Data Handling on The Dataverse Project

Sonia Barbosa, Gustavo Durand, Amber Leahey

Sonia Barbosa
Manager, Harvard Dataverse Repository and Murray Archive

The Dataverse Project
Data Science, IQSS
Harvard University

Gustavo Durand
Technical Lead and Architect of Dataverse

The Dataverse Project
Data Science, IQSS
Harvard University

Amber Leahey
Service Director, Borealis, The Canadian Dataverse Repository / Le dépôt Dataverse canadien

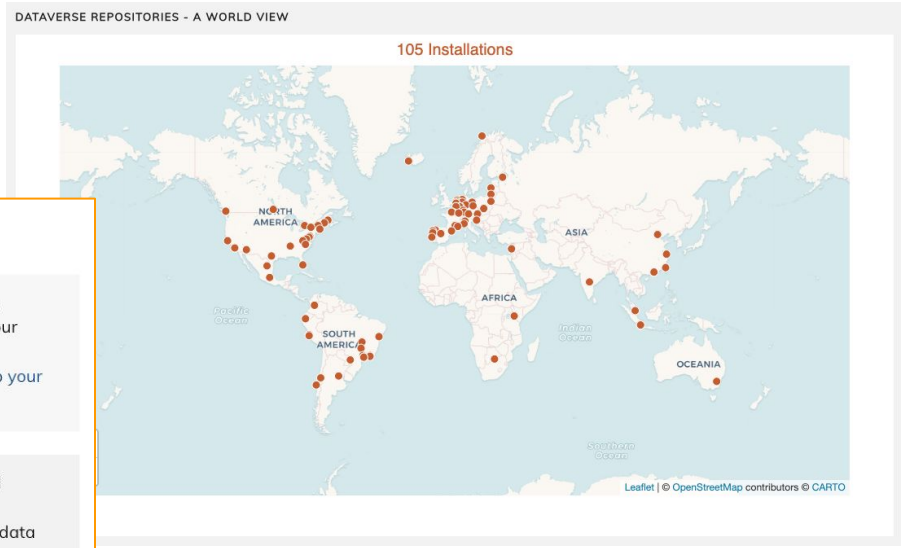
Scholars Portal /University of Toronto Libraries

Introduction

Agenda

- Introduction to Dataverse
- Data Management Principles
- File Handling & Deposit
- Permissions and access
- Overview of File Handling Features
- Management of Various Data Types
- File Versioning
- Files and External Tools / Integrations
- Spotlight: Analysing Survey Data Using Data Explorer
- Thank you & Upcoming Events

Introduction to Dataverse



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 for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools,* or other research and data archival systems with the Dataverse Project. Want to contribute?

Data Management Principles

Data Management Principles

- **Data Governance**: Data Ownership, Stewardship, Policies, Compliance, Ethics
- **Data Quality**: Accuracy, Consistency, Completeness, Reliability, Timeliness, Validity
- **Data Security:** Access control, Encryption, Authentication, Authorization, Auditing and Monitoring, Data Masking
- **Data Privacy:** Consent Management, PII protection, GDPR, HIPAA compliance
- **Data Lifecycle management:** Creation, Storage, Archiving, Disposal, Retention, Versioning
- **Data Integration:** ETL (Extract, Transform, Load), Data Connectors, Data Warehouse, Data Transformation, Data Pipelines
- **Data management:** Consolidation, synchronization, deduplication, governance
- **Metadata management:** Data description, data lineage, data catalog, data dictionary, data taxonomy, data versioning
- **Data backup and recovery:** Regular backup, disaster recovery plan, backup and recovery testing, business continuity plan, data redundancy
- **Data access and retrieval:** Querying/reporting, access APIs, Visualizations, self service and search capabilities
- **Data monitoring and auditing:** Real-time monitoring, Log analysis, Compliance checks, Anomaly detection, Audit trails, Security incident response
- **Data documentation:** Data models, Data schemas, Data dictionaries, Data lineage diagrams, Data change history, Data quality reports

Data Management Principles and Dataverse

Data Organization: Repository access, folder support, and storage for different types of data files and associated metadata.

Data Documentation: Proper metadata and documentation

Metadata Management: Comprehensive metadata to describe their datasets (study level and file level).

Data Sharing: Publish datasets, assign persistent identifiers (DOIs), and specify access rights.

Data Security: Controlling access to data through permissions and embargos to ensure data security and compliance with privacy regulations.

Data Preservation: Long-term data preservation, ensuring that data remains accessible and usable over time.

While providing training and education on maintaining data quality, and providing tools for interoperability and integration.

File Handling and Deposits - Dataverse

Introduction to File Handling & Deposit - Dataverse

- Know **what to share**:
  - Research projects
  - Scientific data outputs
- **Organize files** for deposit
- File **types** and **formats**
- File **upload** options
- File **sizes**
- File-level **metadata support**
  - File naming, file descriptions, file-level tags, file-level edits

File types and formats

Text Data Formats:

- Plain Text
- CSV

Tabular Data Formats:

- CSV (Comma-Separated Values)
- TSV (Tab-Separated Values)
- XLSX (Microsoft Excel)
- ODS (OpenDocument Spreadsheet)
- Excel CSV (Excel-Formatted CSV)
- SPSS
- Stata
- Parquet (for large files)

Geospatial Data Formats:

- GeoJSON: A format for encoding geospatial data using JSON.
- Shapefile: A widely used format for geospatial vector data.

Scientific Data Formats:

- NetCDF (Network Common Data Form): A format for storing scientific data, commonly used in environmental and climate sciences.
- HDF5 (Hierarchical Data Format): A versatile data format used in various scientific fields.

Image Formats:

- PNG (Portable Network Graphics): A widely supported format for lossless image compression.
- SVG (Scalable Vector Graphics): An XML-based format for vector graphics.

Audio and Video Formats:

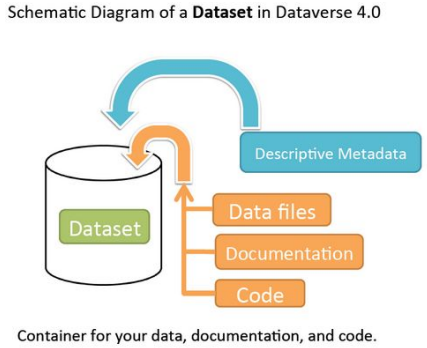
- OGG (Ogg Vorbis): Open audio compression format.
- WebM: An open video format designed for the web.

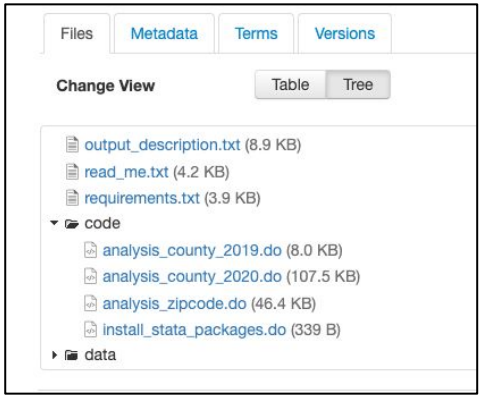
Compression Formats:

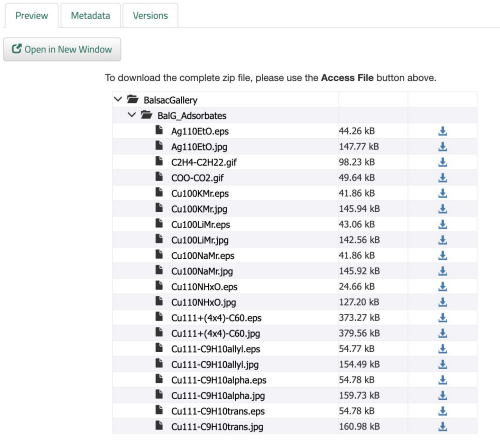
- GZIP: Open compression format for files and data streams.
- 7-Zip: An open, high-compression archive format.

Organize files for deposit

- Use descriptive file names
- Use a clear folder structure
- Use standard file formats
- Include documentation
- Use version control
- Clean the data
- Include metadata
- Use folder subdirectories
- Avoid special characters
- Use consistent file naming and formatting
- Partition if possible
- Ensure security and privacy
- Set appropriate file level permissions and access control
- Test what you have shared and ensure files open and download as expected
- Provide your contact information for your data users
- Specify the terms and conditions for reuse







File sizes and upload options and examples

File formats:

- All formats accepted

Per file size limitations*:

- 2.5GB
- Larger sizes with permission
- Max 1000 files per upload

Tabular Ingest*:

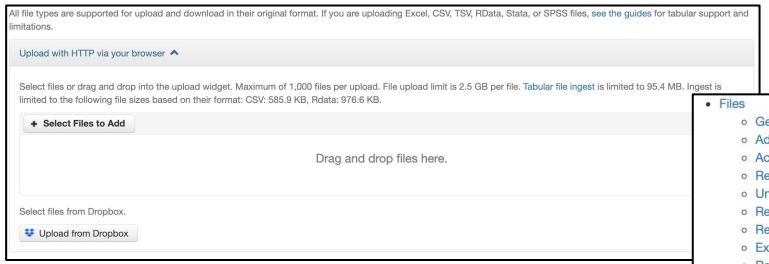
- 95.4 MB
- CSV 585.9KB
- Rdata 976.6KB

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)



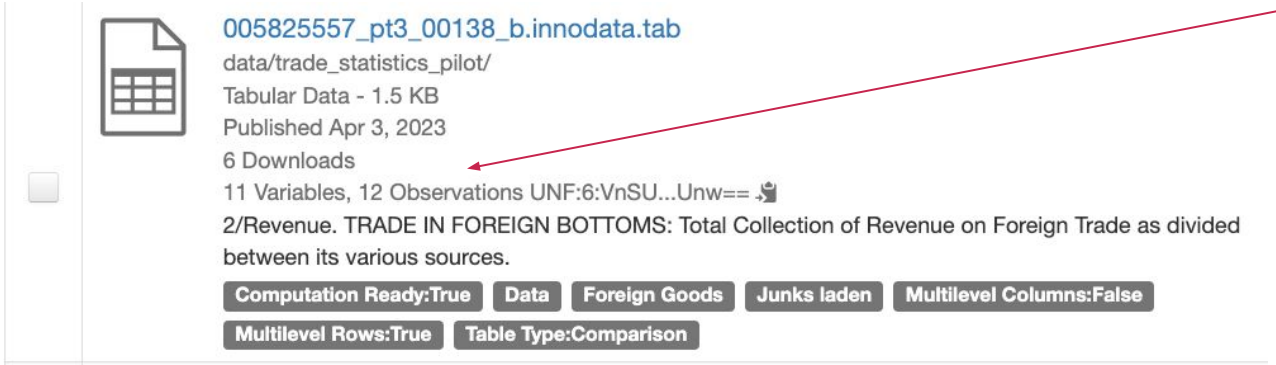
File upload options

Native APIs file handling options

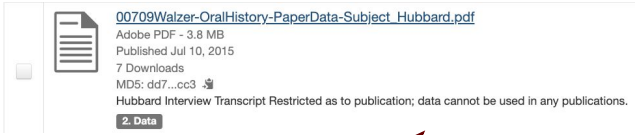
- Files
  - Get JSON Representation of a File
  - Adding Files
  - Accessing (downloading) files
  - Restrict Files
  - Uningest a File
  - Reingest a File
  - Redetect File Type
  - Extract NcML
  - Replacing Files
  - Deleting Files
  - Getting File Metadata
  - Updating File Metadata
  - Editing Variable 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



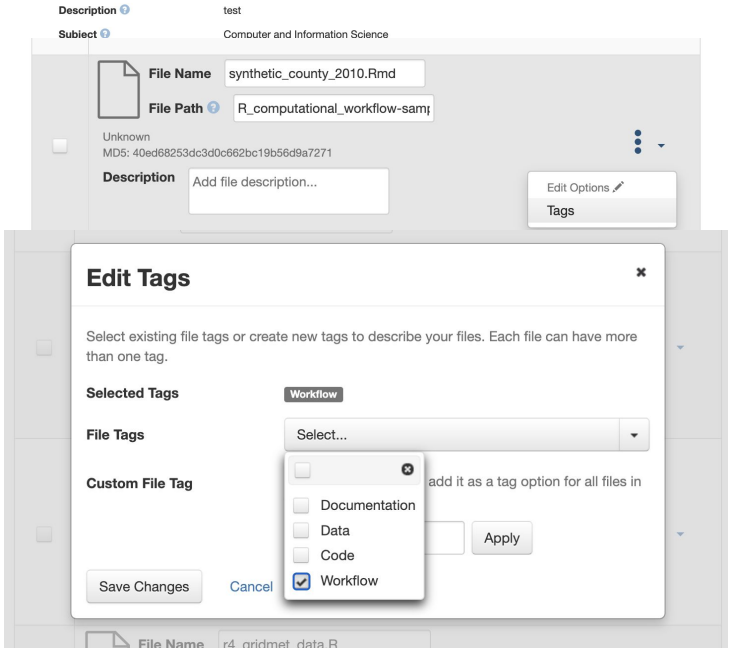
File level "terms"

File level "tags"

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

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]



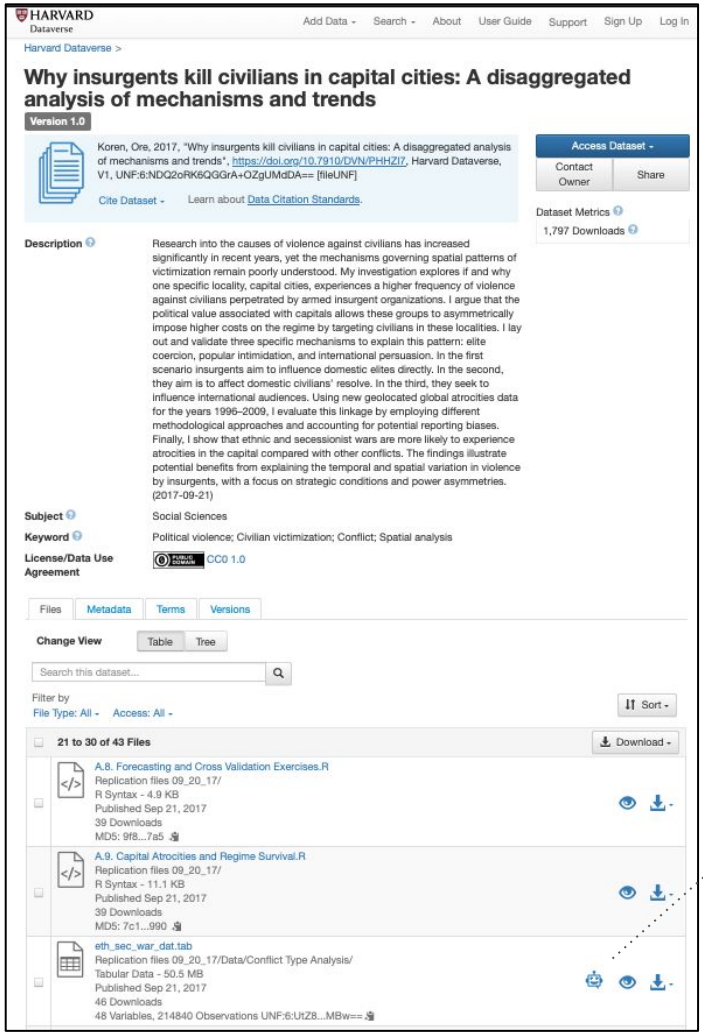
Dataverse & BagIt

- 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)

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

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:nUddO4V8XEZZCnhI+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

What is this data about?

Answer please

This data is about latitude, longitude, and event information.

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

Cool but poor

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

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

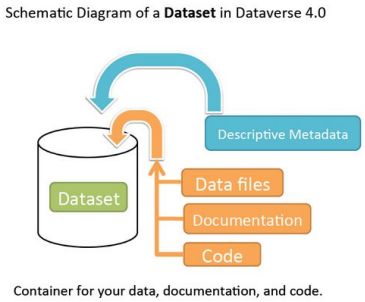
The countries with the highest number of events in the given range of latitudes and longitudes are Afghanistan, Iran, and Pakistan.

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

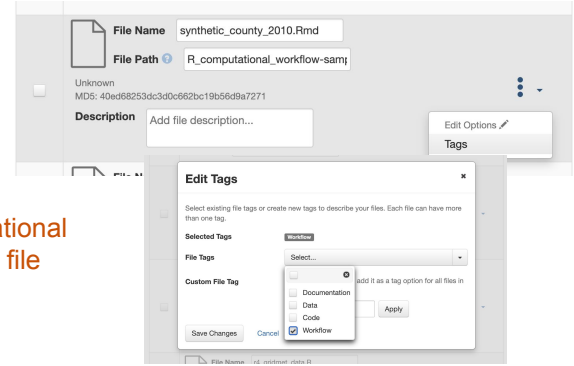
LLM kicks in

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

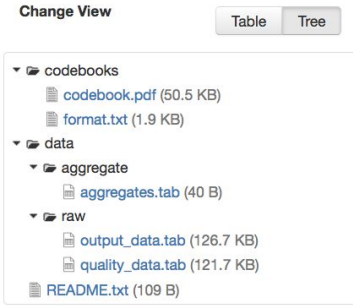
File Types, Format, Documentation



Computational workflow file support



Folder hierarchy support



Tabular data support

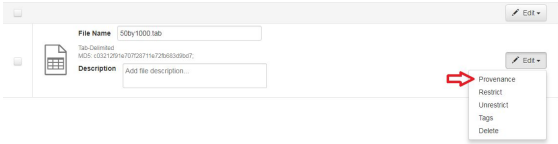
Supported File Formats

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

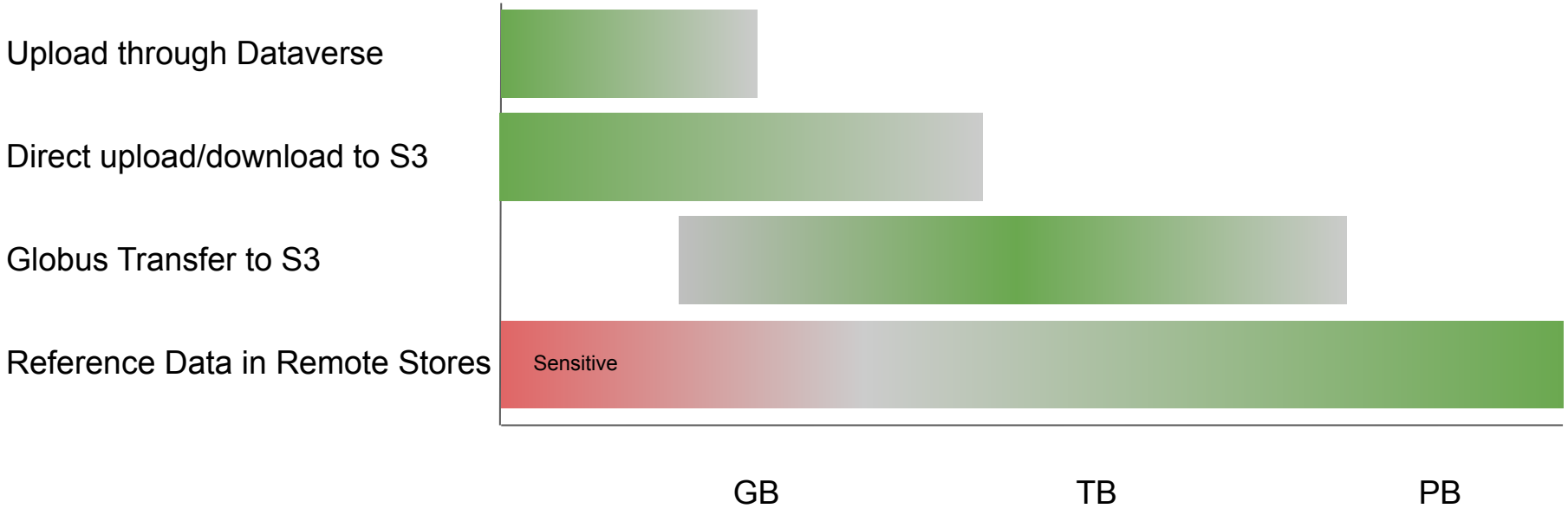


The Dataverse® Project

Coming soon…

Big data support in Dataverse

(overview from 2022)



We are now moving to use Globus API (rather than S3) for store interactions

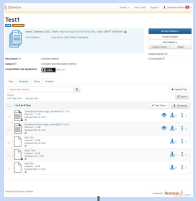
Globus Transfer To Dataverse

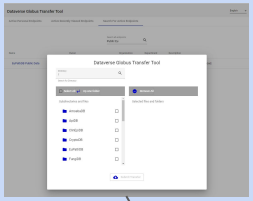
Researcher's Browser

Dataverse Dataset

launch

Transfer **In/Out or Reference**





Globus Endpoints

reliable parallel transfer

Dataverse Server

Dataverse-Globus Transfer App

Globus Store

S3 Store

File Store

Remote Store

Managed Globus Endpoint (e.g. over tape storage)

Globus Service

Considerations

- Storage costs
- Initial data transfer costs
- File access costs
- File size limitations
- File count limitations
- Access protocol
- Backups
- User-data and application-data interaction requirements
- Dataset representation
- Usage monitoring
- Admin overhead

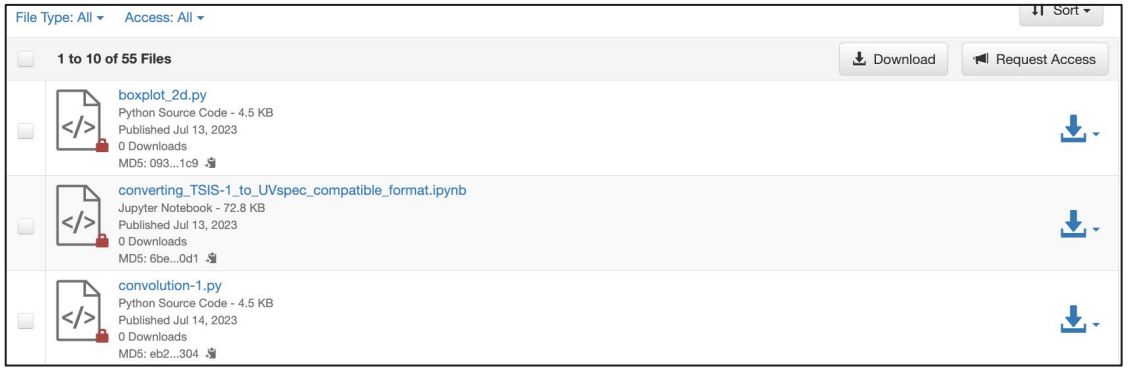
Permissions and Access

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

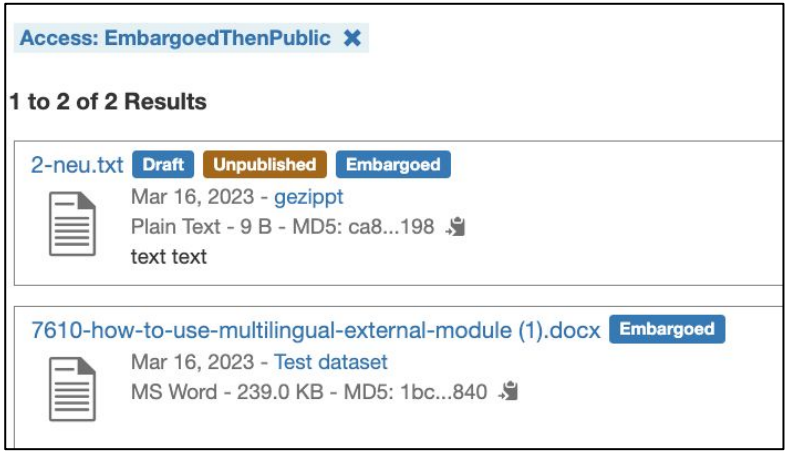


Permissions & Access: Flexible Permission System

- 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)

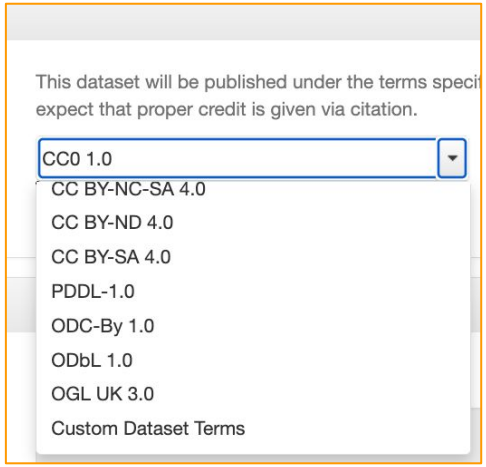
Access control:

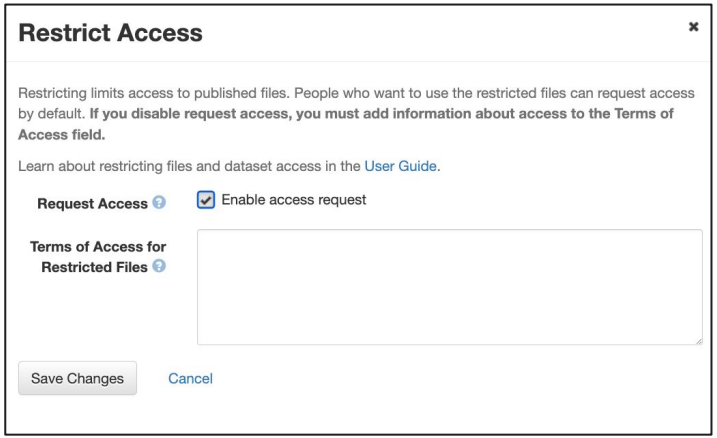
- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed



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:

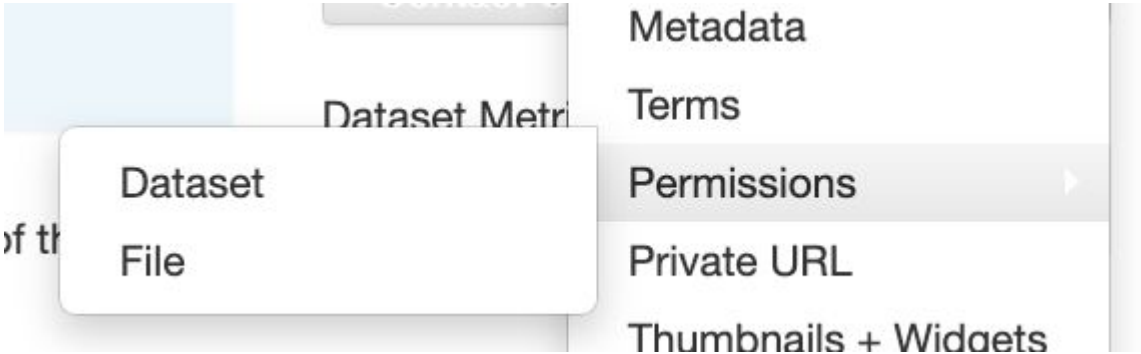
- Custom terms
- Restricted data must clearly define terms of access
- “Request access” workflow
- Public, Restricted, Embargoed
- Licenses, Custom Licenses

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

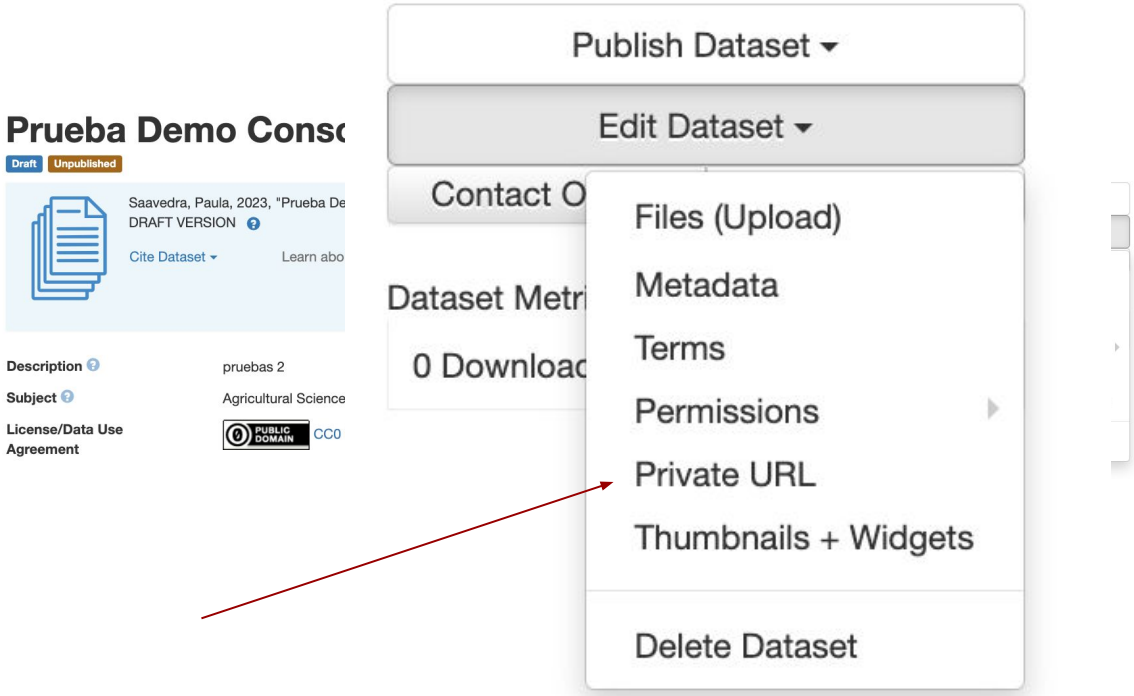


Permissions & Access: Flexible Permission System

- 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 options
- Private URL for DRAFT dataset

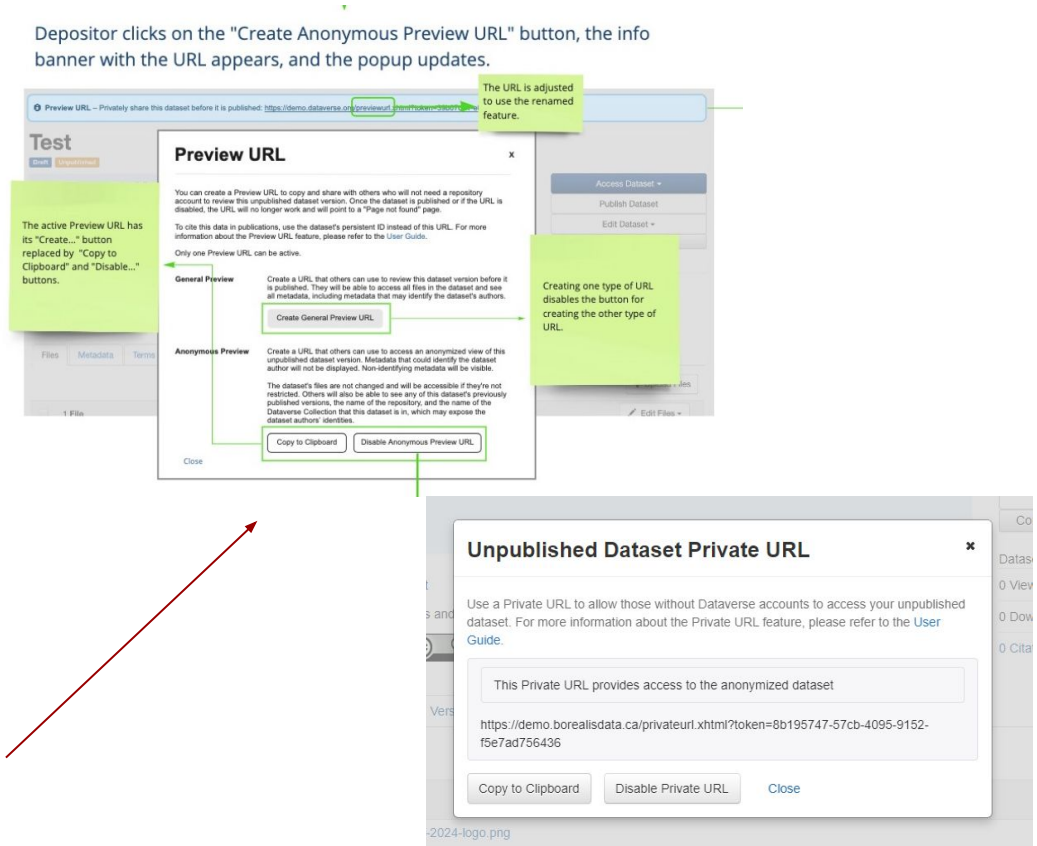


Permissions & Access: Flexible Permission System

- 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
- Private URL for DRAFT dataset
- Anonymous Peer Review



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

Enhanced File Handling of Data Types

File Handling Checklist

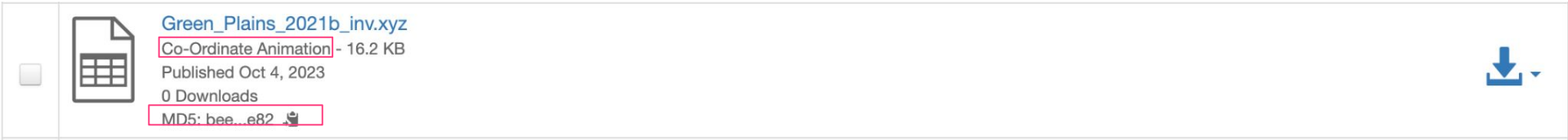
Enhanced file handling and data management features include:

- Deposit & sharing (file identification, checksum (MD5), packaging)
- File Preservation (normalization)
- Tabular data ingest (tabular data formats)
- Geospatial and dimensional data (GeoJSON, Shapefile, GeoTIFF, NetCDF & HDF5)
- Additional file metadata extraction (FITS, NetCDF (NcML) and others)
- File previewers & data analysis integrations

Dataverse Format Checklist	Deposit & Sharing (Identification, MD5 checksum)	Preservation (Normalization)	Exploration (Preview and extraction)	Analysis (Tools & integrations)
CSV	Y	Y	Y	Y
TXT	Y	Y	Y	Y
MP4	Y	N	Y	Y
TIFF	Y	N	N	N
Quicktime (.MOV)	Y	N	Y	Y
Excel (XLSX)	Y	Y	Y	Y
SPSS	Y	Y	Y	Y
STATA	Y	Y	Y	Y
SAS	Y	N	N	N
Shapefile	Y	N	Y	Y
GeoTIFF	Y (beta)	N	Y (beta)	Y

File Preserveration

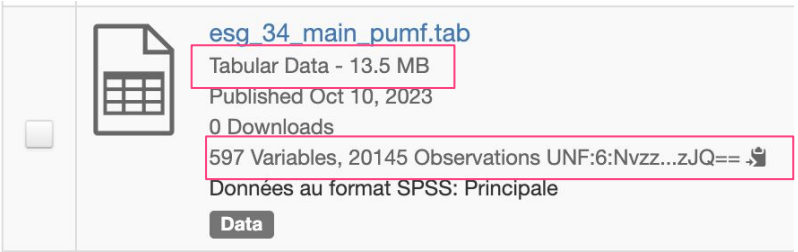
- Format identification (MIME Type)
- Checksum (MD5)
- Format normalization
- File metadata
- Folder preservation
- Exporting options for preservation system integration (BagIt, Archivematica)



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

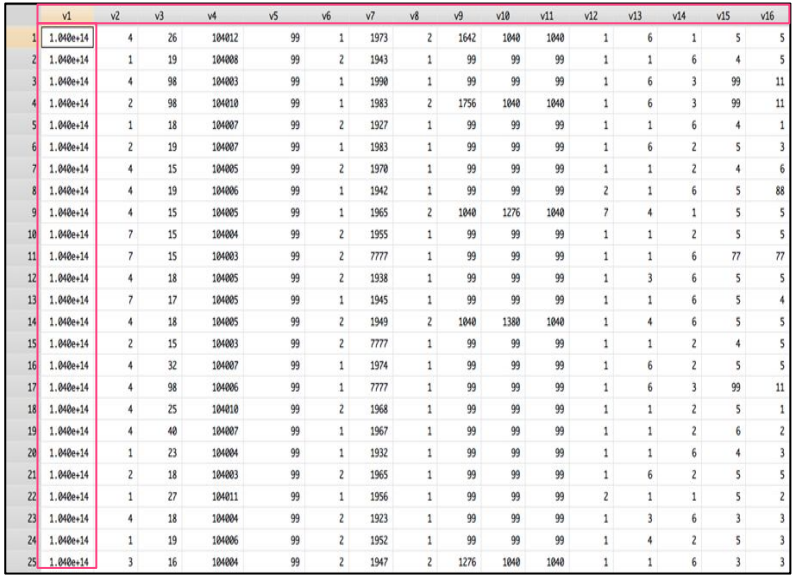


Reuses and integrations

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

Tabular data: depositing files

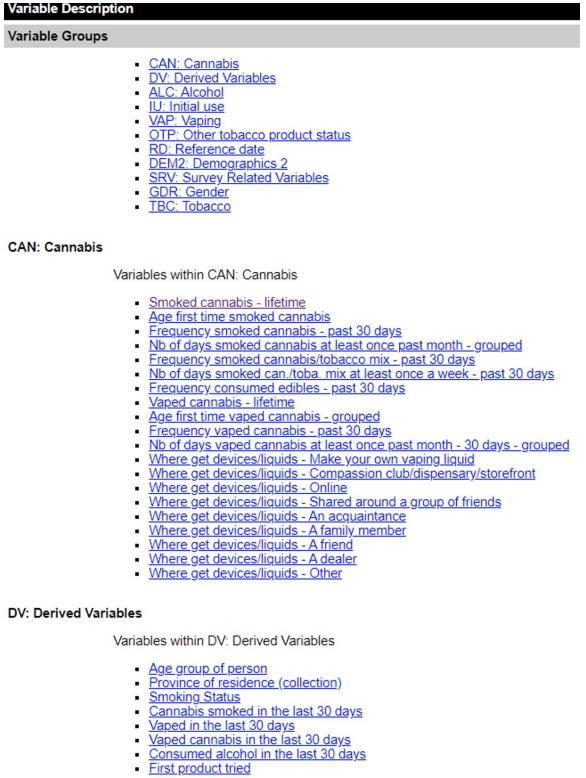
- Upload tabular files (SPSS, STATA, R, CSV, Excel)
- Data ingest tips:
  - Responses/ observations as rows
  - Variables/concepts as columns
  - No notes / improper formatting
  - Does not support multiple spreadsheets in one file
  - Excel (XLSX) is supported
  - CSV requires documentation
  - Encode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



Variable Description

Variable Groups

- CAN: Cannabis
- DV: Derived Variables
- ALC: Alcohol
- IU: Initial use
- VAP: Vaping
- OTP: Other tobacco product status
- RD: Reference date
- DEM2: Demographics 2
- SRV: Survey Related Variables
- GDR: Gender
- TBC: Tobacco

CAN: Cannabis

Variables within CAN: Cannabis

- Smoked cannabis - lifetime
- Age first time smoked cannabis
- Frequency smoked cannabis - past 30 days
- Nb of days smoked cannabis at least once past month - grouped
- Frequency smoked cannabis/tobacco mix - past 30 days
- Nb of days smoked can./toba. mix at least once a week - past 30 days
- Frequency consumed edibles - past 30 days
- Vaped cannabis - lifetime
- Age first time vaped cannabis - grouped
- Frequency vaped cannabis - past 30 days
- Nb of days vaped cannabis at least once past month - 30 days - grouped
- Where get devices/liquids - Make your own vaping liquid
- Where get devices/liquids - Compassion club/dispensary/storefront
- Where get devices/liquids - Online
- Where get devices/liquids - Shared around a group of friends
- Where get devices/liquids - An acquaintance
- Where get devices/liquids - A family member
- Where get devices/liquids - A friend
- Where get devices/liquids - A dealer
- Where get devices/liquids - Other

DV: Derived Variables

Variables within DV: Derived Variables

- Age group of person
- Province of residence (collection)
- Smoking Status
- Cannabis smoked in the last 30 days
- Vaped in the last 30 days
- Vaped cannabis in the last 30 days
- Consumed alcohol in the last 30 days
- First product tried

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

Geospatial Data

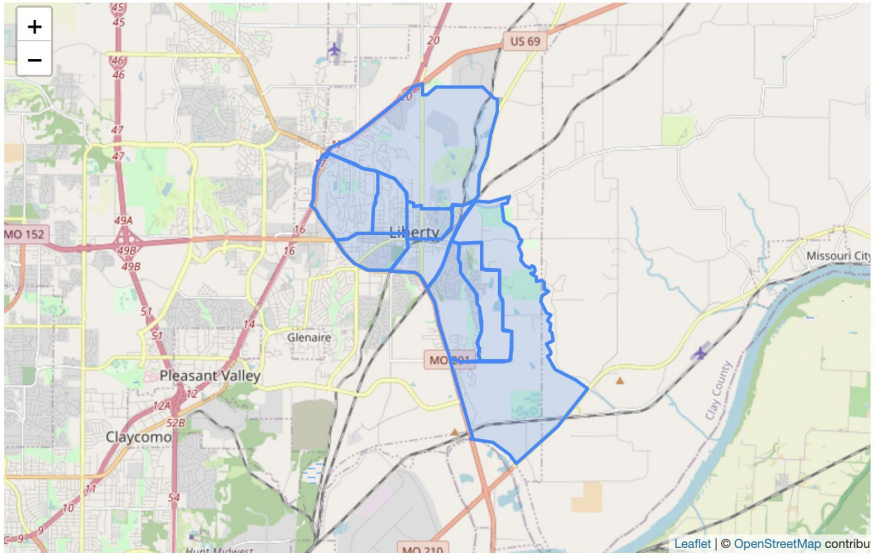
cd_2006_orig_CBF.zip
Shapefile as ZIP Archive - 27.2 MB
Published Apr 4, 2023
0 Downloads
MD5: c0e...22e
UNI-CEN Digital Boundary File for Census Division (CD), 2006, in Esri Shapefile format and NAD83 CSRS / EPSG:3348 projection. This file is part of the CBF-Original Shorelines series. Retrieved from: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

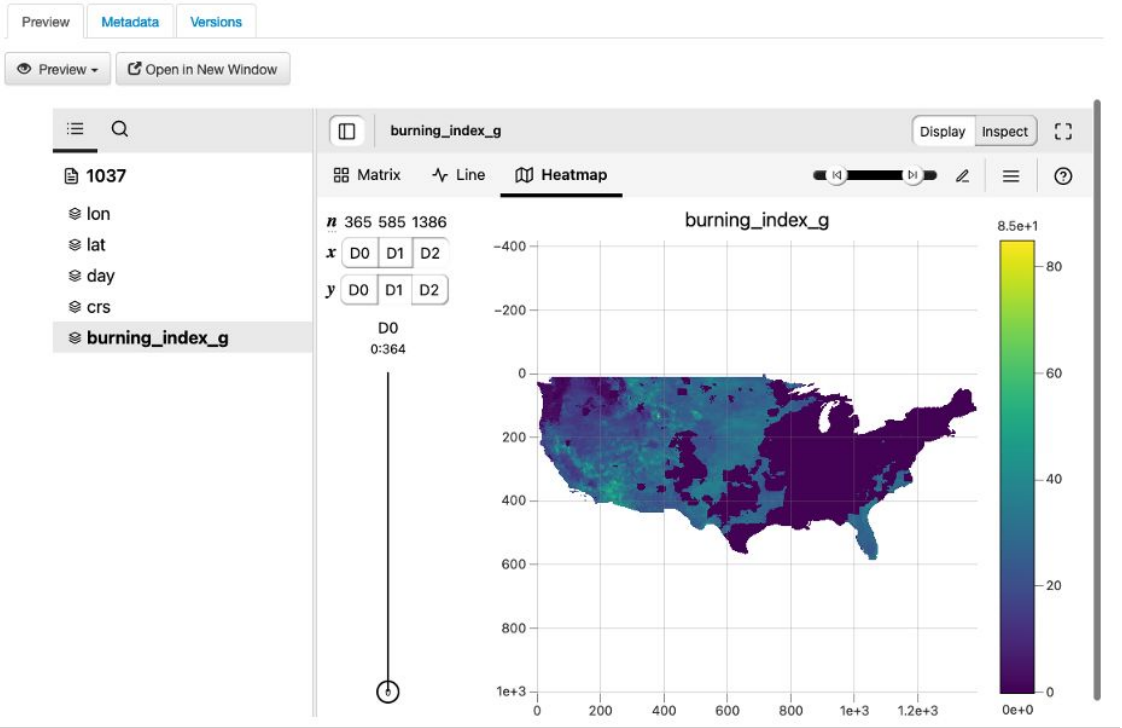
Open in New Window



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

NetCDF & HDF5 Previewer

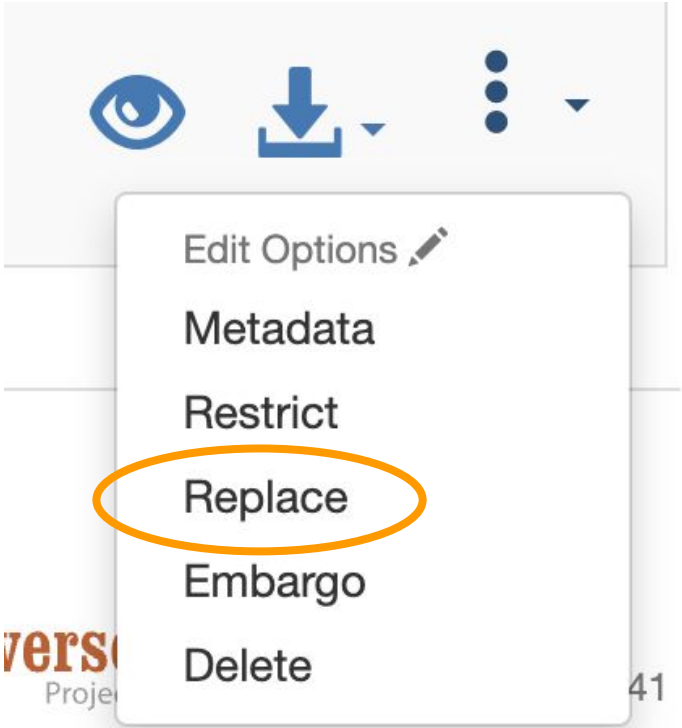
- Common scientific data formats
- Metadata extracted in NcML (XML) format and saved as an auxiliary file
- H5Web visualization and file preview tool
  - Preview data
  - Explore and visualize data



File Versioning

Versioning

- A dataset begins as an unpublished DRAFT; when published becomes v1.0
- Future versions can be minor (e.g. v1.1) or major, (e.g v2.0)
- File metadata can be changed per version, as needed
- File “versioning” is provided via file replacement - when uploading a new version of the file, it is linked to the previous version



Persistent Identifiers

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

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:

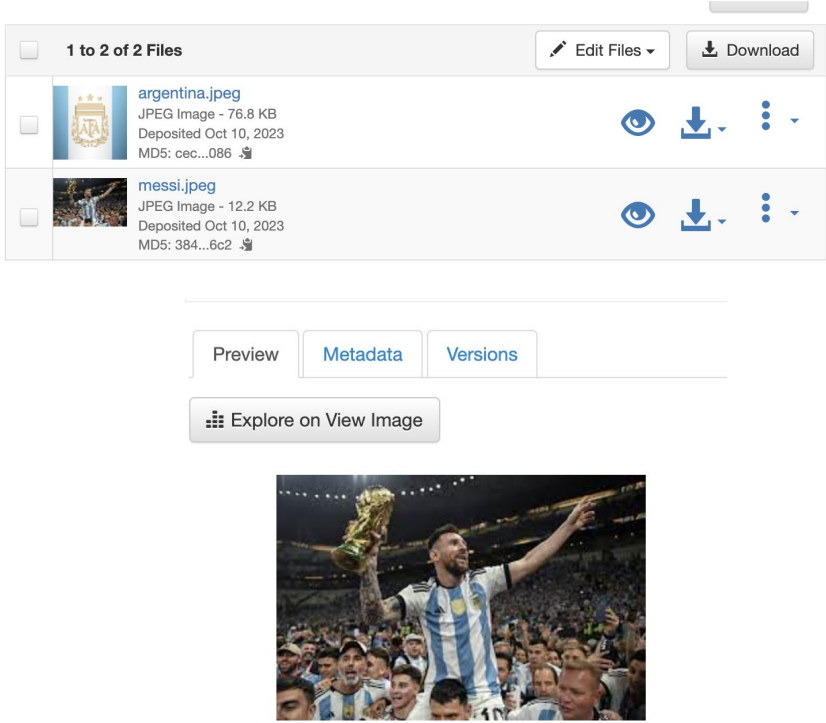
External Tools Manifest

- External tools must be expressed in an external tool manifest file
- Can be uploaded to a Dataverse installation via API

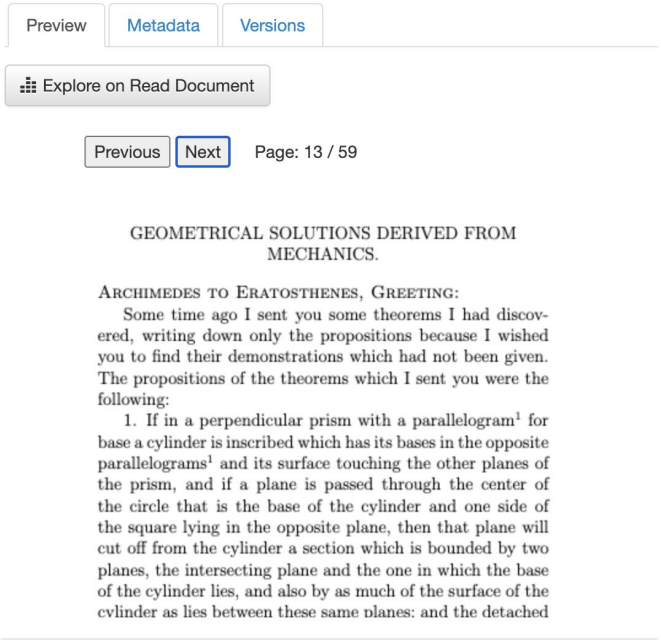
```
{
  "displayName": "Awesome Tool",
  "description": "The most awesome tool.",
  "type": "explore",
  "toolUrl": "https://awesometool.com",
  "toolParameters": {
    "queryParameters": [
      {
        "fileid": "{fileId}"
      },
      {
        "key": "{apiToken}"
      }
    ]
  }
}
```

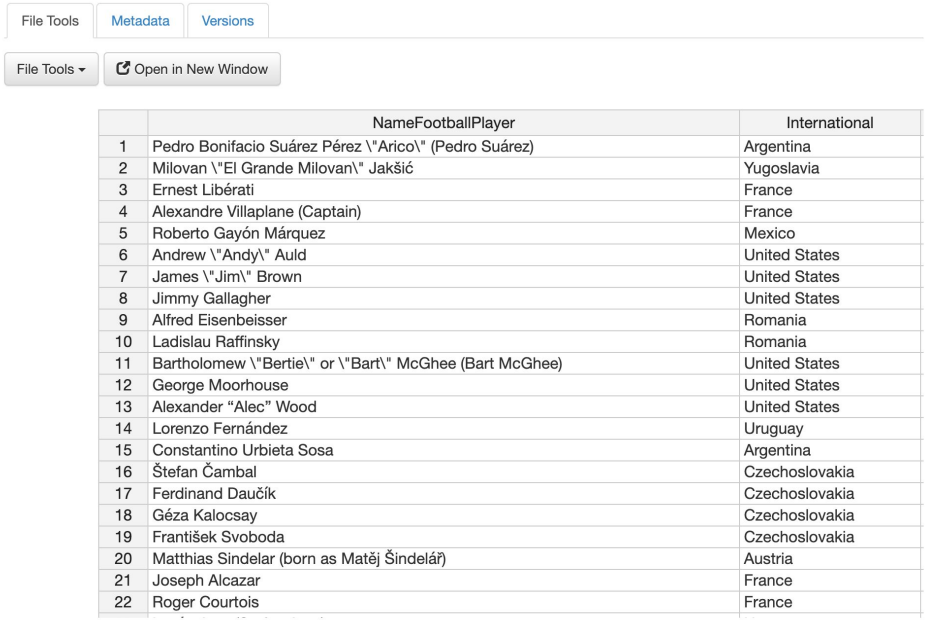
File Previewers

- 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
  - GeoJSON
  - Zip files
  - NcML files
- Previewers are available through the preview (eye) icon on Dataset pages
- And also embedded as a tab on Datafile pages



File Previewers (more examples)

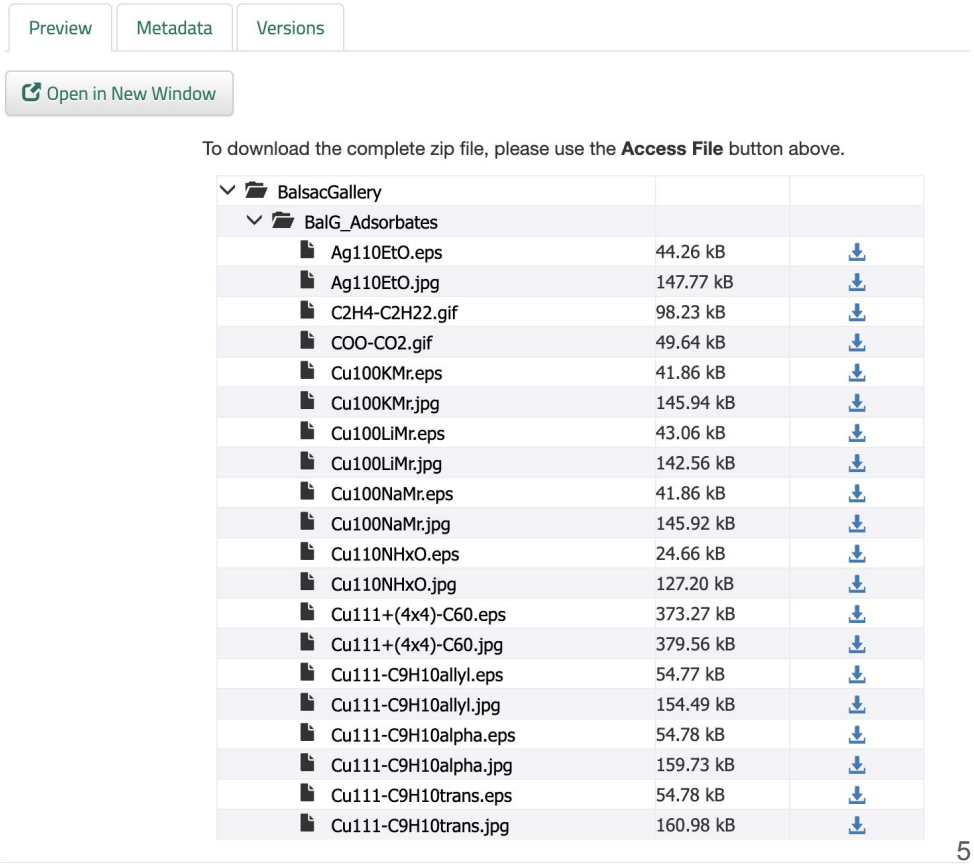




	NameFootballPlayer	International
1	Pedro Bonifacio Suárez Pérez \"Arico\" (Pedro Suárez)	Argentina
2	Milovan \"El Grande Milovan\" Jakšić	Yugoslavia
3	Ernest Libérati	France
4	Alexandre Villaplane (Captain)	France
5	Roberto Gayón Márquez	Mexico
6	Andrew \"Andy\" Auld	United States
7	James \"Jim\" Brown	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

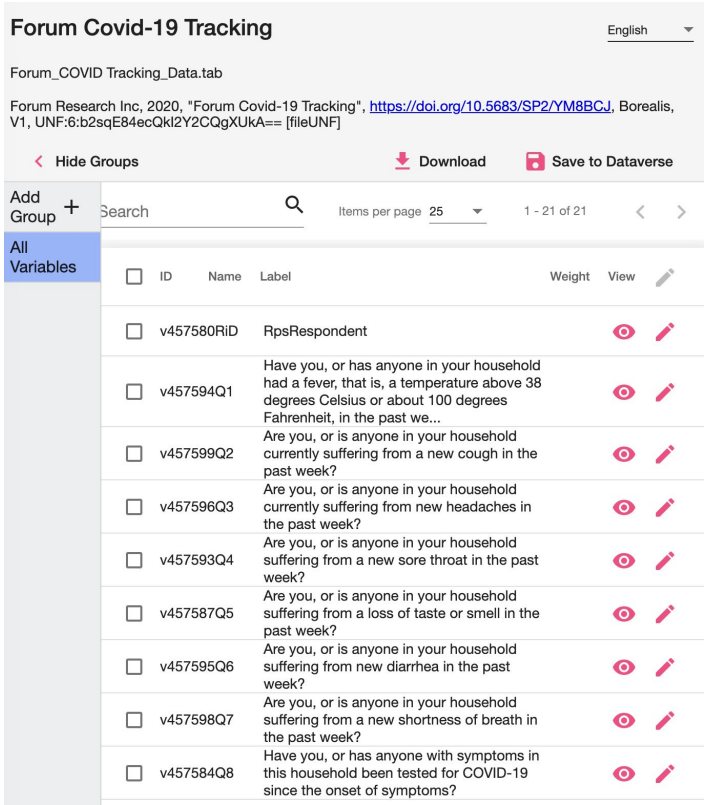
Zip File Previewer +

- 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 and downloader too



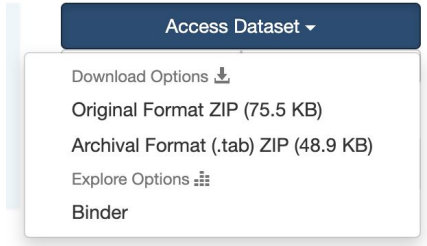
File Exploration, Curation, and Query Tools

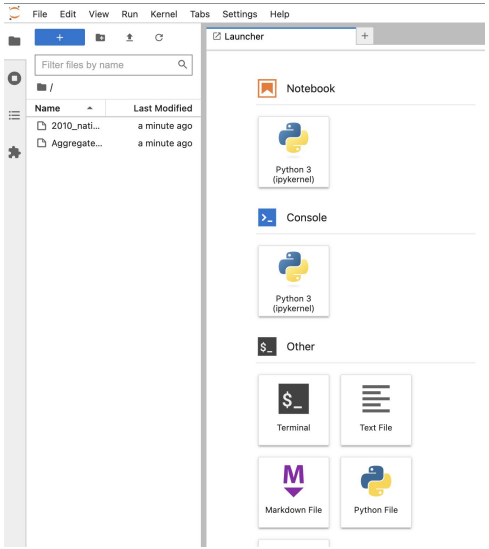
- File level **explore** tools provide a variety of features from data visualization to statistical analysis
- File level **configure** tools allow (authorized) users to send metadata about the file back to Dataverse
- File level **query** tools allow the user to ask questions (e.g. natural language 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



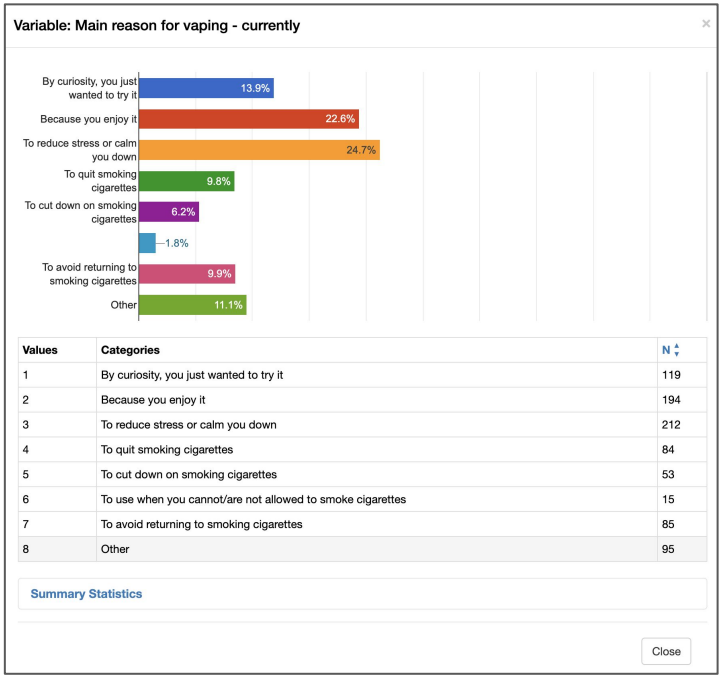


Spotlight: Survey Data Analysis using Data Explorer

- 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

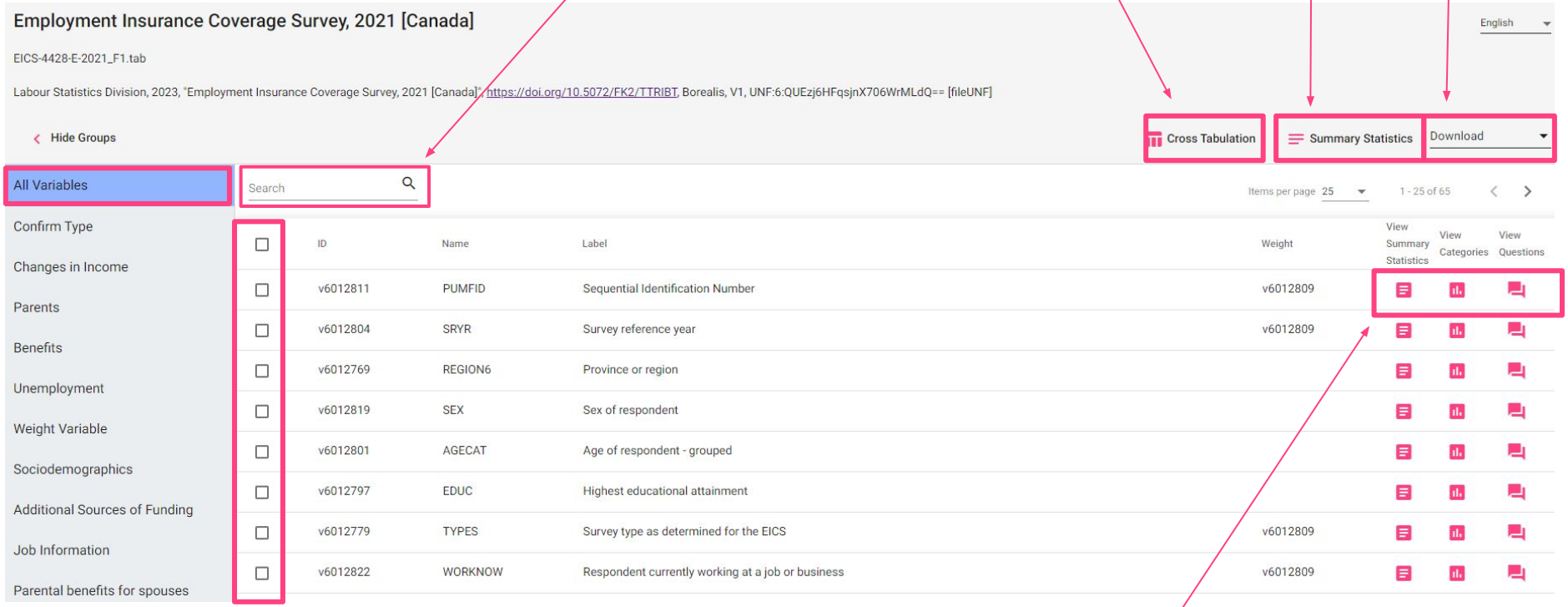
Data Explorer

Search for variables

Cross tabulate two or more variables

View multiple summary statistics

Subset and download

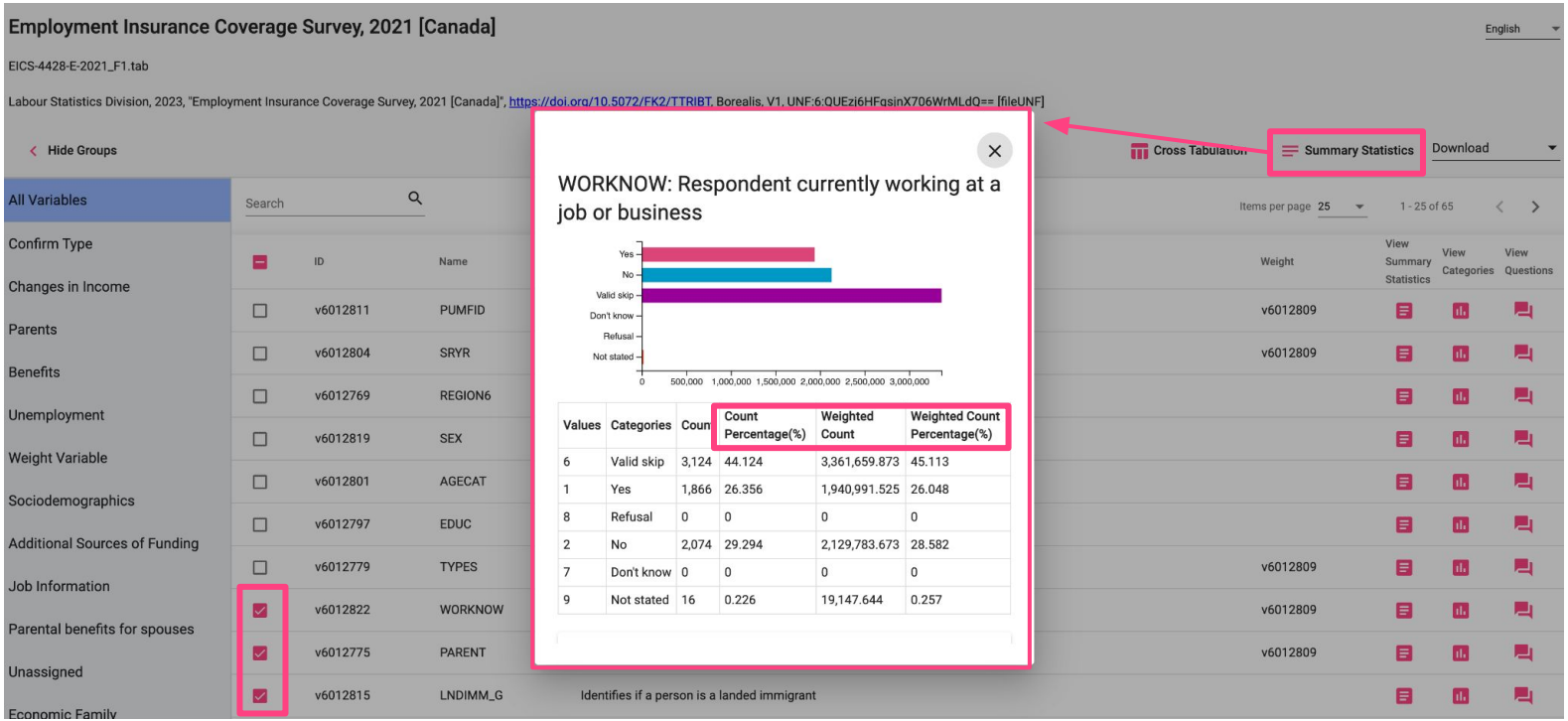


Browse by Variable Groups

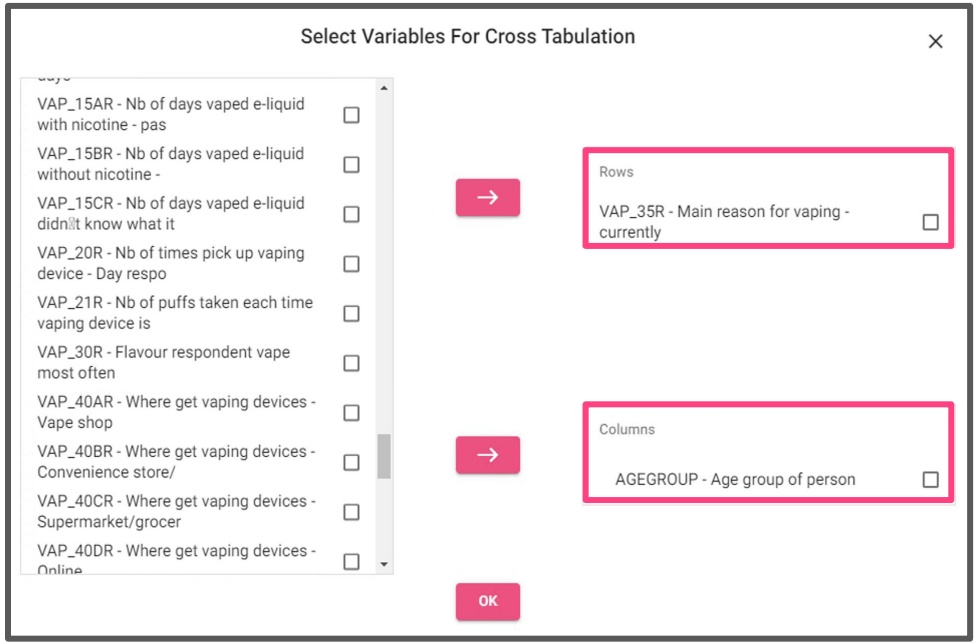
Select / Unselect Variables

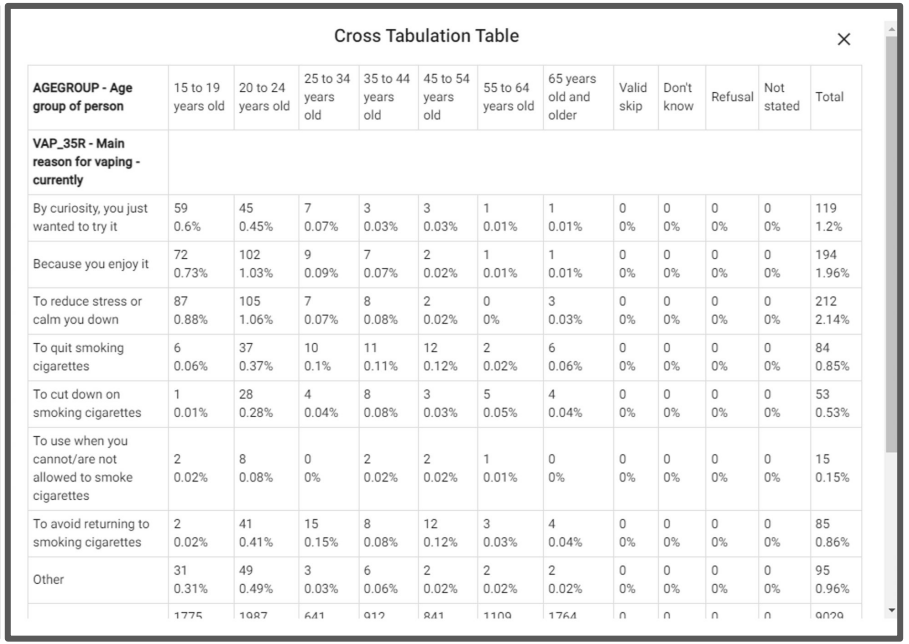
View summary statistics , frequencies, question information

Summary Statistics



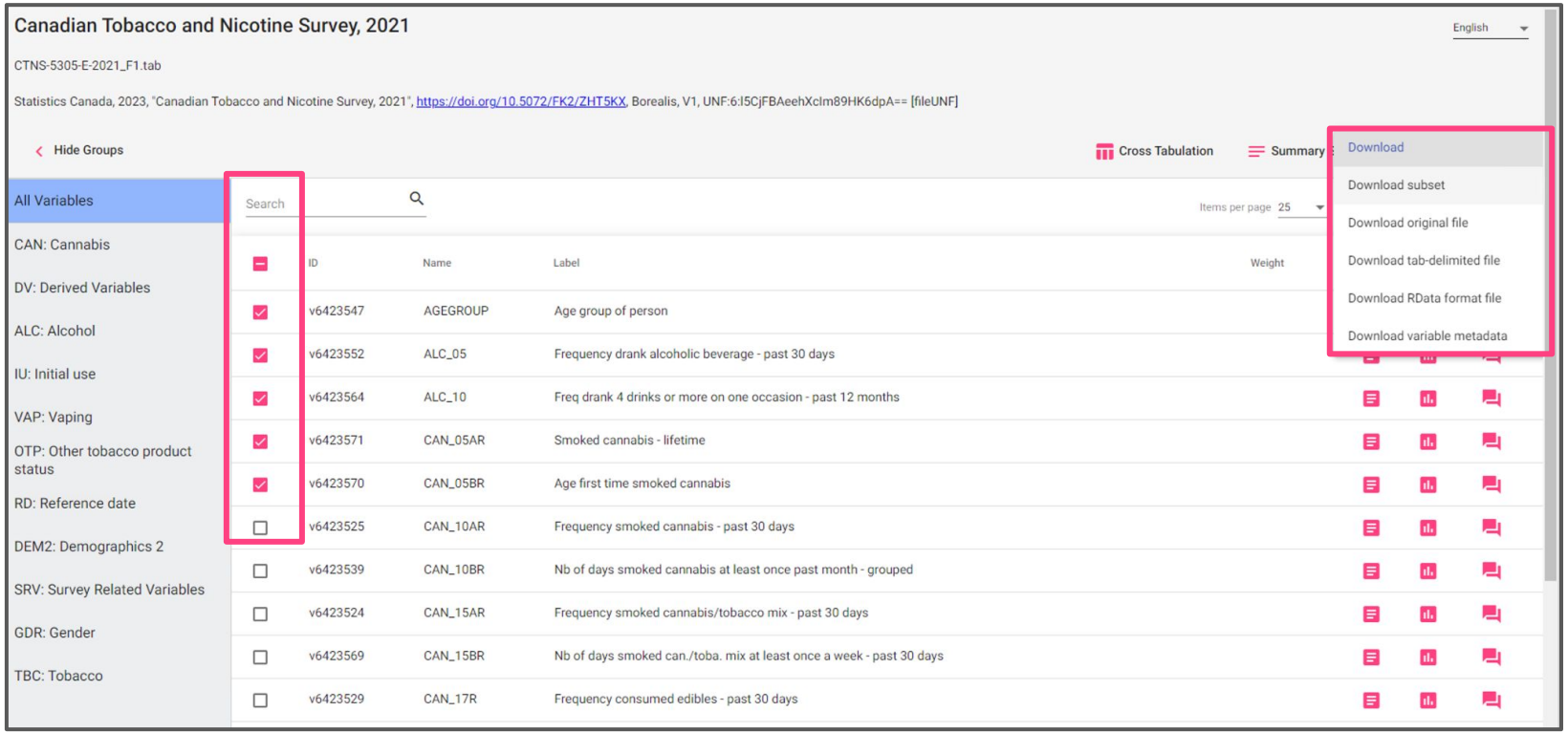
Analyse data





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

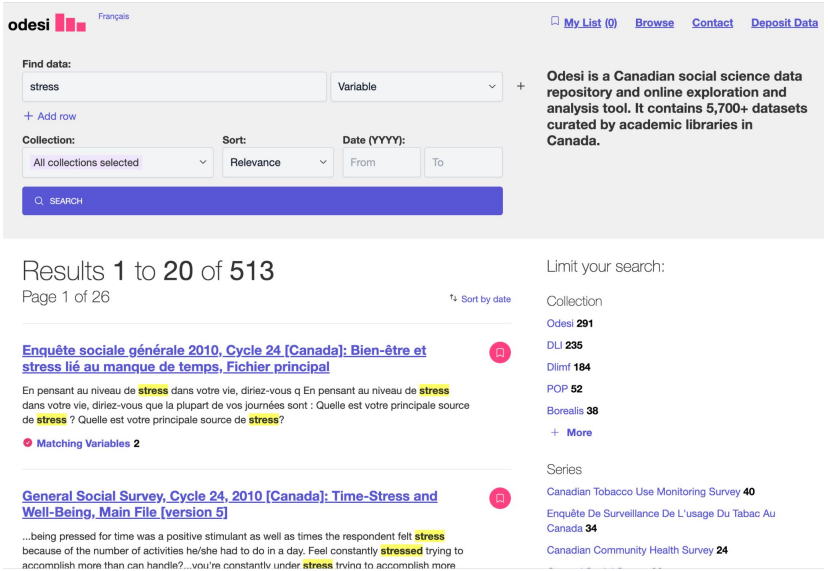
Download and subset



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. Odesi)

Data catalogs



Thank you and resources!

The Dataverse Project, IQSS, Harvard University:
https://dataverse.org/ and Guides

The Harvard Dataverse Repository support and policies website:
https://support.dataverse.harvard.edu/

Upcoming Trainings, **CHECK FOR UPCOMING WEBINARS!**
https://support.dataverse.harvard.edu/calendar/upcoming

Contact us:
support@dataverse.harvard.edu