11 Disasters
Nov 23, 2022
Aldrich, Daniel, 2022, "Replication Data for: How Social Infrastructure Saves Lives: A Quantitative Analysis of Japan's 3/11 Disasters", https://doi.org/10.7910/DVN/YA0EYS, Harvard Dataverse, V1, UNF:6:fe9B3jEgPRcE25dCykAt6g== [fileUNF]
Observers have long debated how societies should invest resources to safeguard citizens and property, especially in the face of increasing shocks and crises. This article explores how social infrastructure -the spaces and places that help build and maintain social ties and trust,...

Brewing Social Capital: A Survey of Breweries and Social Ties across the United States
Dec 20, 2020
Aldrich, Daniel; Schulz, Eric, 2020, "Brewing Social Capital: A Survey of Breweries and Social Ties across the United States",


陽明交大 NYCU Dataverse
Add Data
Search
User Gu
NYCU Dataverse
Metrics
1,741 Downloads
Wei-Chen (Walon) Chiu Dataverse
Search this dataverse...
Advanced Search
Dataverses (1)
Datasets (177)
Files (4,802)
Dataverse Category
Researcher (1)
Publication Year
1 to 10 of 178 Results
DenseNet-Powered Radiologist Early Warning System by RGB Imaging
Jan 31, 2024
Shyh-Wei Chen; Yu-Heng Hsieh; Jyun-Kai Chen; Shin-Fu Wang; We
Ching-Tsorng Tsai; Jyh-Wen Chai; Shyan-Ming Yuan, 2024, "Dense
https://doi.org/10.57770/I2YL4C, NCTU Dataverse, V1
This is a research paper of radiologist early warning system using dnesenet

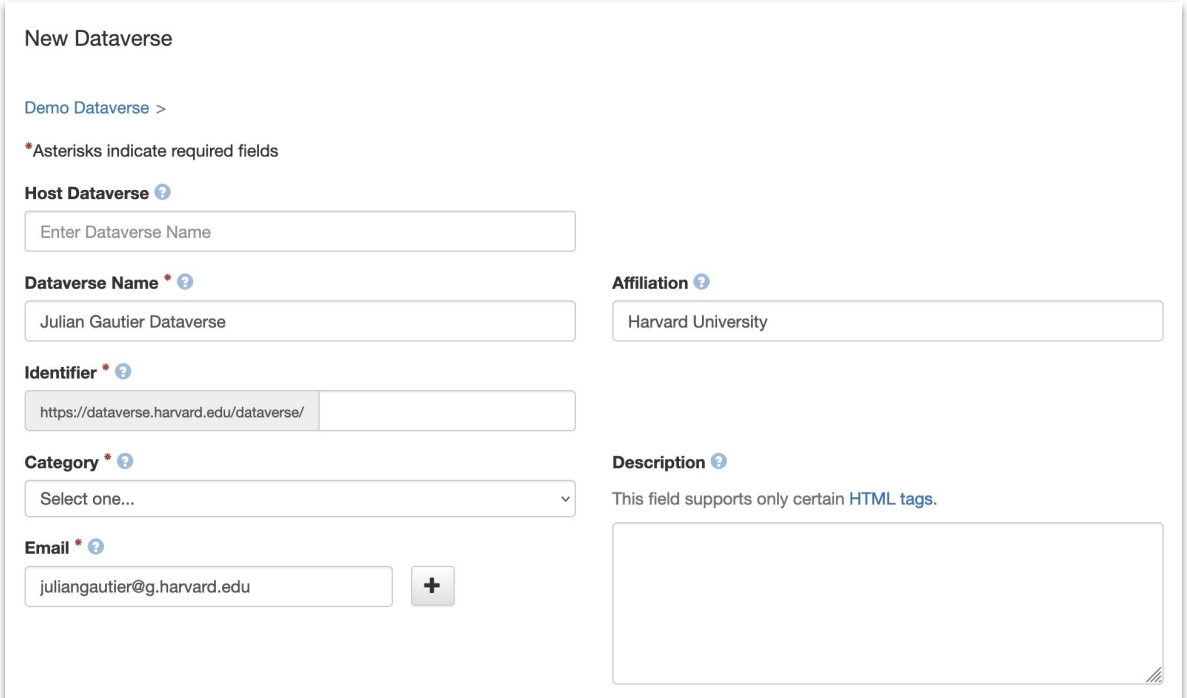

Author Name
Wei-Chen Chiu (29)
Tan, Chih Shan (21)
Sheng-Hsiung Yang (19)
Yau-Huei Wei (15)
Y. H. Wei (14)
Yi-Hsuan Tsai (13)
Min Sun (7)
Shyan-Ming Yuan (7)
Yu-Ting Wu (7)
Yuan, Shyan-Ming (6)
Fu-En Wang (4)
Jin-Wu Tsai (4)
Mario Fritz (4)
Y. T. Wu (4)
Yi-Ju Tseng (4)
Less...
More...


NYCU Dataverse
Wei-Chen (Walon) Chiu Dataverse
Visit my personal website
(Computer Science, NYCU)
NYCU Dataverse >
Contact
Share
Associate Professor in Computer Science at National Chiao Tung University
Frank S Chen Junior Chair Professor cacaFly Junior Chair Professor
Doctor of Engineering Science (Dr.-Ing.) Max Planck Institute for Informatics
Search this dataverse...
Advanced Search
Dataverses (0)
Datasets (3)
Files (1,074)
Publication Year
Subject
Computer and Information Science (3)
Author Affiliation
1 to 3 of 3 Results
Sort
Replication Data for: Using gamification to create and label photos that are challenging for computer vision and people
Jun 14, 2022
Piotr Kotlinski; Xi-Jing Chang; Yang Chih-Yun; Wei-Chen Chiu; Yung-Ju Chang, 2022, "Replication Data for: Using gamification to create and label photos that are challenging for computer vision and people", https://doi.org/10.57770/KSMGBP, NYCU Dataverse, V1
It would be hard to overstate the importance of Computer Vision (CV), applications of which can be found from self-driving cars, through facial recognition to augmented reality and the healthcare industry. Recent years have witnessed dramatic progress in visual-object recognition...
Replication Data for: Improving the Kinect by Cross-Modal Stereo
Jun 14, 2022
Wei-Chen Chiu; Ulf Blanke; Mario Fritz, 2022, "Replication Data for: Improving the Kinect by Cross-Modal Stereo", https://doi.org/10.57770/OIZBEF, NYCU Dataverse, V1
Kinect's active sensing strategy is very well suited to produce robust and high-frame rate depth maps for human pose estimation. But the shift to the robotics domain surfaced applications under a wider set of operation condition it wasn't originally designed for. We see the senso...
Replication Data for: Multi-Class Video Co-Segmentation with a Generative Multi-Video Model
Jun 14, 2022
Wei-Chen Chiu; Mario Fritz, 2022, "Replication Data for: Multi-Class Video Co-Segmentation with a Generative Multi-Video Model", https://doi.org/10.57770/AGBDYQ, NYCU Dataverse, V1
Video data provides a rich source of information that is available to us today in large quantities e.g. from online resources. Tasks like segmentation benefit greatly from the analysis of spatio-temporal motion patterns in videos and recent advances in video segmentation has show...


Wei-Chen (Walon) Chiu; Taiwanese: Khu Uî-Sîn
Home
About Me
Calendar
Associate Professor in Computer Science at National Chiao Tung University
Frank S Chen Junior Chair Professor cacaFly Junior Chair Professor
Doctor of Engineering Science (Dr.-Ing.) Max Planck Institute for Informatics
邱維辰
walon [at] cs [dot] nctu [dot] edu [dot] tw
walon [dot] chiu [at] gmail [dot] com
Curriculum Vitae
Google Scholar
[Office] Room 526, Engineering Building C, Guangfu Campus, NCTU
[Telephone] +886-3-5712121 ext.54730
Information about my research group:
[Name] Enriched Vision Applications Lab
[Location] Room 548, Engineering Building C, Guangfu Campus, NCTU

Metadata you can use to describe collections

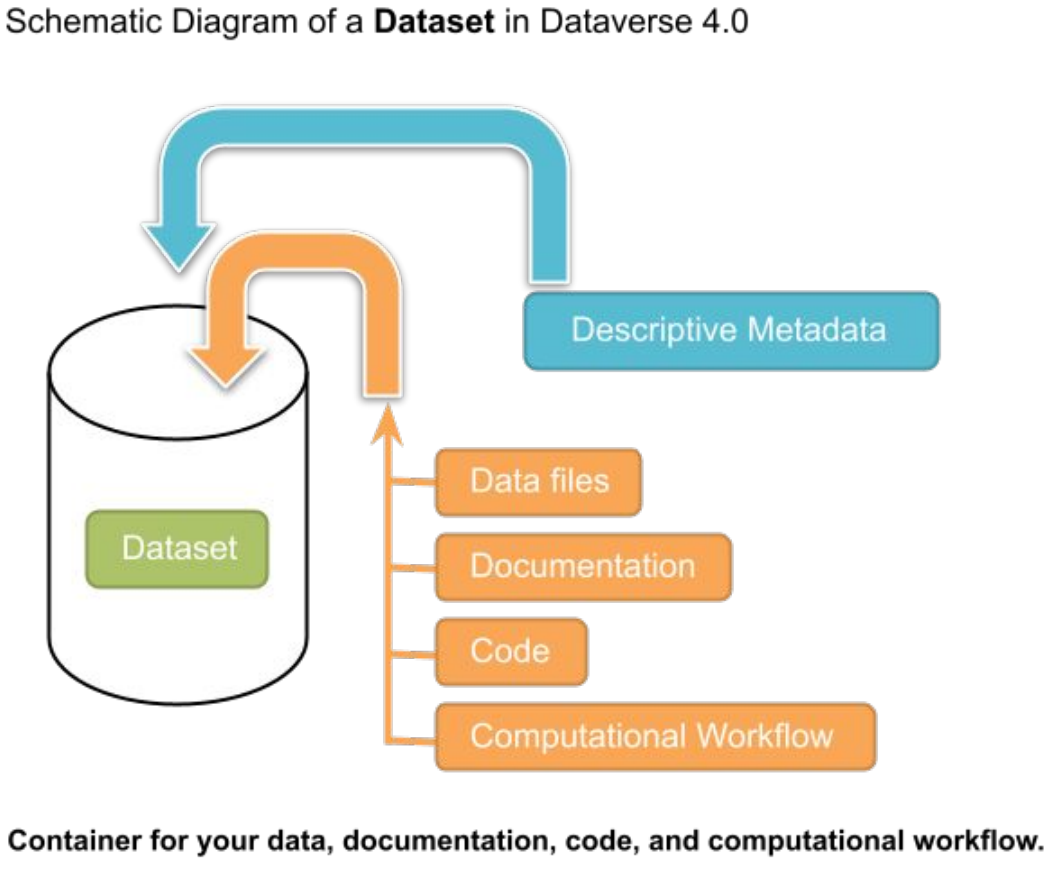
- Names
- Affiliations
- Identifiers of collection URLs
- Categories
- Emails
- Descriptions

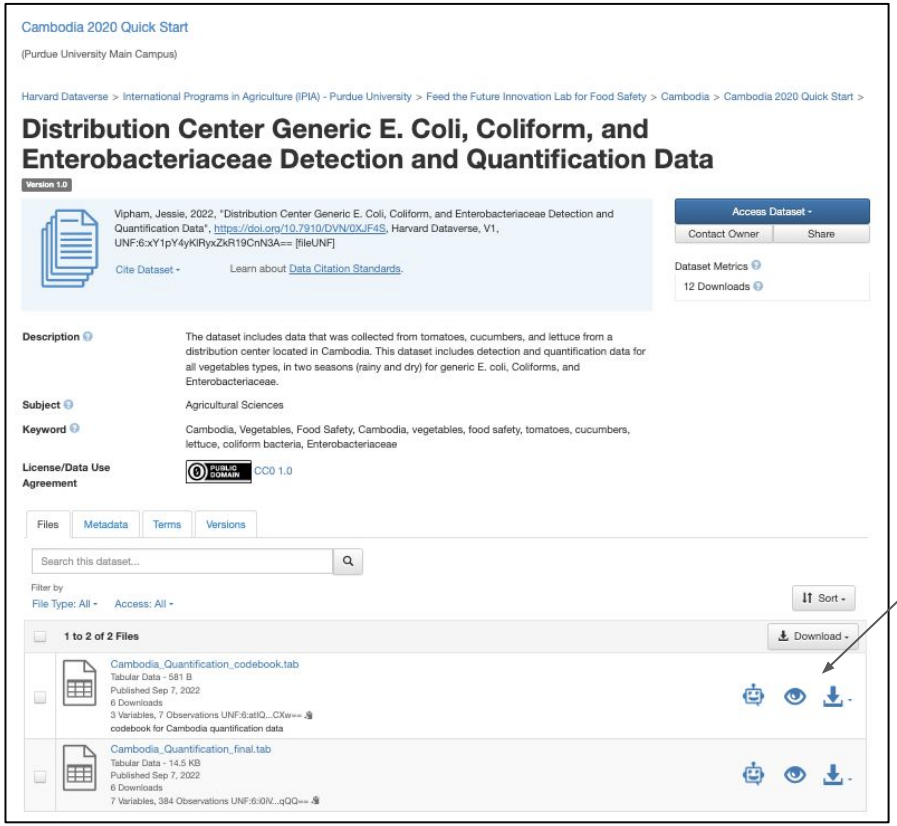


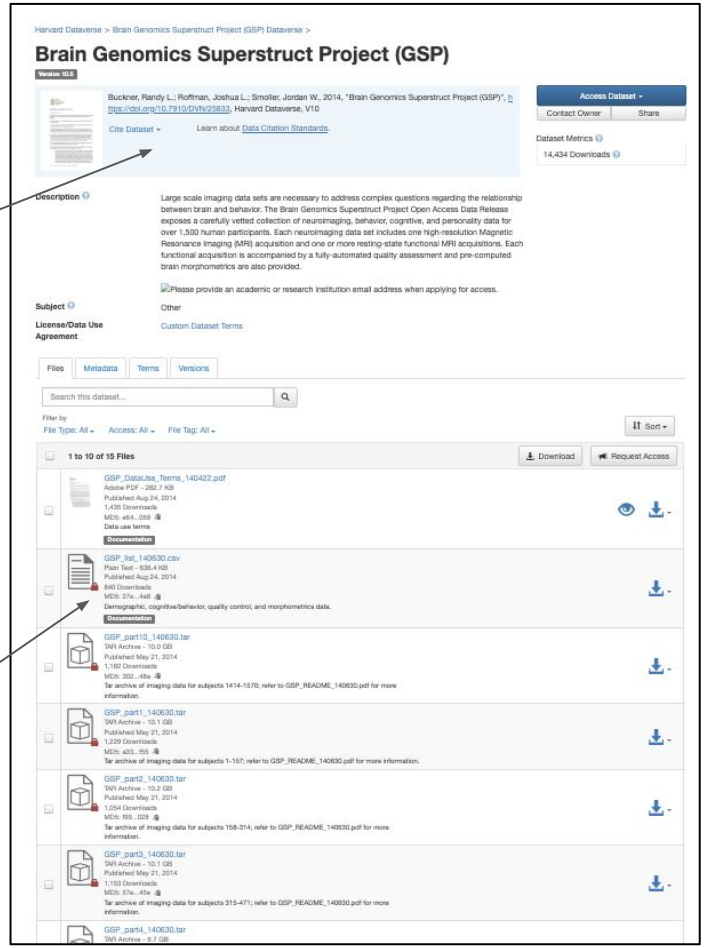
The User Experience: Datasets

Deposit a Dataset

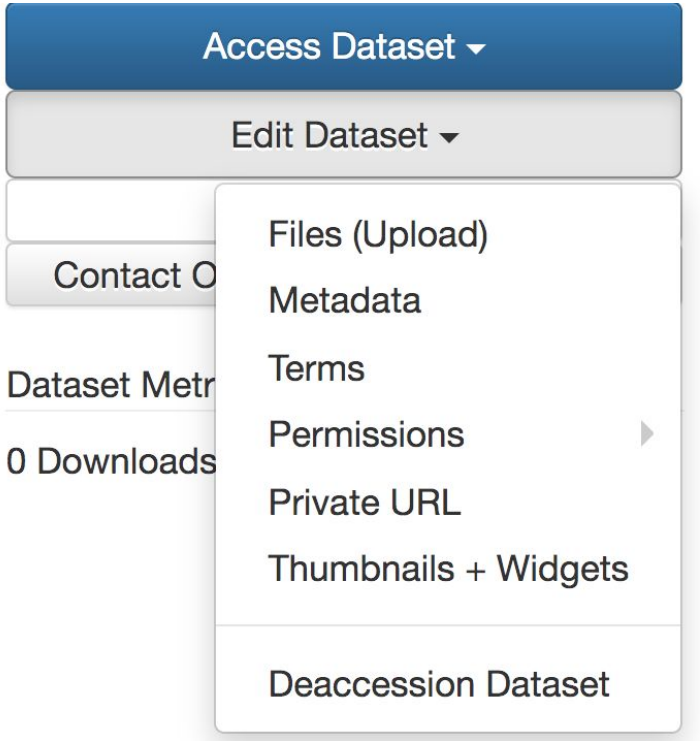
- Supported Metadata
- Adding a New Dataset
- File Upload
- File Handling
- Restricted Files
- Edit Files
- Replace Files
- Terms
- Roles & Permissions
- Data Provenance
- Thumbnails + Widgets
- Publish Dataset
- Submit for Review
- Private URL to Review Unpublished Dataset
- Embargoes
- Dataset Versions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




Cambodia 2020 Quick Start
(Purdue University Main Campus)
Harvard Dataverse > International Programs in Agriculture (IPIA) - Purdue University > Feed the Future Innovation Lab for Food Safety > Cambodia > Cambodia 2020 Quick Start >
Distribution Center Generic E. Coli, Coliform, and Enterobacteriaceae Detection and Quantification Data
Version 1.0
Vipham, Jessie, 2022, "Distribution Center Generic E. Coli, Coliform, and Enterobacteriaceae Detection and Quantification Data", https://doi.org/10.7910/DVN/0XJF4S, Harvard Dataverse, V1, UNF:6:xY1pY4yKlRyxZkR19CnN3A== [fileUNF]
Cite Dataset
Learn about Data Citation Standards.
Access Dataset
Contact Owner
Share
Dataset Metrics
12 Downloads
Description
The dataset includes data that was collected from tomatoes, cucumbers, and lettuce from a distribution center located in Cambodia. This dataset includes detection and quantification data for all vegetables types, in two seasons (rainy and dry) for generic E. coli, Coliforms, and Enterobacteriaceae.
Subject
Agricultural Sciences
Keyword
Cambodia, Vegetables, Food Safety, Cambodia, vegetables, food safety, tomatoes, cucumbers, lettuce, coliform bacteria, Enterobacteriaceae
License/Data Use Agreement
CC0 1.0
Files
Metadata
Terms
Versions
Search this dataset...
Filter by
File Type: All
Access: All
Sort
1 to 2 of 2 Files
Download
Cambodia_Quantification_codebook.tab
Tabular Data - 581 B
Published Sep 7, 2022
6 Downloads
3 Variables, 7 Observations UNF:6:atIQ...CXw==
codebook for Cambodia quantification data
Cambodia_Quantification_final.tab
Tabular Data - 14.5 KB
Published Sep 7, 2022
6 Downloads
7 Variables, 384 Observations UNF:6:i0lV...qQQ==


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.
Access Dataset
Contact Owner
Share
Dataset Metrics
14,434 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.
Please provide an academic or research institution email address when applying for access.
Subject
Other
License/Data Use Agreement
Custom Dataset Terms
Files
Metadata
Terms
Versions
Search this dataset...
Filter by
File Type: All
Access: All
File Tag: All
Sort
1 to 10 of 15 Files
Download
Request Access
GSP_DataUse_Terms_140422.pdf
Adobe PDF - 282.7 KB
Published Aug 24, 2014
1,435 Downloads
Data use terms
Documentation
GSP_list_140630.csv
Plain Text - 636.4 KB
Published Aug 24, 2014
640 Downloads
Demographic, cognitive/behavior, quality control, and morphometrics data.
Documentation
GSP_part10_140630.tar
TAR Archive - 10.0 GB
Published May 21, 2014
1,182 Downloads
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
1,229 Downloads
Tar archive of imaging data for subjects 1-157; refer to GSP_README_140630.pdf for more information.
GSP_part2_140630.tar
TAR Archive - 10.2 GB
Published May 21, 2014
1,254 Downloads
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
1,153 Downloads
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

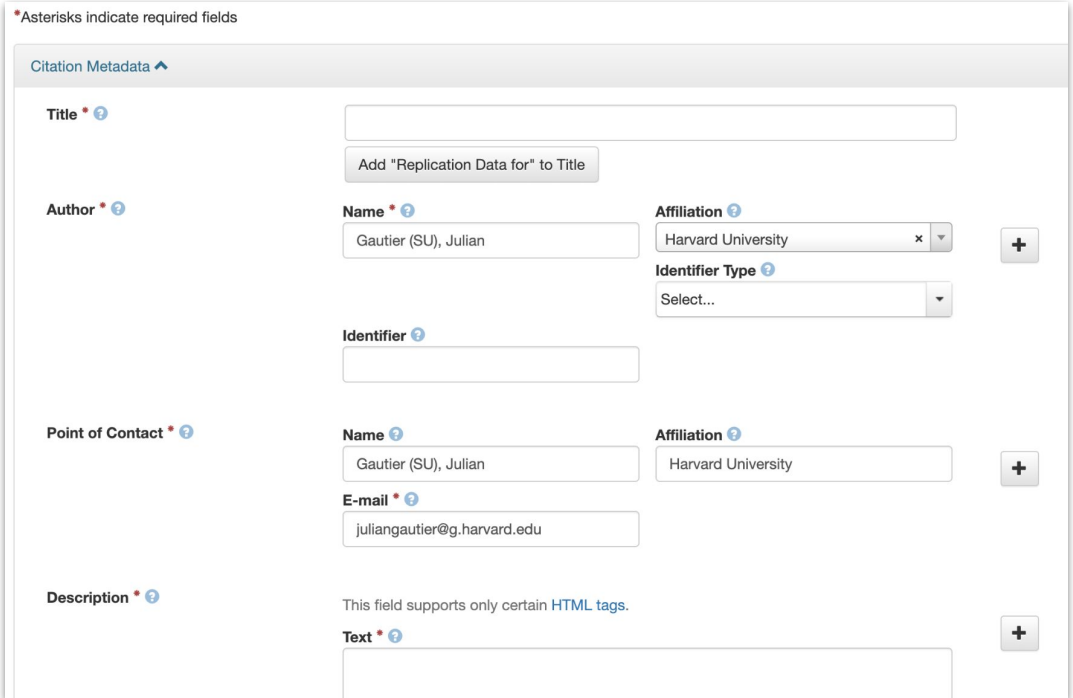
Datasets - Editing Options



- A **Private URL** can be generated for DRAFT datasets to share a dataset with a coauthor or journal, before publishing.

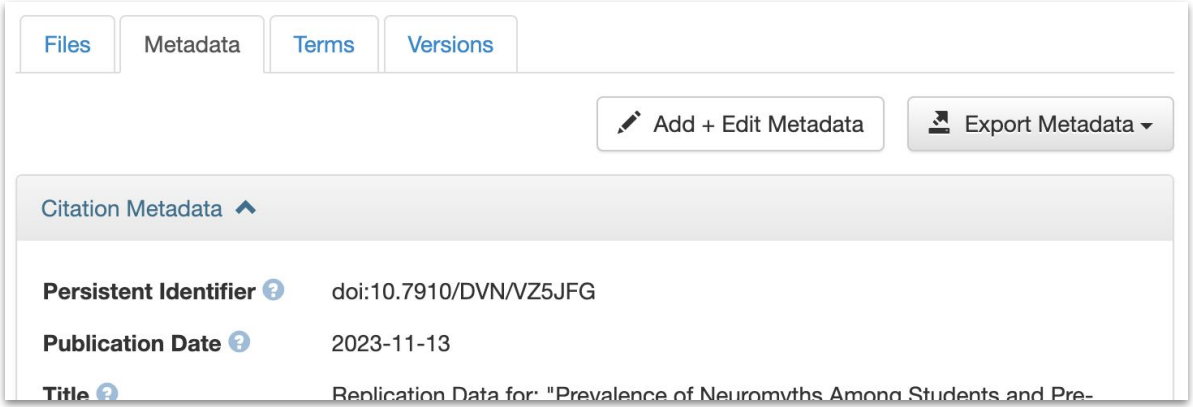
Metadata you can use to describe datasets

Using many fields shown in collapsible panels, such as “Citation Metadata” and “Terms”



Metadata Dataverse creates for datasets

- Persistent IDs
- Creation dates
- Publication dates
- Version numbers
- Sum of file downloads
- Sum of page views (in repositories with Make Data Count enabled)



Metadata Dataverse creates for datasets

- Differences among versions

	Dataset Version	Summary	Contributors	
☐	12.0	**Citation Metadata:** Description (1 Changed); Notes (Changed); **Additional Citation Metadata:** (3 Changed); **Files (Added: 98; Removed: 84);** View Details	Julian Gautier	202
☐	11.0	**Citation Metadata:** Description (1 Changed); Notes (Changed); **Additional Citation Metadata:** (4 Changed); **Files (Added: 84; Removed: 57);** View Details	Julian Gautier	202
☐	10.0	**Citation Metadata:** Notes (Changed); Description (1 Changed); **Additional Citation Metadata:** (2 Removed, 2 Changed); **Files (Added: 57; Removed: 50);** View Details	Julian Gautier	202
☐	9.0	**Citation Metadata:** Description (1 Changed); **Additional Citation Metadata:** (1 Added); **Files (Added: 50; Removed: 37);** View Details	Julian Gautier	202
☐	8.1	**Citation Metadata:** Description (1 Changed); View Details	Julian Gautier	202
☐	8.0	**Citation Metadata:** Description (1 Changed); **Additional Citation Metadata:** (1 Added, 1 Changed); **Files (Added: 1; Replaced: 1);** View Details	Julian Gautier	202
☐	7.0	**Citation Metadata:** Description (1 Changed); **Files (Added: 35; Removed: 36);** View Details	Julian Gautier	202

The User Experience: Files

File sizes and upload options and examples

File formats:

- All formats accepted

Per file size limitations*:

- Set by repository

Tabular Ingest*:

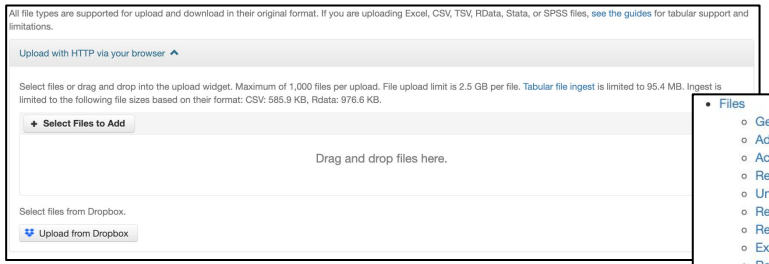
- Set by repository

Total project size limitations*:

- 1 TB (non Harvard)
- 2.5TB (Harvard)
- Paid options available for larger projects

Policies*:

- No identifiable data
- Copyright adherence
- No protected password files
- Terms required (CC0 by default)



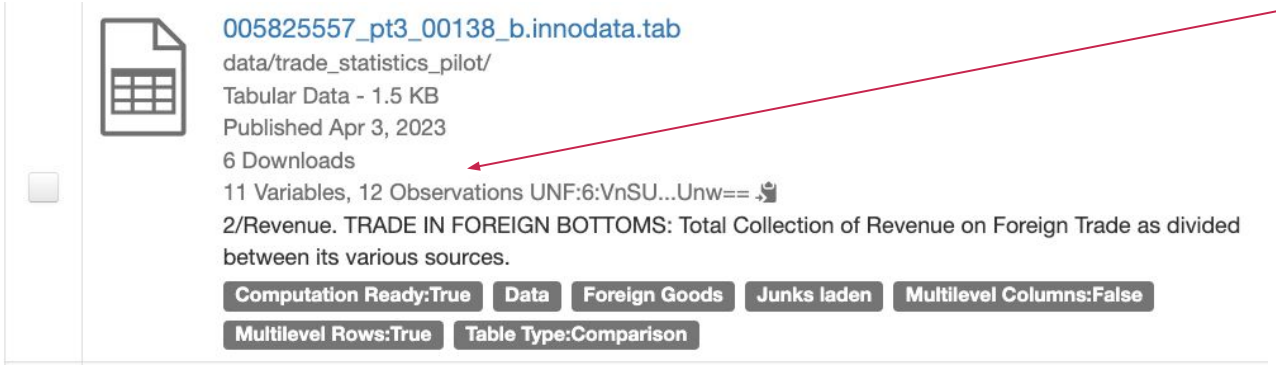
File upload options

Native APIs file handling options

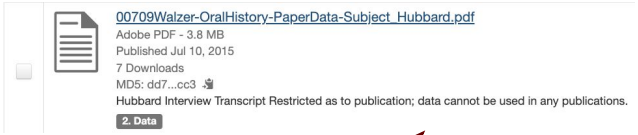
- Files
  - Get JSON Representation of a File
  - Adding Files
  - Accessing (downloading) files
  - Restrict Files
  - Uningest a File
  - Reingest a File
  - Redetect File Type
  - Extract NcML
  - Replacing Files
  - Deleting Files
  - Getting File Metadata
  - Updating File Metadata
  - Editing Variable Level Metadata
  - Provenance
    - Get Provenance JSON for an uploaded file
    - Get Provenance Description for an uploaded file
    - Create/Update Provenance JSON and provide related entity name for an uploaded file
    - Create/Update Provenance Description for an uploaded file
    - Delete Provenance JSON for an uploaded file
  - Datafile Integrity
  - Get External Tool Parameters
  - Get Fixity Algorithm

***Determined by repository**. Unless otherwise stated, these policies are for Harvard repository.

File level metadata considerations and features



Metadata extracted from tabular and other files with tabular ingest feature: Observations, Variables

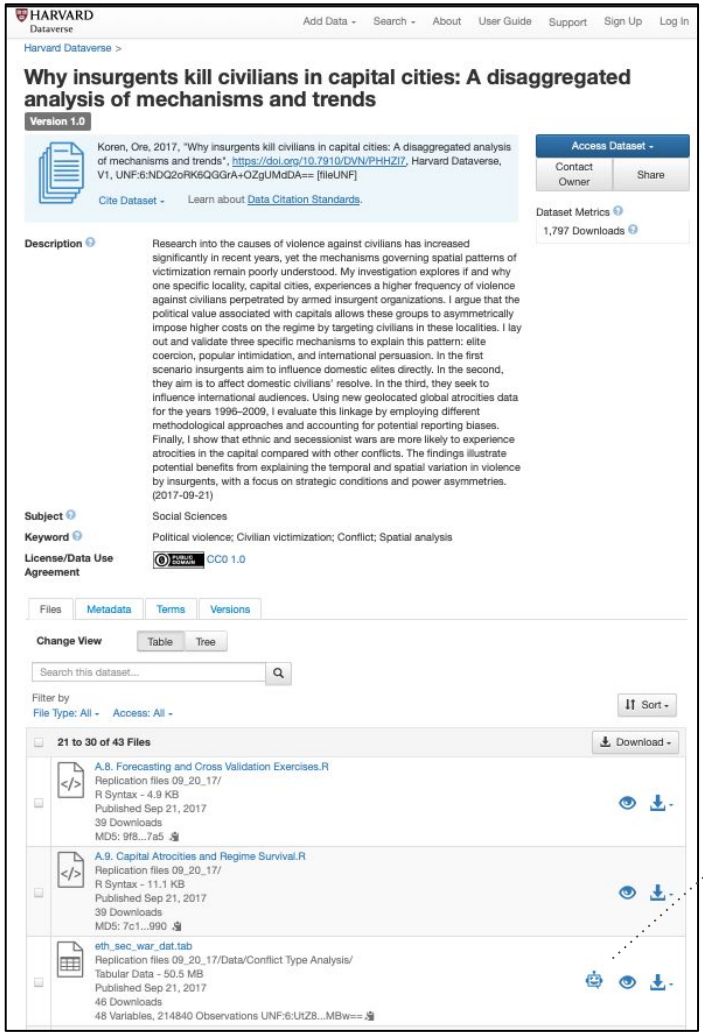


File level "terms"

File level "tags"

Plan to populate additional metadata for files: description, edit file names, add tags

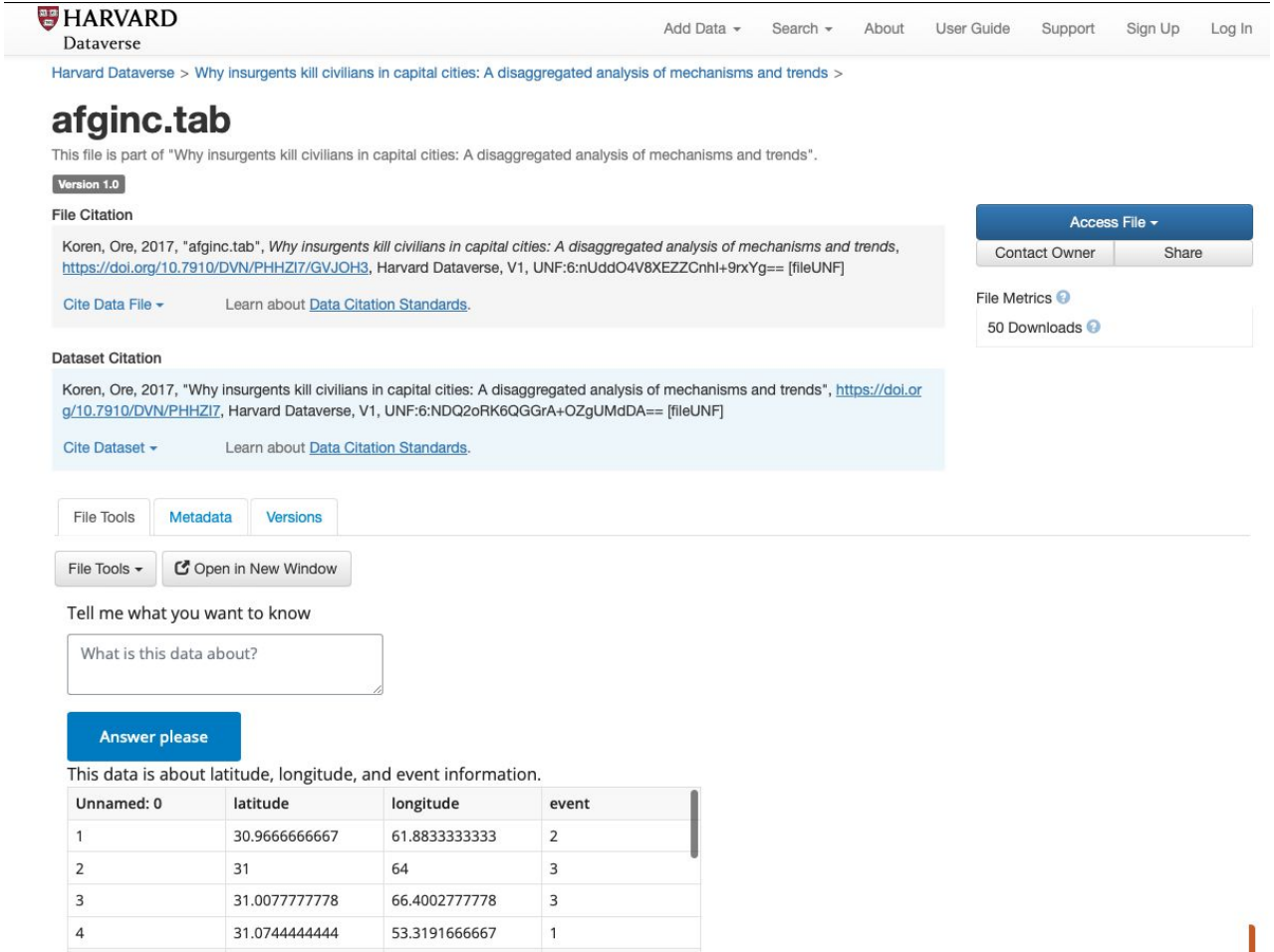
Towards AI integration



https://doi.org/10.7910/DVN/PHHZI7

This chatbot only sees the tabular data but is clueless about the metadata

Tell me what this data are about



Cool but poor

This chatbot only sees the tabular data but is clueless about the metadata

HARVARD UNIVERSITY

Tell me the range of latitudes and longitudes with the highest number of events

HARVARD Dataverse

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

Harvard Dataverse > Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends >

afginc.tab

This file is part of "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends".

Version 1.0

File Citation

Koren, Ore, 2017, "afginc.tab", *Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends*, https://doi.org/10.7910/DVN/PHHZI7/GVJOH3, Harvard Dataverse, V1, UNF:6:nUddO4V8XEZZCnhl+9rxYg== [fileUNF]

Cite Data File ▾ Learn about Data Citation Standards.

Access File ▾ | Contact Owner | Share

File Metrics

50 Downloads

Dataset Citation

Koren, Ore, 2017, "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends", https://doi.org/10.7910/DVN/PHHZI7, Harvard Dataverse, V1, UNF:6:NDQ2oRK6QGGrA+OZgUMdDA== [fileUNF]

Cite Dataset ▾ Learn about Data Citation Standards.

File Tools | Metadata | Versions

File Tools ▾ Open in New Window

Tell me what you want to know

longitudes with the highest number of events

Answer please

The range of latitudes and longitudes with the highest number of events is 30.9666666666667, 61.8833333333333 to 32.0, 67.25.

Unnamed: 0	latitude	longitude	event
1	30.9666666667	61.8833333333	2
2	31	64	3
3	31.0077777778	66.4002777778	3
4	31.0744444444	53.3191666667	1

ok-ish

This chatbot only sees the tabular data but is clueless about the metadata

Map the range of latitudes and longitudes with the highest number of events to the names of countries

HARVARD Dataverse

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

Harvard Dataverse > Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends >

afginc.tab

This file is part of "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends".

Version 1.0

File Citation

Koren, Ore, 2017, "afginc.tab", *Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends*, https://doi.org/10.7910/DVN/PHHZI7/GVJOH3, Harvard Dataverse, V1, UNF:6:nUddO4V8XEZZCnhl+9rxYg== [fileUNF]

Cite Data File ▾ Learn about Data Citation Standards.

Access File ▾ | Contact Owner | Share

File Metrics

50 Downloads

Dataset Citation

Koren, Ore, 2017, "Why insurgents kill civilians in capital cities: A disaggregated analysis of mechanisms and trends", https://doi.org/10.7910/DVN/PHHZI7, Harvard Dataverse, V1, UNF:6:NDQ2oRK6QGGrA+OZgUMdDA== [fileUNF]

Cite Dataset ▾ Learn about Data Citation Standards.

File Tools | Metadata | Versions

File Tools ▾ | Open in New Window

Tell me what you want to know

map the range of latitudes and longitudes with the highest number of events to the names of countries

Answer please

The countries with the highest number of events in the given range of latitudes and longitudes are Afghanistan, Iran, and Pakistan.

Unnamed: 0	latitude	longitude	event
1	30.9666666667	61.8833333333	2
2	31	64	3
3	31.0077777778	66.4002777778	3
4	31.0744444444	53.3191666667	1

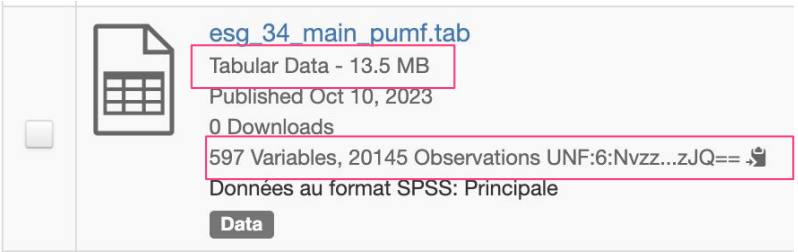
LLM kicks in

This chatbot only sees the tabular data but is clueless about the metadata

Tabular Data Ingest

Tabular data formats (CSV, Excel, SPSS, Stata, R)

- Tabular data ingest
  - Conversion to Tab-delimited text
  - Metadata extraction of variables
  - Data Documentation Initiative (DDI) metadata standard

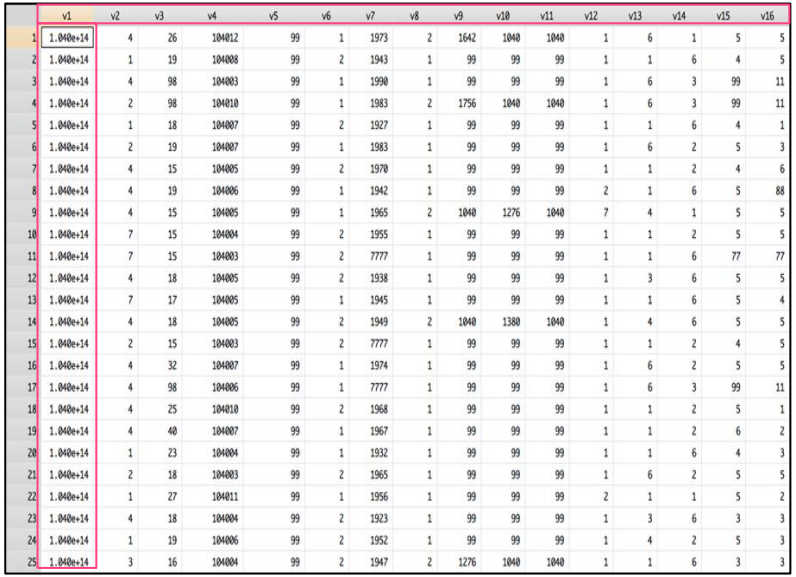


Reuses and integrations

- Enables variable metadata searching
- Preview and exploration
- Summary statistics and analysis
- Download as original format, Tab-delimited or R format

Tabular data: depositing files

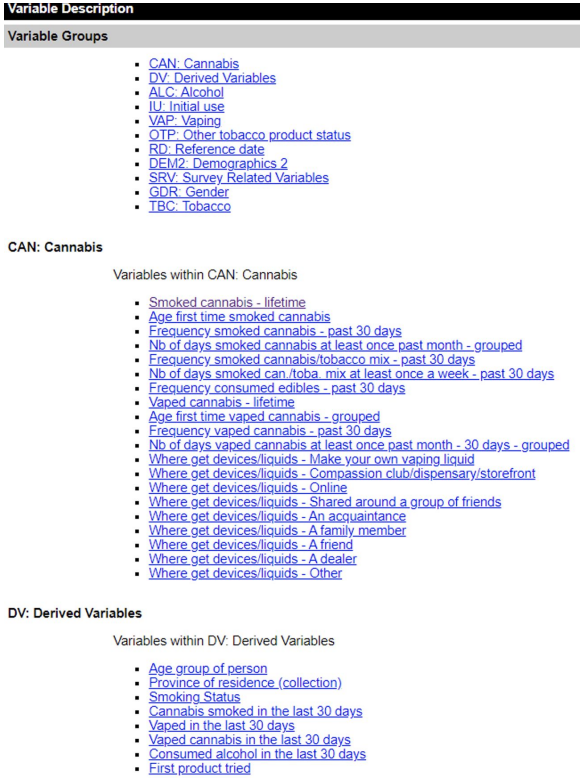
- Upload tabular files (SPSS, STATA, R, CSV, Excel)
- Data ingest tips:
  - Responses/ observations as rows
  - Variables/concepts as columns
  - No notes / improper formatting
  - Does not support multiple spreadsheets in one file
  - Excel (XLSX) is supported
  - CSV requires documentation
  - Encode in original language
  - Ingest errors can occur



	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
1	1.040e+14	4	26	104012	99	1	1973	2	1642	1040	1040	1	6	1	5	5
2	1.040e+14	1	19	104008	99	2	1943	1	99	99	99	1	1	6	4	5
3	1.040e+14	4	98	104003	99	1	1990	1	99	99	99	1	6	3	99	11
4	1.040e+14	2	98	104010	99	1	1983	2	1756	1040	1040	1	6	3	99	11
5	1.040e+14	1	18	104007	99	2	1927	1	99	99	99	1	1	6	4	1
6	1.040e+14	2	19	104007	99	1	1983	1	99	99	99	1	6	2	5	3
7	1.040e+14	4	15	104005	99	2	1970	1	99	99	99	1	1	2	4	6
8	1.040e+14	4	19	104006	99	1	1942	1	99	99	99	2	1	6	5	88
9	1.040e+14	4	15	104005	99	1	1965	2	1040	1276	1040	7	4	1	5	5
10	1.040e+14	7	15	104004	99	2	1955	1	99	99	99	1	1	2	5	5
11	1.040e+14	7	15	104003	99	2	7777	1	99	99	99	1	1	6	77	77
12	1.040e+14	4	18	104005	99	2	1938	1	99	99	99	1	3	6	5	5
13	1.040e+14	7	17	104005	99	1	1945	1	99	99	99	1	1	6	5	4
14	1.040e+14	4	18	104005	99	2	1949	2	1040	1380	1040	1	4	6	5	5
15	1.040e+14	2	15	104003	99	2	7777	1	99	99	99	1	1	2	4	5
16	1.040e+14	4	32	104007	99	1	1974	1	99	99	99	1	6	2	5	5
17	1.040e+14	4	98	104006	99	1	7777	1	99	99	99	1	6	3	99	11
18	1.040e+14	4	25	104010	99	2	1968	1	99	99	99	1	1	2	5	1
19	1.040e+14	4	40	104007	99	1	1967	1	99	99	99	1	1	2	6	2
20	1.040e+14	1	23	104004	99	1	1932	1	99	99	99	1	1	6	4	3
21	1.040e+14	2	18	104003	99	2	1965	1	99	99	99	1	6	2	5	5
22	1.040e+14	1	27	104011	99	1	1956	1	99	99	99	2	1	1	5	2
23	1.040e+14	4	18	104004	99	2	1923	1	99	99	99	1	3	6	3	3
24	1.040e+14	1	19	104006	99	2	1952	1	99	99	99	1	4	2	5	3
25	1.040e+14	3	16	104004	99	2	1947	2	1276	1040	1040	1	1	6	3	3

Tabular data: variable metadata

- Assigns Unique identifiers
- Structured metadata for names, labels, codes, values
- Data contents:
  - Variables
  - Questionnaires & interviews (question text, universes)
  - Recodes, codes, values and notes
- Weighting
- Summary statistics
- Export variable metadata
- Export HTML and XML Codebook



Source: Canadian Tobacco and Nicotine Survey, 2021. DDI Codebook (Odesi)

Geospatial Data

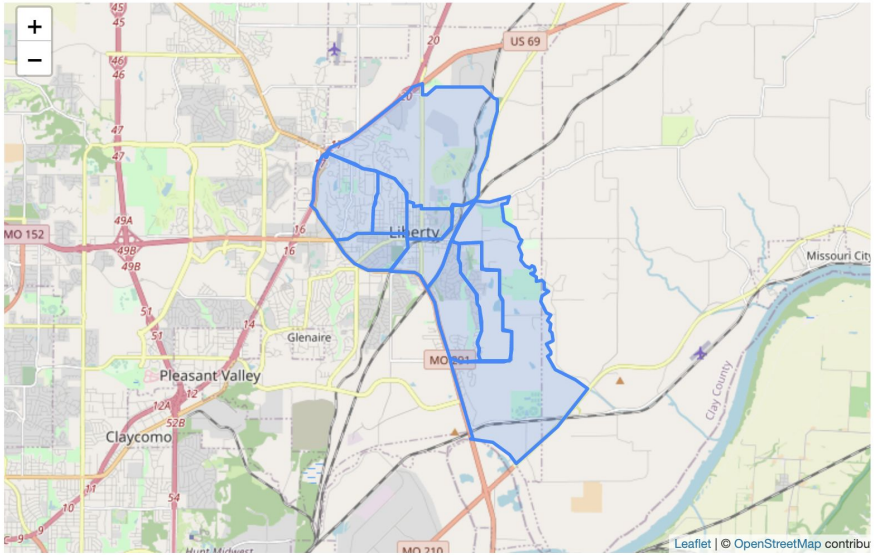
cd_2006_orig_CBF.zip
Shapefile as ZIP Archive - 27.2 MB
Published Apr 4, 2023
0 Downloads
MD5: c0e...22e
UNI-CEN Digital Boundary File for Census Division (CD), 2006, in Esri Shapefile format and NAD83 CSRS / EPSG:3348 projection. This file is part of the CBF-Original Shorelines series. Retrieved from: https://borealisdata.ca/dataverse/unicen. For more information about the project, visit https://observatory.uwo.ca/unicen.

pr_1871_orig_CBF.geojson
GeoJSON - 21.5 MB
Published Apr 4, 2023
2 Downloads
MD5: 14f...fa6
UNI-CEN Digital Boundary File for Province/Territory (PR), 1871, in geojson format and WGS84 / EPSG:4326 projection. This file is part of the CBF-Original Shorelines series. Retrieved from: https://borealisdata.ca/dataverse/unicen. For more information about the project, visit https://observatory.uwo.ca/unicen.

GeoJSON Previewer

Preview | Metadata | Versions

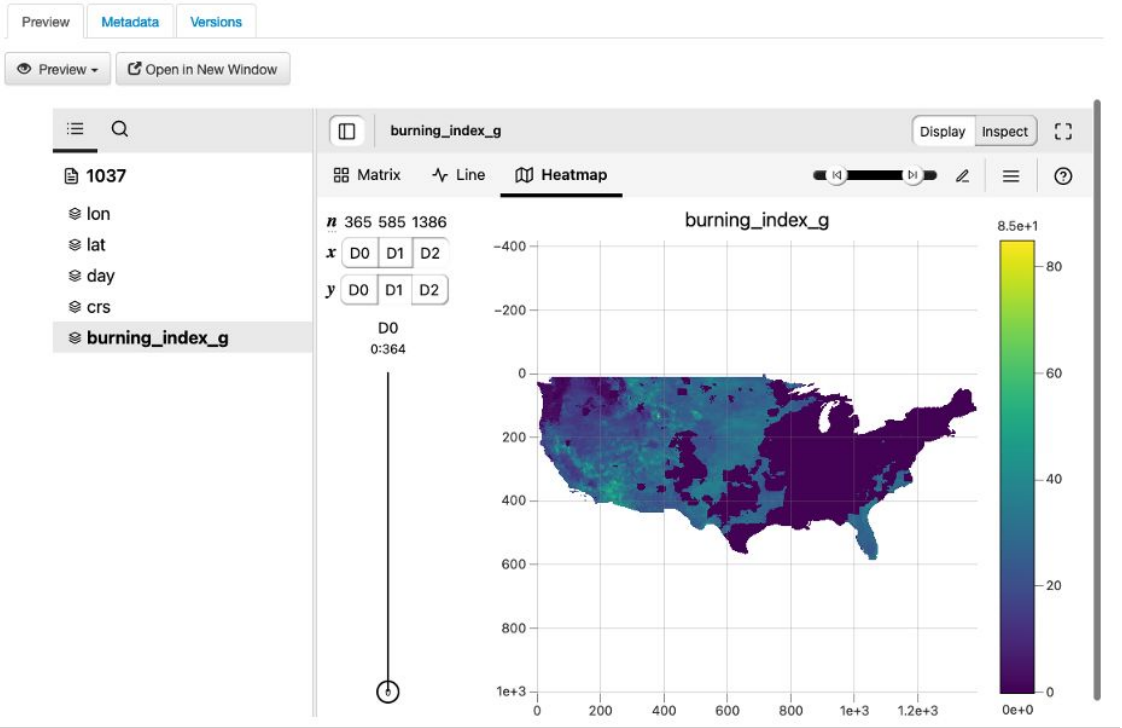
Open in New Window



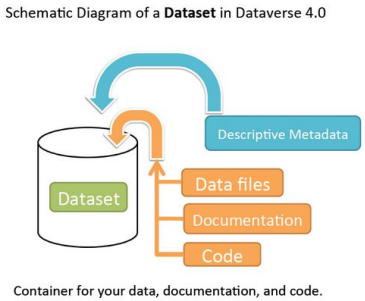
Previewers originally developed by QDR and maintained at https://github.com/GlobalDataverseCommunityConsortium/dataverse-previewers. Feedback and contributions welcome.

NetCDF & HDF5 Previewer

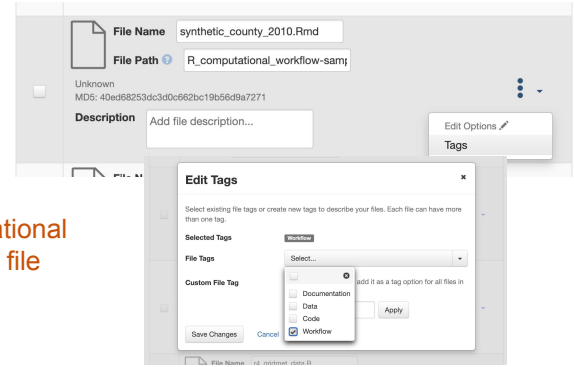
- Common scientific data formats
- Metadata extracted in NcML (XML) format and saved as an auxiliary file
- H5Web visualization and file preview tool
  - Preview data
  - Explore and visualize data



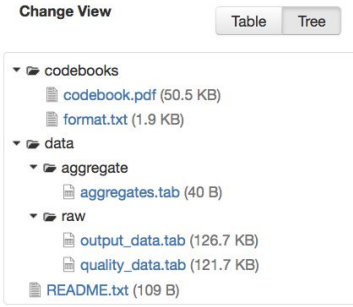
File Types, Format, Documentation



Computational workflow file support



Folder hierarchy support



Tabular data support

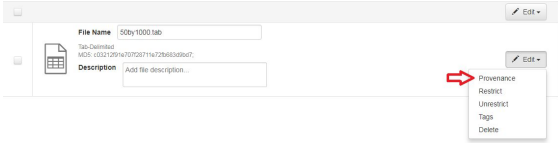
Supported File Formats

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

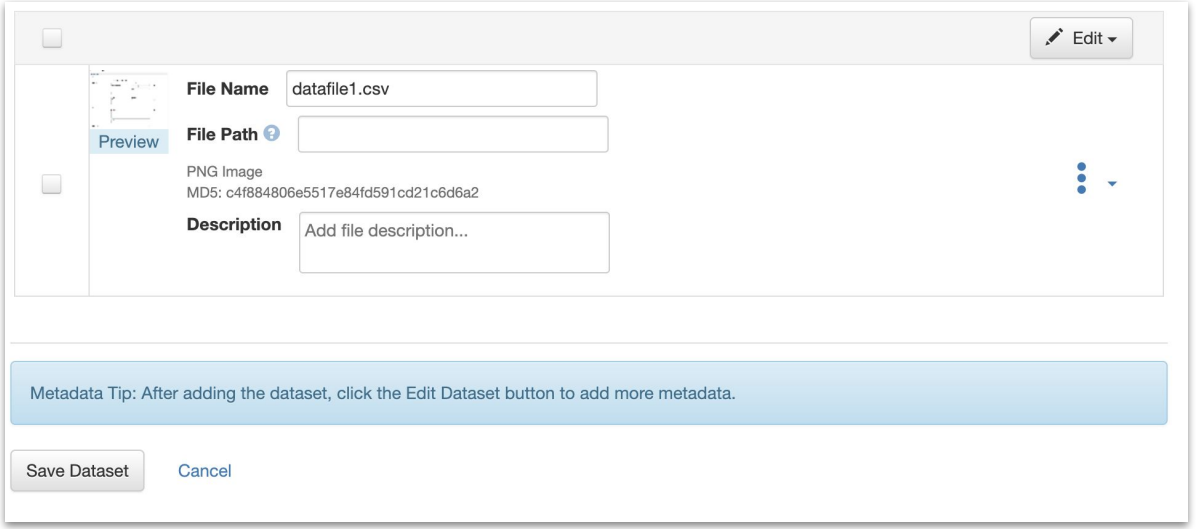
File level Provenance support



The Dataverse® Project

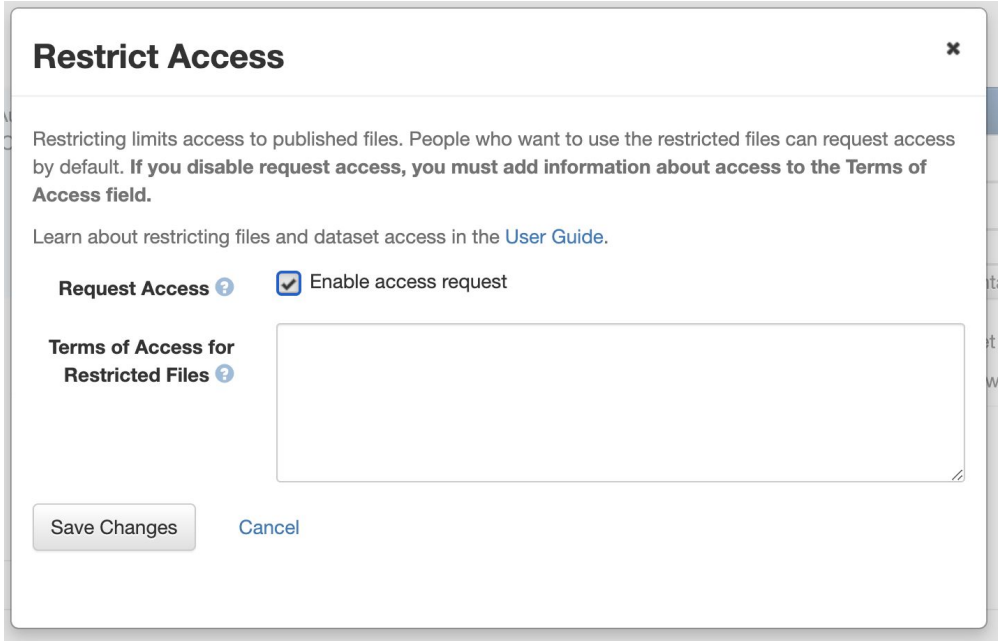
Metadata you can use to describe files

- File names
- File paths
- Descriptions



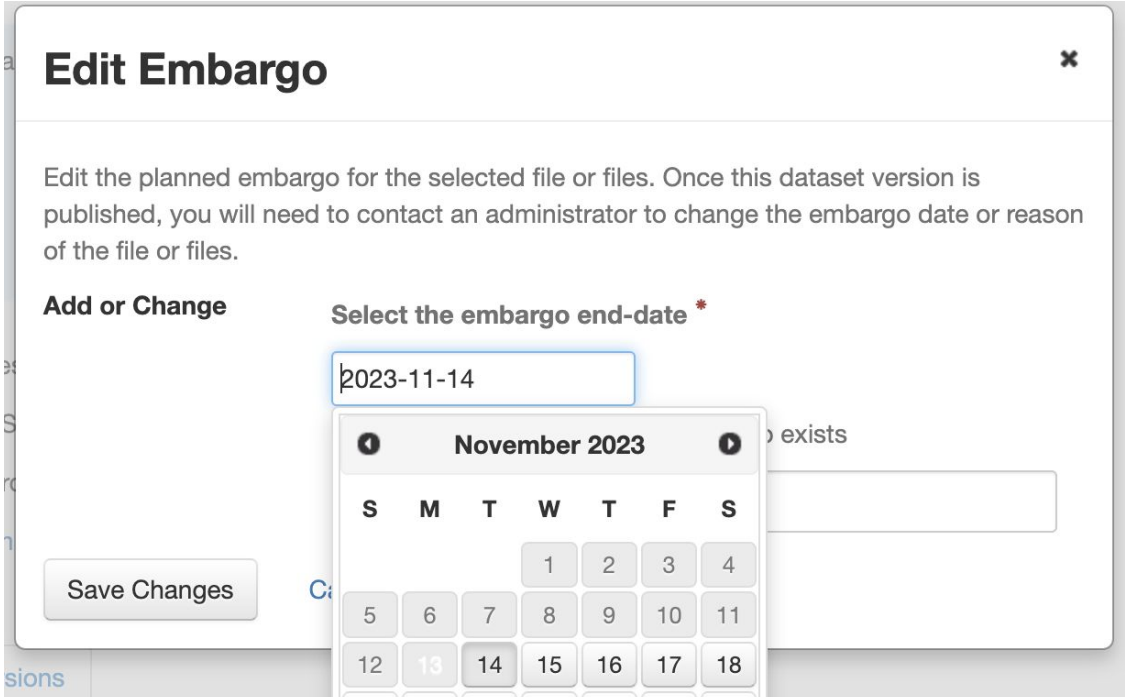
Metadata you can use to describe files

- Tags
- Restrictions
- Embargo dates and reasons



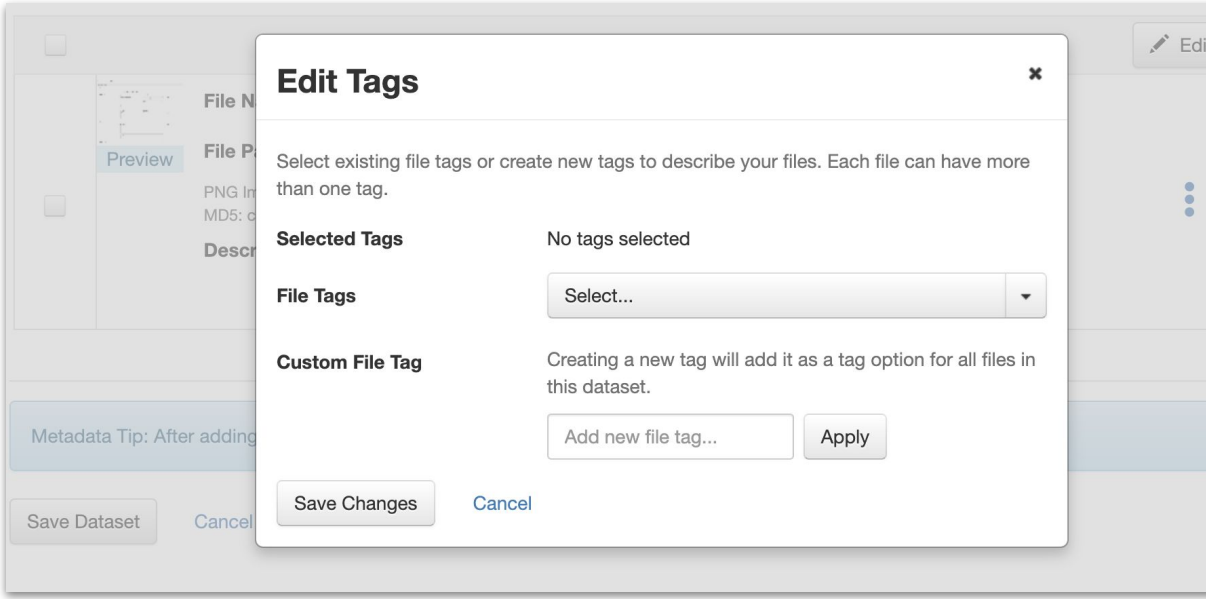
Metadata you can use to describe files

- Tags
- Restrictions
- Embargo dates and reasons



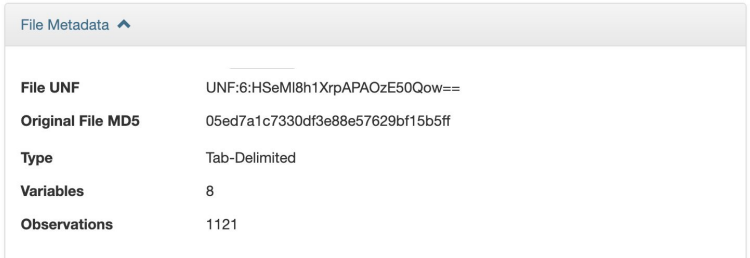
Metadata you can use to describe files

- Tags
- Restrictions
- Embargo dates and reasons

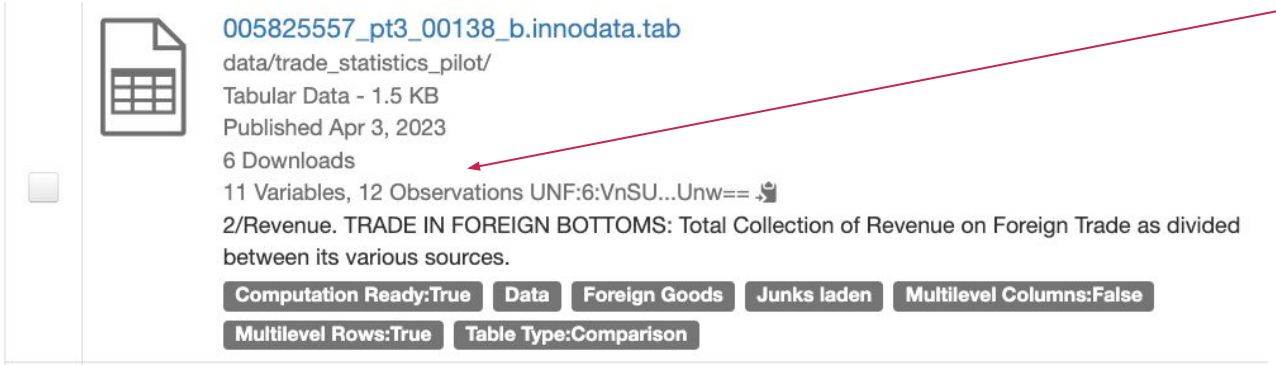


Metadata Dataverse creates for ingested tabular files

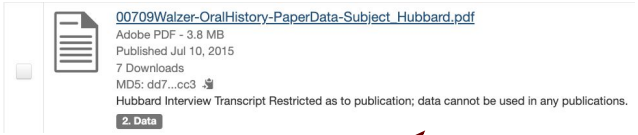
- Counts of variables and observations
- Variable statistics
- Checksums (UNFs)



File level metadata considerations and features



Metadata extracted from tabular and other files with tabular ingest feature: Observations, Variables



File level "terms"

File level "tags"

Plan to populate additional metadata for files: description, edit file names, add tags

The User Experience: Files and External Tools

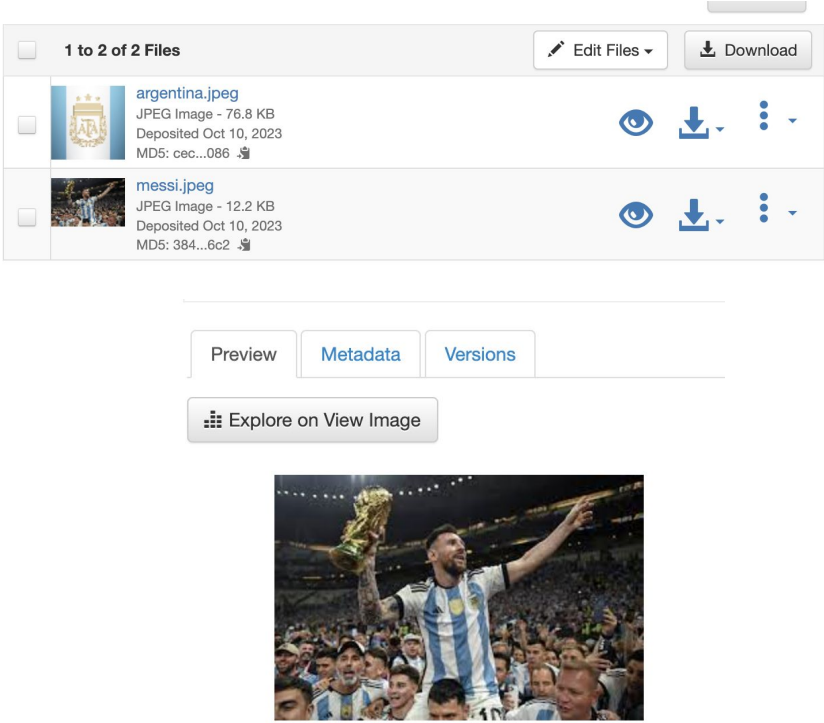
External Tools

- External tools can provide additional features that are not part of the Dataverse Software itself, such as data file previews, visualization, and curation
- Communicate with Dataverse via its robust APIs
- Can be defined for Datasets or Files
- Can be Read-only or Read / Write

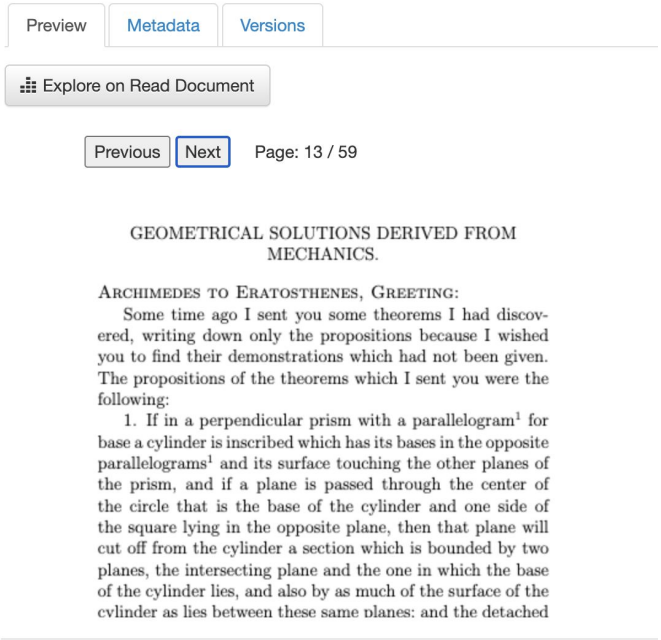
Contents:

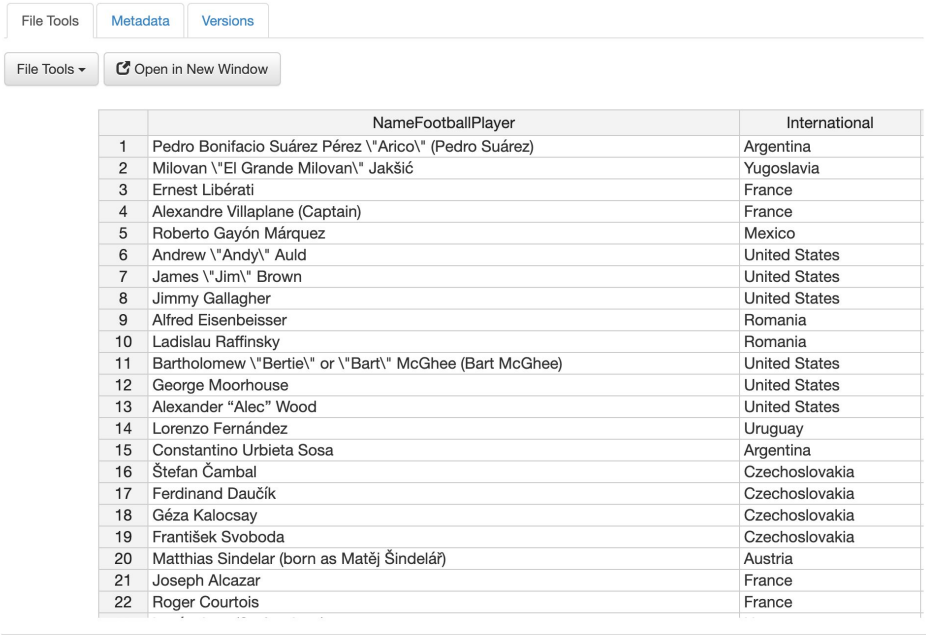
File Previewers

- A set of tools that display the content of files, allowing them to be viewed without downloading the file, including
  - audio
  - html
  - Hypothes.is annotations
  - images
  - PDF
  - text
  - video
  - tabular data
  - spreadsheets



File Previewers (more examples)

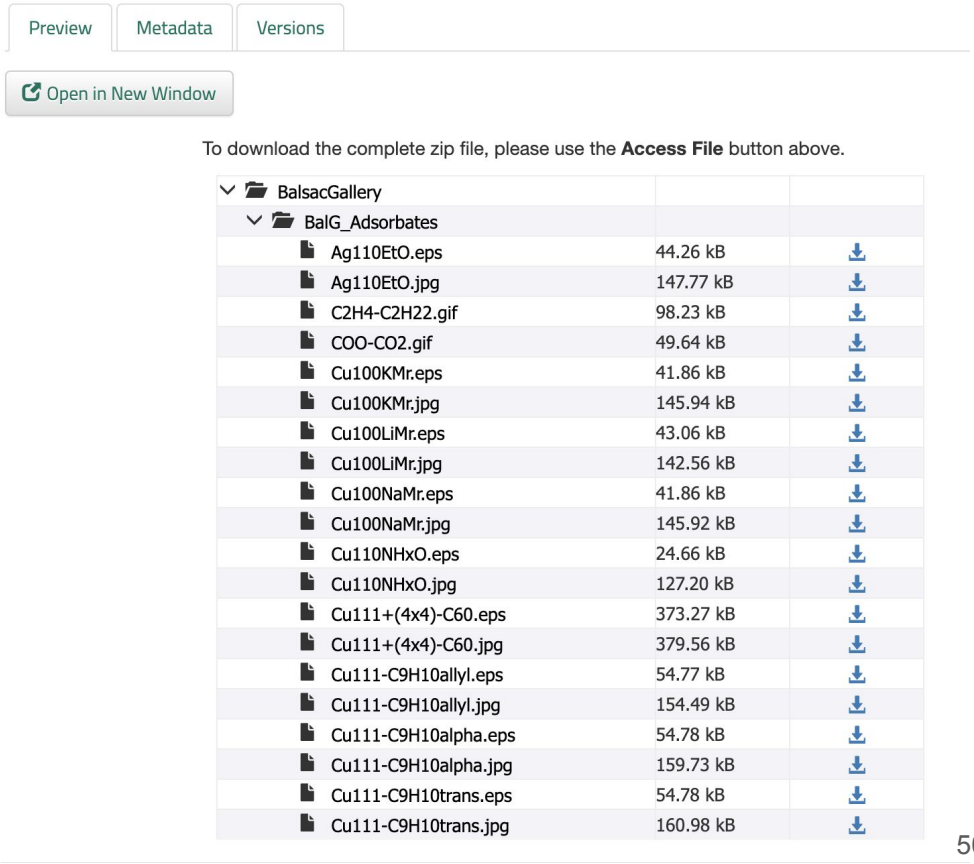




	NameFootballPlayer	International
1	Pedro Bonifacio Suárez Pérez \"Arico\" (Pedro Suárez)	Argentina
2	Milovan \"El Grande Milovan\" Jakšić	Yugoslavia
3	Ernest Libérati	France
4	Alexandre Villaplane (Captain)	France
5	Roberto Gayón Márquez	Mexico
6	Andrew \"Andy\" Auld	United States
7	James \"Jim\" Brown	United States
8	Jimmy Gallagher	United States
9	Alfred Eisenbeisser	Romania
10	Ladislau Raffinsky	Romania
11	Bartholomew \"Bertie\" or \"Bart\" McGhee (Bart McGhee)	United States
12	George Moorhouse	United States
13	Alexander “Alec” Wood	United States
14	Lorenzo Fernández	Uruguay
15	Constantino Urbieta Sosa	Argentina
16	Štefan Čambal	Czechoslovakia
17	Ferdinand Daučík	Czechoslovakia
18	Géza Kalocsay	Czechoslovakia
19	František Svoboda	Czechoslovakia
20	Matthias Sindelar (born as Matěj Šindelář)	Austria
21	Joseph Alcazar	France
22	Roger Courtois	France

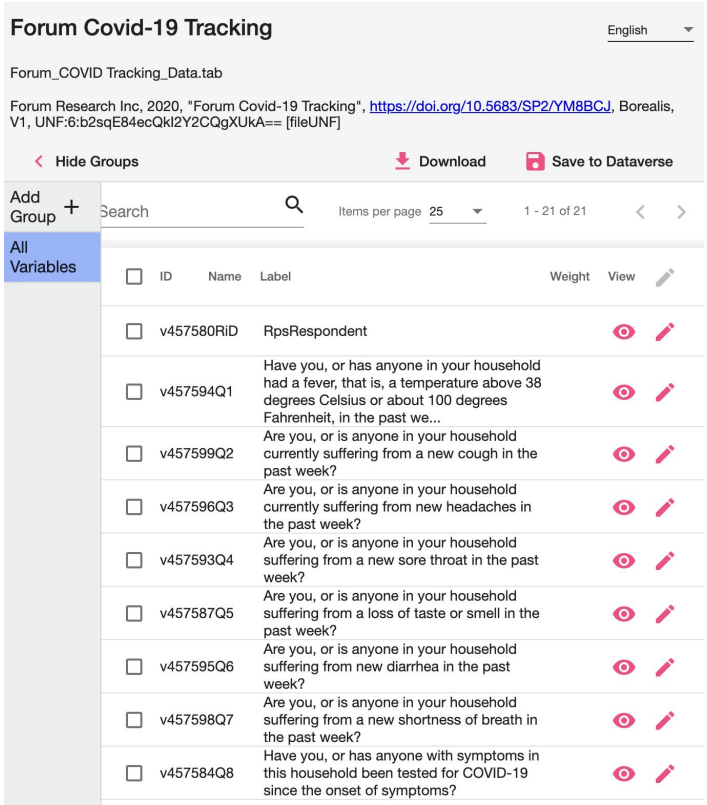
Zip File Previewer +

- A previewer that will show you the internal content and structure of a zip file (or electronic lab notebook)
- Uses the Range functionality in our Access api; so it's not just a viewer, it's an individual file unpacker



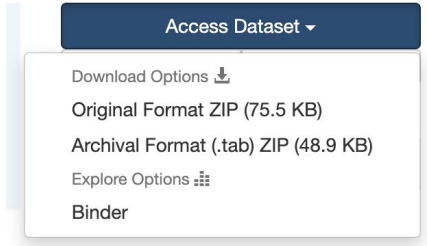
File Exploration, Curation, and Query Tools

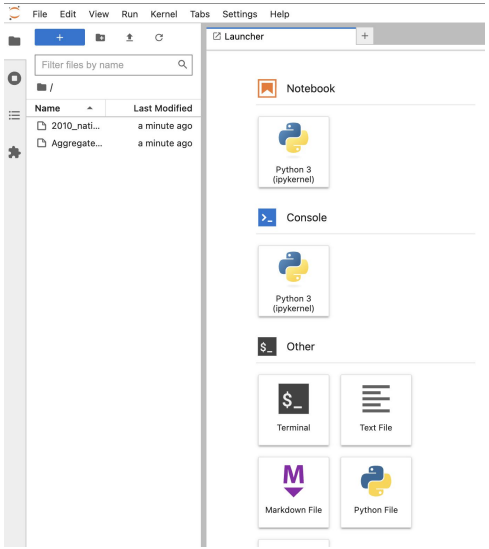
- File level **explore** tools provide a variety of features from data visualization to statistical analysis
- File level **configure** tools allow (authorized) users to send metadata about the file back to Dataverse
- File level **query** tools allow the user to ask questions (e.g. natural language queries) of a data table's contents without having to download the file



Dataset External Tools

- Dataset level **explore** tools allow the user to explore all the files in a dataset - common use case is reproducibility
  - **WholeTale** - creates reproducible research packages based on popular tools such as Jupyter and RStudio
  - **Binder** - spins up custom computing environments in the cloud (including Jupyter notebooks)
- Dataset level **configure** tools allow (authorized) users to send metadata about the dataset back to Dataverse
  - **Turbo Curator** *(coming soon)* - provides recommendations to improve the dataset metadata



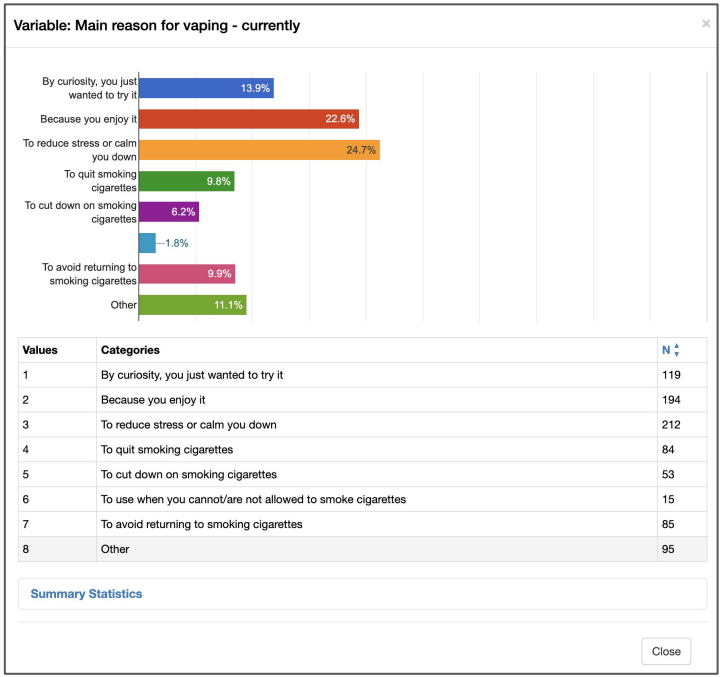


Spotlight: Survey Data Analysis using Data Explorer

- Data Explorer 2.0 Github (open-source external Dataverse tool)
- Contributed by Borealis (Canada)

Data Analysis using Data Explorer

- Helps users understand the data
  - Preview data and metadata for variables, questions, notes, universes
  - Get started with data analysis; digging, filtering, summary statistics
- Analyse and explore data
  - Frequencies charts
  - Cross-tabulation builder
  - Subsetting & download
- Analysing in Data Explorer promotes:
  - Data reuse
  - Transparency and reproducibility
  - Teaching data concepts
  - Open access and sharing



Source: VAP_35R, Canadian Tobacco and Nicotine Survey, 2021.
Odesi https://odesi.ca

Connecting to Data Explorer

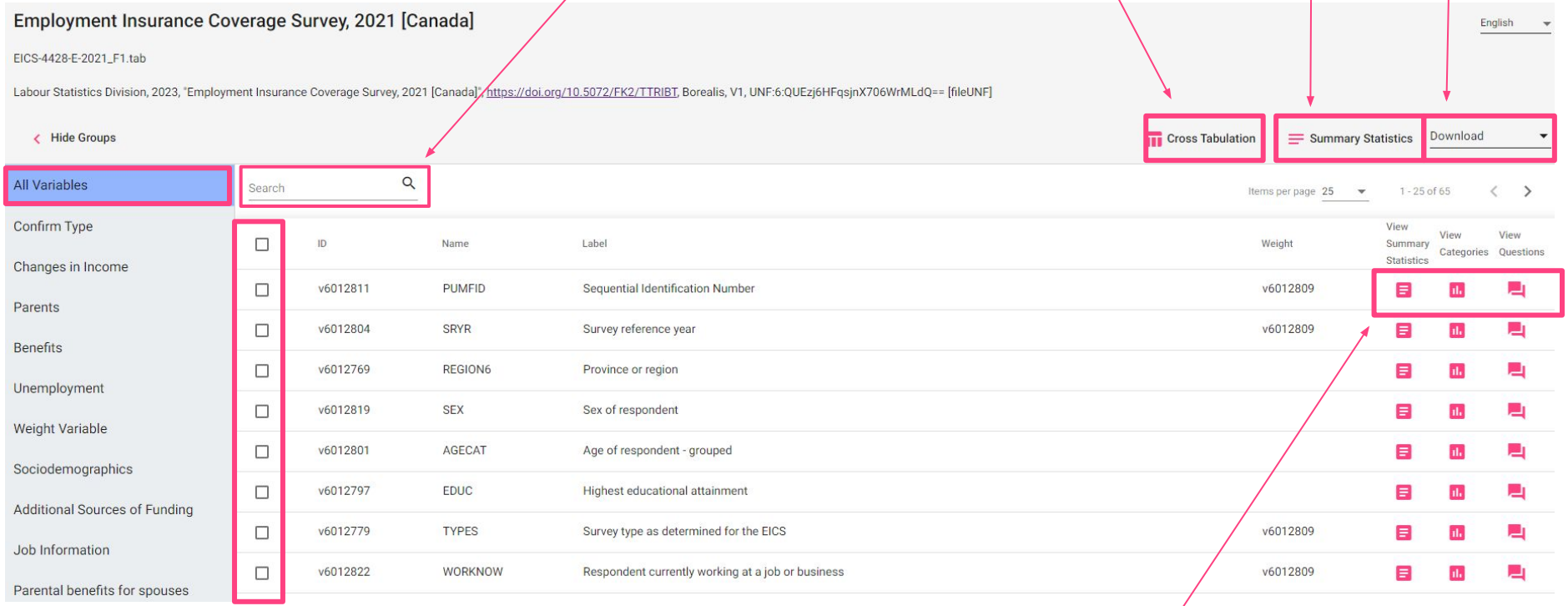
Data Explorer

Search for variables

Cross tabulate two or more variables

View multiple summary statistics

Subset and download

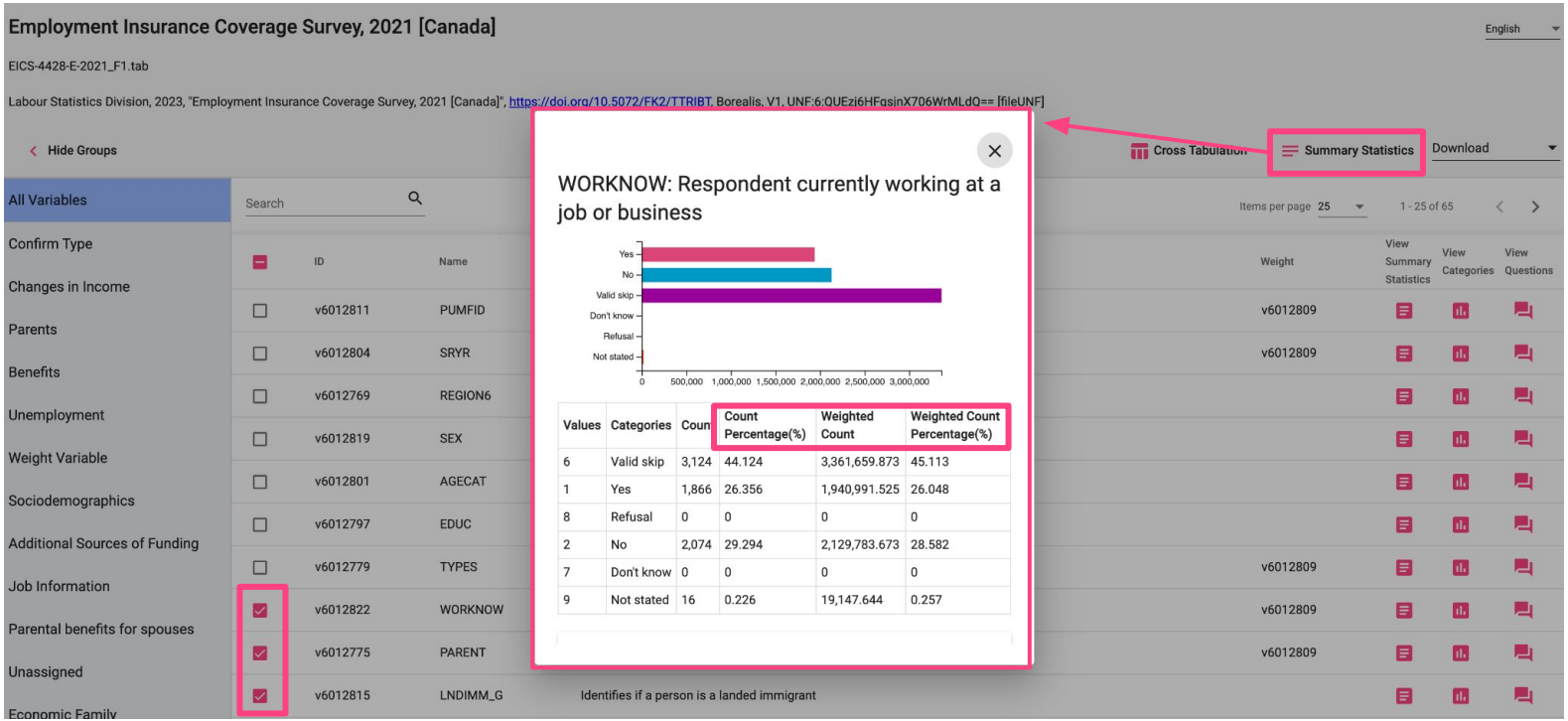


Browse by Variable Groups

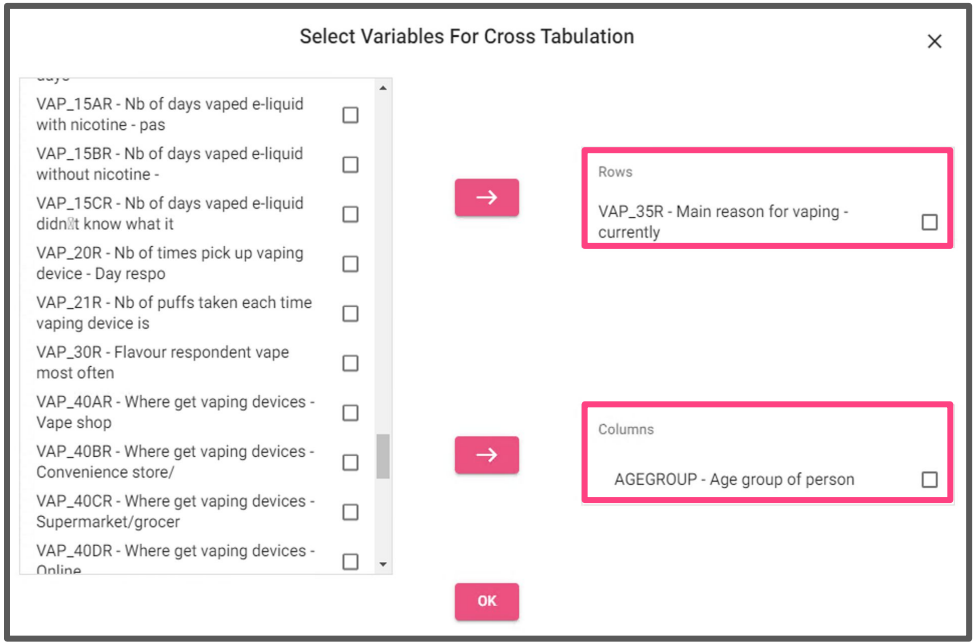
Select / Unselect Variables

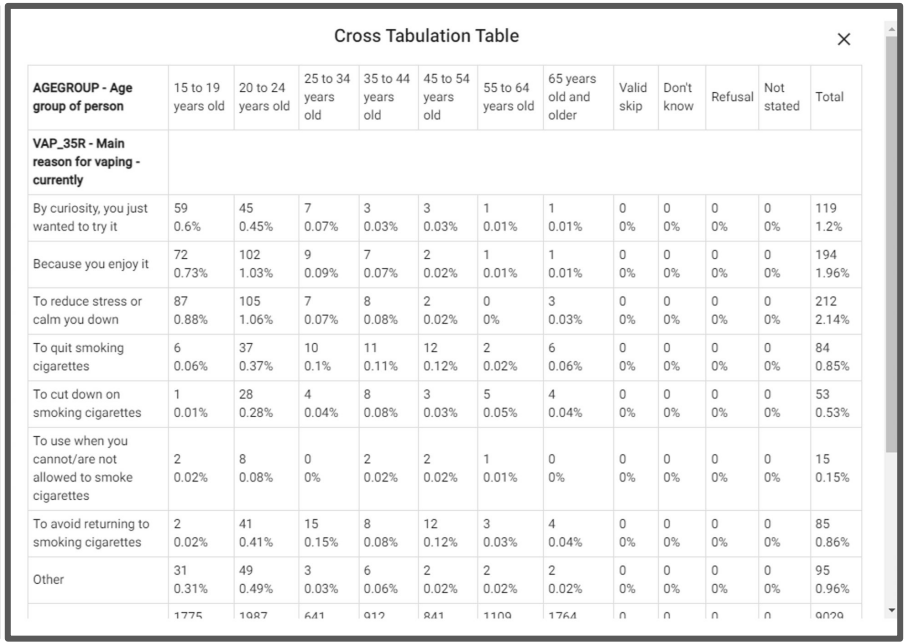
View summary statistics , frequencies, question information

Summary Statistics



Analyse data

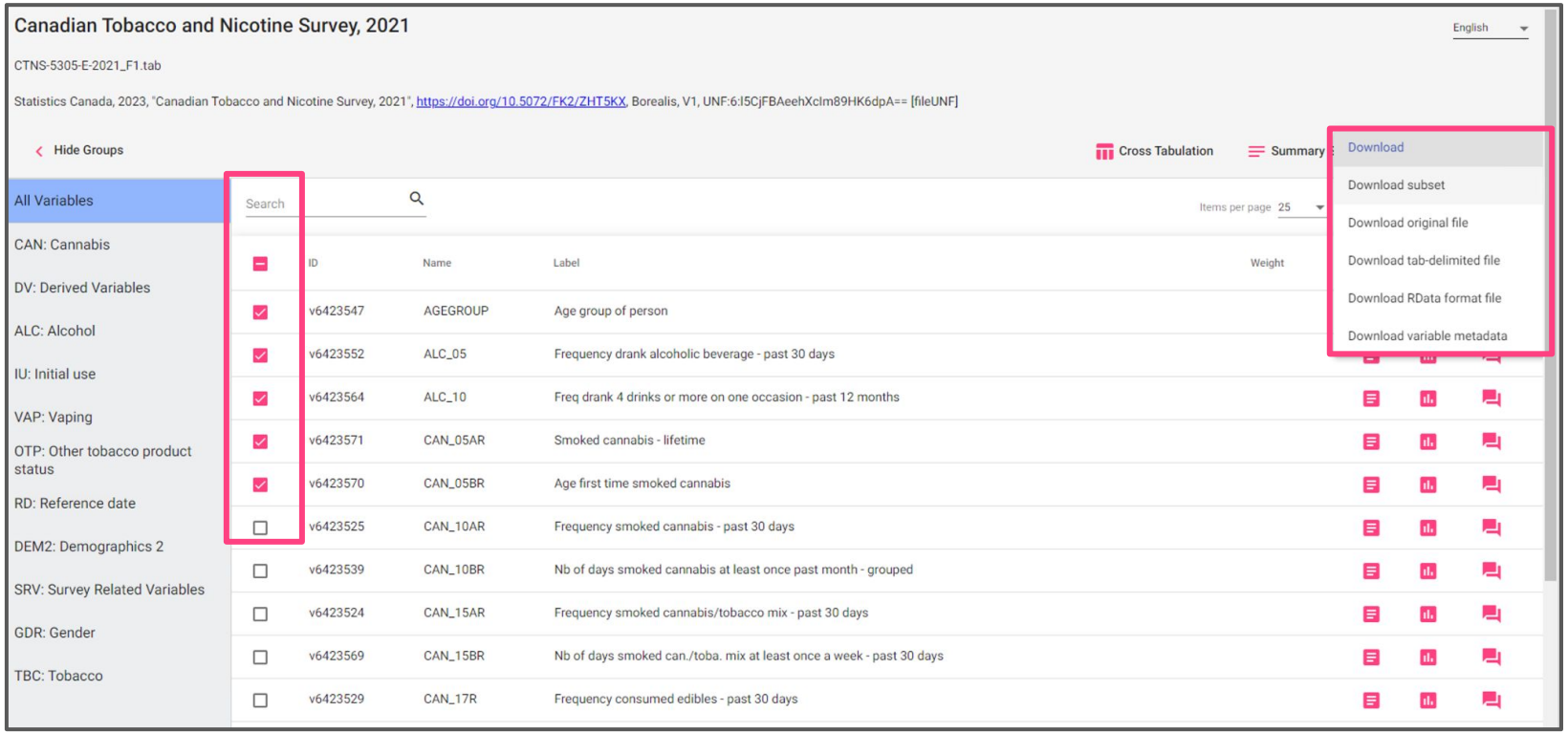




Cross Tabulation Table

AGEGROUP - Age group of person	15 to 19 years old	20 to 24 years old	25 to 34 years old	35 to 44 years old	45 to 54 years old	55 to 64 years old	65 years old and older	Valid skip	Don't know	Refusal	Not stated	Total
VAP_35R - Main reason for vaping - currently												
By curiosity, you just wanted to try it	59 0.6%	45 0.45%	7 0.07%	3 0.03%	3 0.03%	1 0.01%	1 0.01%	0 0%	0 0%	0 0%	0 0%	119 1.2%
Because you enjoy it	72 0.73%	102 1.03%	9 0.09%	7 0.07%	2 0.02%	1 0.01%	1 0.01%	0 0%	0 0%	0 0%	0 0%	194 1.96%
To reduce stress or calm you down	87 0.88%	105 1.06%	7 0.07%	8 0.08%	2 0.02%	0 0%	3 0.03%	0 0%	0 0%	0 0%	0 0%	212 2.14%
To quit smoking cigarettes	6 0.06%	37 0.37%	10 0.1%	11 0.11%	12 0.12%	2 0.02%	6 0.06%	0 0%	0 0%	0 0%	0 0%	84 0.85%
To cut down on smoking cigarettes	1 0.01%	28 0.28%	4 0.04%	8 0.08%	3 0.03%	5 0.05%	4 0.04%	0 0%	0 0%	0 0%	0 0%	53 0.53%
To use when you cannot/are not allowed to smoke cigarettes	2 0.02%	8 0.08%	0 0%	2 0.02%	2 0.02%	1 0.01%	0 0%	0 0%	0 0%	0 0%	0 0%	15 0.15%
To avoid returning to smoking cigarettes	2 0.02%	41 0.41%	15 0.15%	8 0.08%	12 0.12%	3 0.03%	4 0.04%	0 0%	0 0%	0 0%	0 0%	85 0.86%
Other	31 0.31%	49 0.49%	3 0.03%	6 0.06%	2 0.02%	2 0.02%	2 0.02%	0 0%	0 0%	0 0%	0 0%	95 0.96%
	1775	1987	641	912	841	1109	1764	0	0	0	0	9029

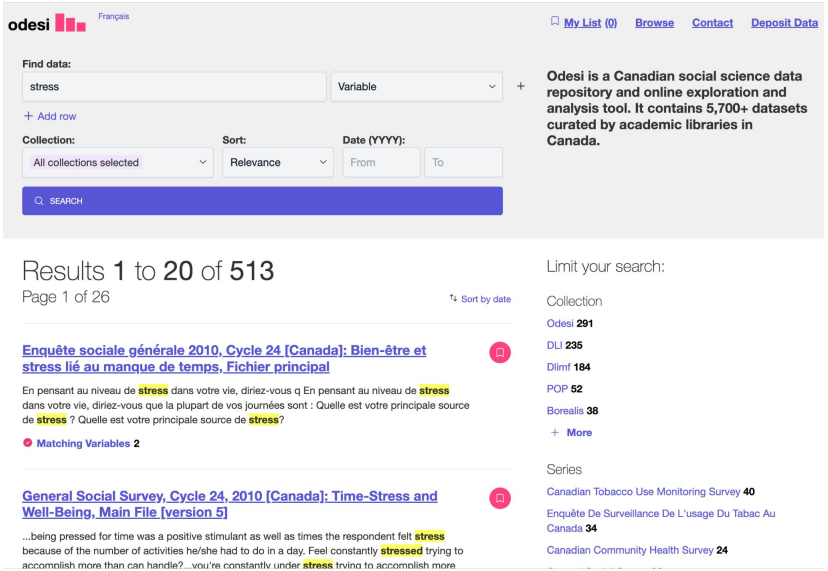
Download and subset



Further Integrations

- Depositing and sharing tabular data in Dataverse
  - Improvements across tabular data ingest and Data Explorer (weighting cross-tabs, exporting charts, more visualizations, Data Documentation Initiative (DDI) support)
- Using Dataverse APIs you can build your own data exploration tools
- Data Explorer is an open, ready to use tool
- Data Explorer can be built into integrated data search sites (e.g.

Data catalogs



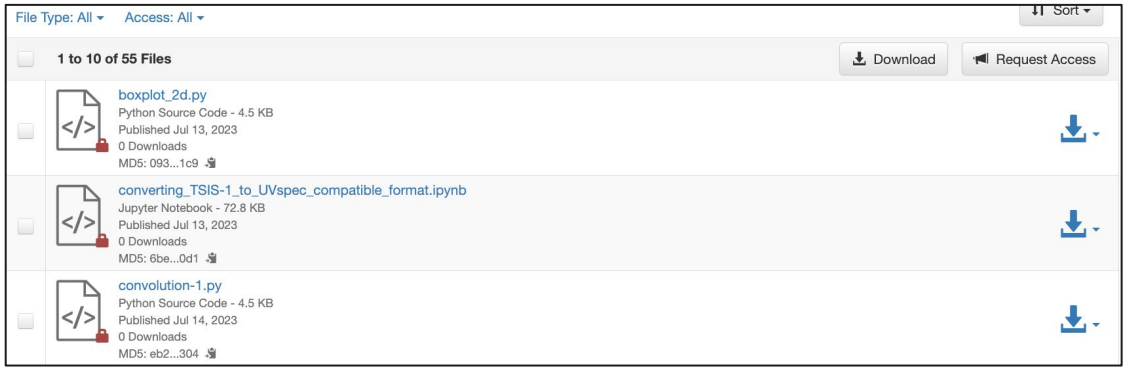
The User Experience: Permissions and Access

Permissions & Access: Flexible Permission System

- Supports multiple workflows by controlling who can deposit files on your behalf, and who can download files
- 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)

Access control:

- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow

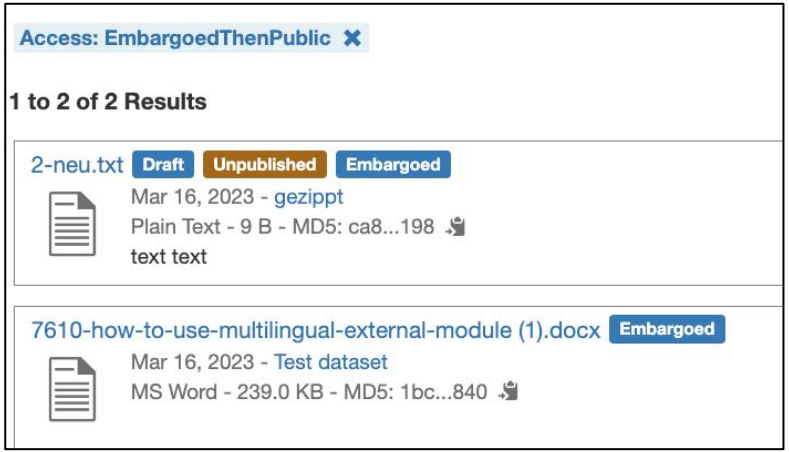


Permissions & Access: Flexible Permission System

- Supports multiple workflows by controlling who can deposit files on your behalf, and who can download files
- 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)

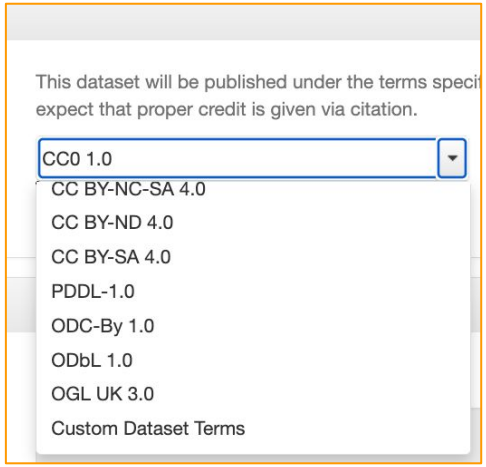
Access control:

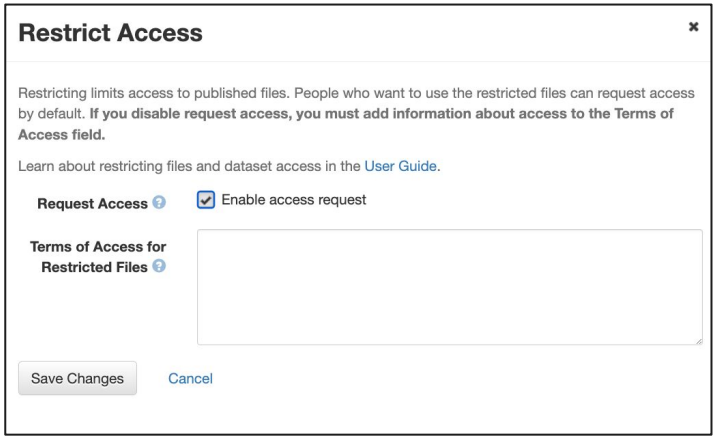
- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed



Permissions & Access: Flexible Permission System

- Supports multiple workflows by controlling who can deposit files on your behalf, and who can download files
- 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)





Access control:

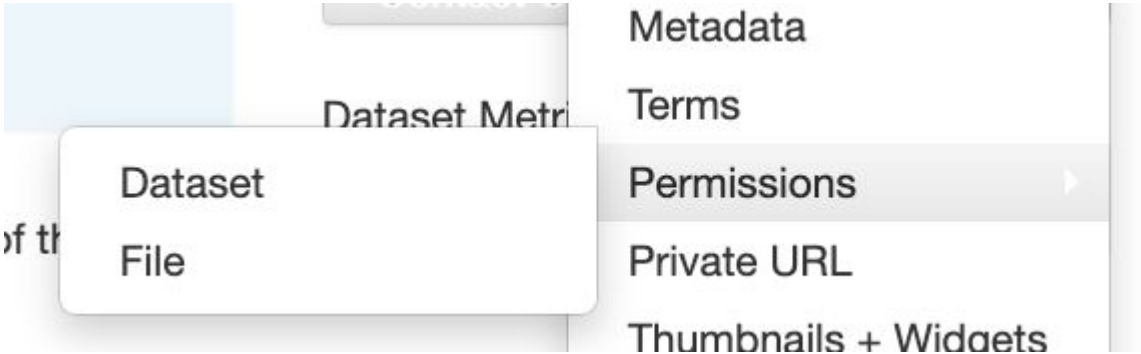
- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed
- Licenses, Custom Licenses

Permissions & Access: Flexible Permission System

- Supports multiple workflows by controlling who can deposit files on your behalf, and who can download files
- 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)

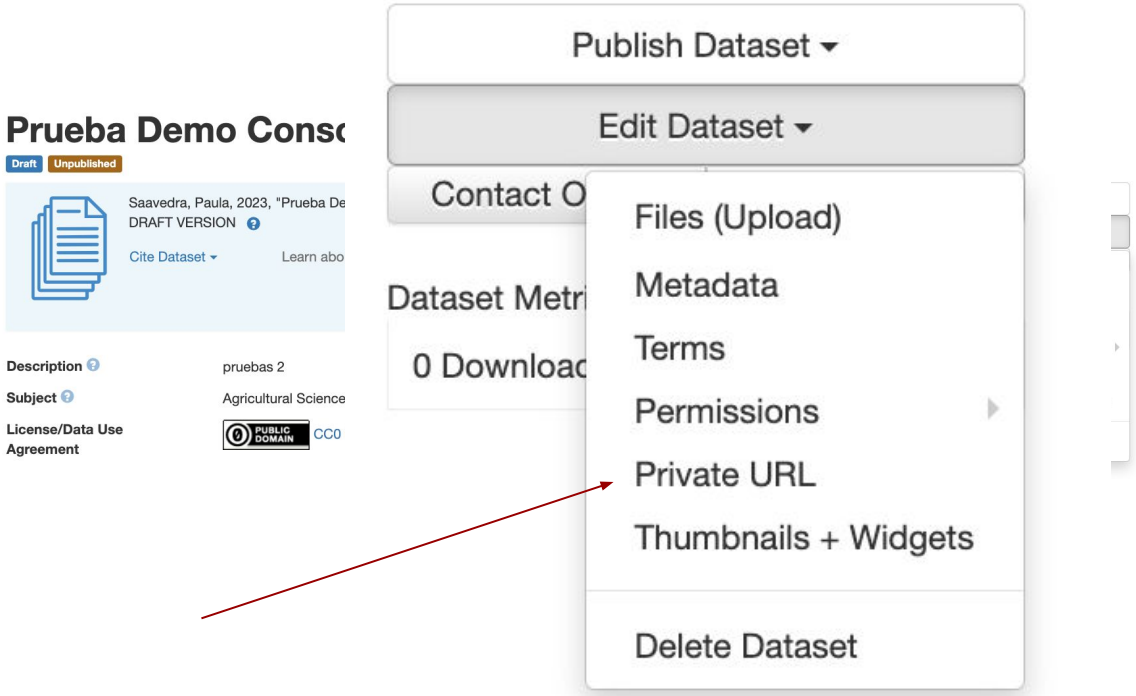
Access control:

- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed
- Licenses, Custom Licenses
- Dataset or File level download options



Permissions & Access: Flexible Permission System

- Supports multiple workflows by controlling who can deposit files on your behalf, and who can download files
- 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)



Access control:

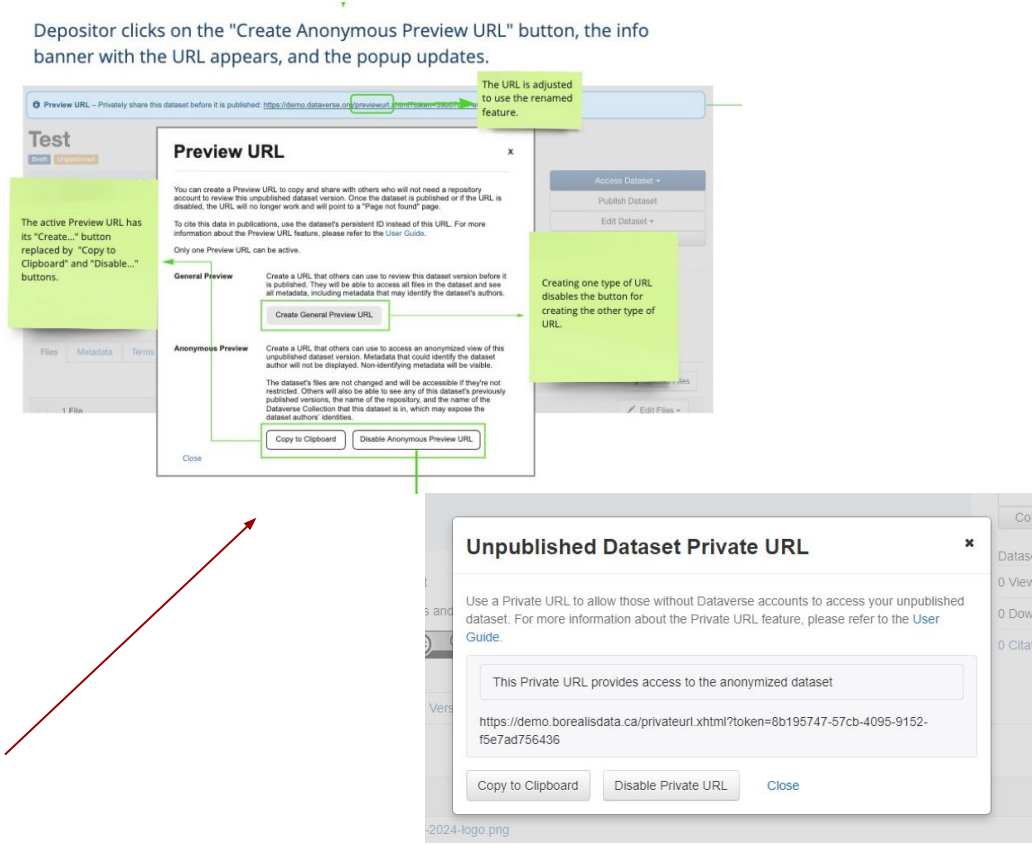
- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed
- Licenses, Custom Licenses
- Dataset or File level download

Permissions & Access: Flexible Permission System

- Supports multiple workflows by controlling who can deposit files on your behalf, and who can download files
- 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)

Access control:

- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed
- Licenses, Custom Licenses
- Dataset or File level download

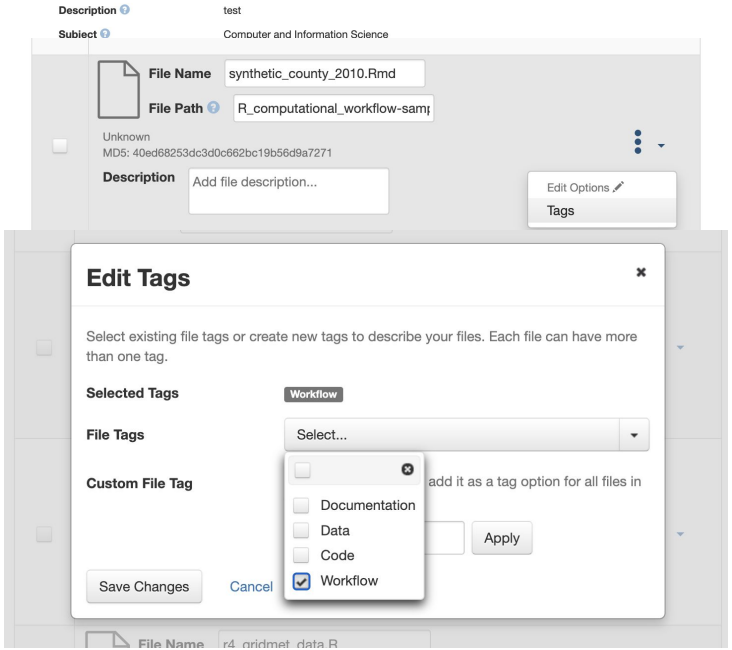


https://github.com/IQSS/dataverse/issues/8185

The User Experience: Data Exploration

Computational workflow and files support

- External tools exist to support **workflows** and **reproducibility**
- specific metadata to better support discoverability (e.g. the new “Dataset Feature” facet)
- Automatic checksum validation on **BagIt** file upload (based on the BagIt manifest)
- **Binder** support (to run Jupyter notebooks) [soon a connection to NERC/MOC]



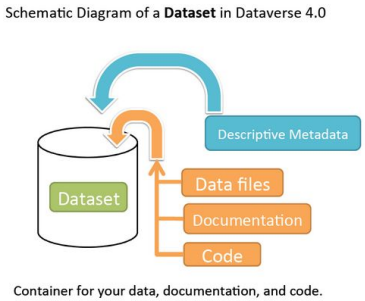
Dataverse & BagIt

- Generates BagIt zip file with complete metadata & all datafiles for a Dataverse dataset
- Conforms to RDA recommendations & includes complete JSON-LD/RDF metadata using OAI-ORE structure
- Imports BagIt packages as datasets, providing round-trip, export/import capability.

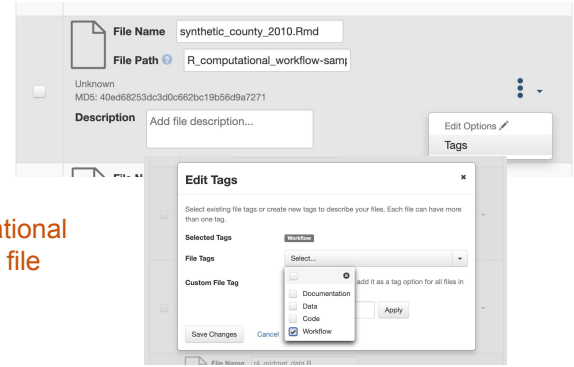
Dataverse Archives Using BagIT

- Texas Data Repository (Duracloud/Chronopolis)
- Qualitative Data Repository (Google Cloud)
- Harvard Data Commons Workflow developed for depositing archival bags into the Harvard DRS repo (not in production yet)

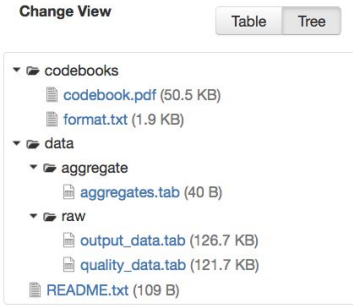
File Types, Format, Documentation



Computational workflow file support



Folder hierarchy support



Tabular data support

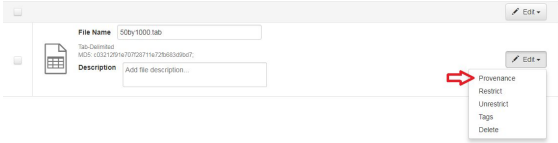
Supported File Formats

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

File level Provenance support



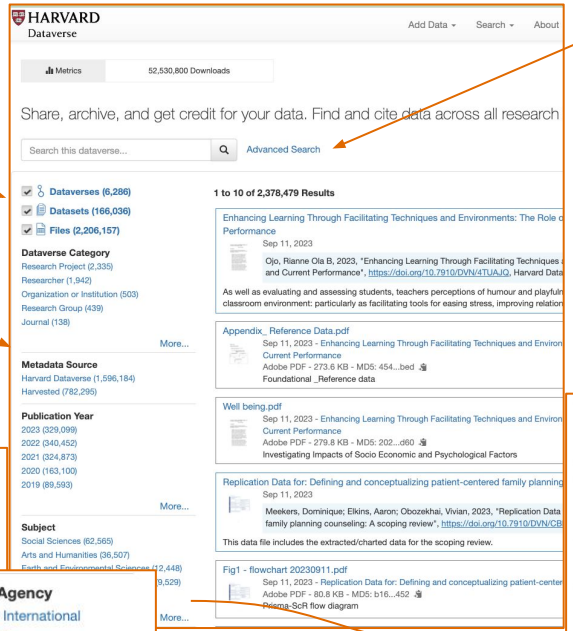
The Dataverse® Project

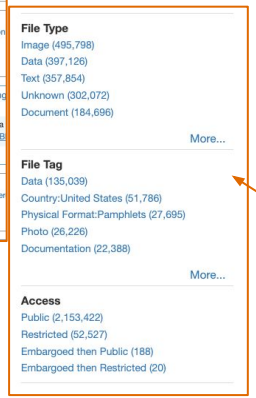
The User Experience: Search and Browse Functionality

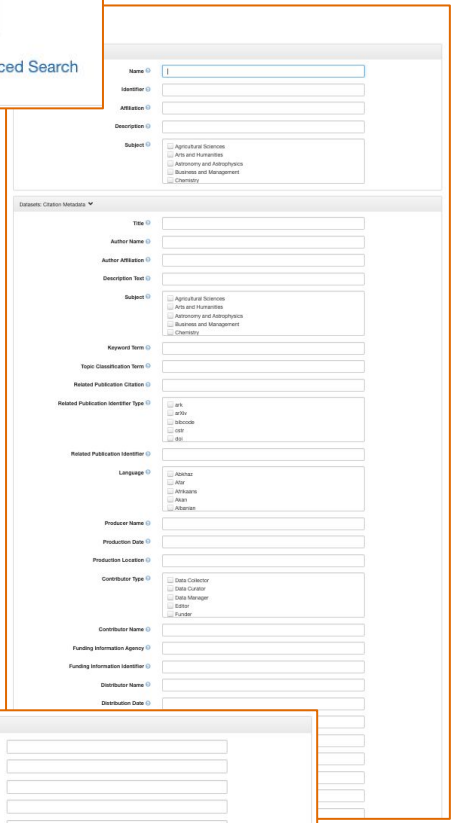
Search & browse

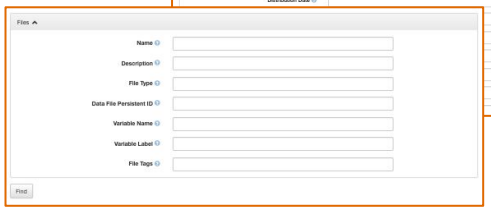
Search:

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









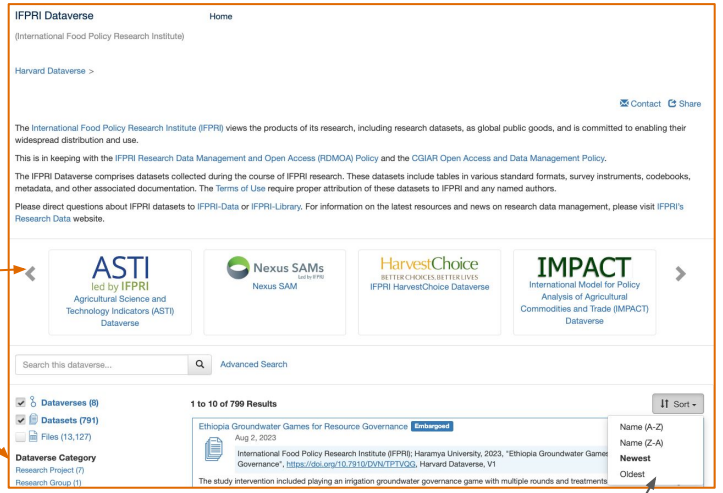
Search & browse

Dataset details:

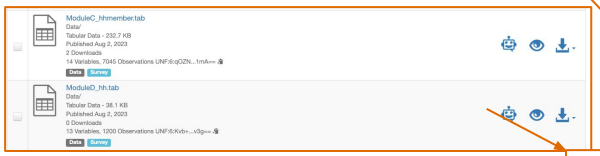
- Detailed dataset page
- Metadata
- Description
- Authors
- Citation
- Download options
- Documentation, readme, code…
- Multiple format download options

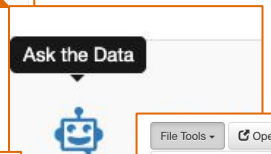
Browse:

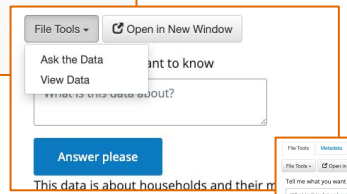
- Categories/Subjects
- Featured Datasets/collections

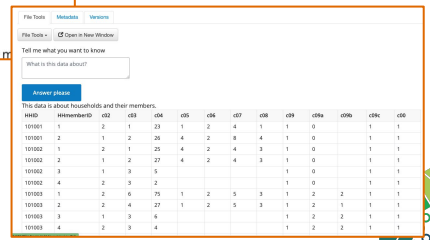


Browse: Recently added









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

References:

Dataverse - Goals, Roadmaps, and Releases

Introduction to Metadata in Dataverse Repositories: For Researchers and Support Staff

Introduction to Data Handling on the Dataverse Platform

Dataverse Software - New Features and Future Plans