Workshop: How to set up and configure a Dataverse repository that suits your needs

Alina Danciu, Geneviève Michaud, Baptiste Rouxel, Tom Villette

SciencesPo
CENTRE DE DONNÉES SOCIO-POLITIQUES

Workshop: How to set up and configure a Dataverse repository that suits your needs

IASSIST 2022
07-06-2022
DOI:10.5281/zenodo.6625972

Alina Danciu, Geneviève Michaud, Baptiste Rouxel
CDSP, Sciences Po & CNRS

How to set up and configure a Dataverse repository that suits your needs

1. Data Documentation Initiative (DDI) and FAIR
2. The Dataverse project in brief
3. What's in a data repository?
4. Assess your needs
5. Main Dataverse features
6. A Dataverse bingo
7. Desirable improvements
8. Extending functionalities
9. Stay tuned / contribute to the Dataverse project

Data Documentation Initiative (DDI) and FAIR

Quizz : What is FAIR ?

- An elaborate contemporary folk dance involving the energetic flapping of the jaws and waving of hands, followed by a prolonged period of inactivity
- A specific set of universally agreed practices for sharing research data, implemented by adhering to well-defined specifications applying equally across all domains
- A compelling article published in *Nature* in 2016, describing the basic principles which should be followed for sharing research data in sciences of all kinds

Quizz : What is FAIR ?

- ~~An elaborate contemporary folk dance involving the energetic flapping of the jaws and waving of hands, followed by a prolonged period of inactivity~~
- A specific set of universally agreed practices for sharing research data, implemented by adhering to well-defined specifications applying equally across all domains
- A compelling article published in *Nature* in 2016, describing the basic principles which should be followed for sharing research data in sciences of all kinds

Quizz : What is FAIR ?

- ~~An elaborate contemporary folk dance involving the energetic flapping of the jaws and waving of hands, followed by a prolonged period of inactivity~~
- ~~A specific set of universally agreed practices for sharing research data, implemented by adhering to well-defined specifications applying equally across all domains~~
- A compelling article published in *Nature* in 2016, describing the basic principles which should be followed for sharing research data in sciences of all kinds

Quizz : What is FAIR ?

- ~~An elaborate contemporary folk dance involving the energetic flapping of the jaws and waving of hands, followed by a prolonged period of inactivity~~
- ~~A specific set of universally agreed practices for sharing research data, implemented by adhering to well-defined specifications applying equally across all domains~~

A compelling article published in *Nature* in 2016, describing the basic principles which should be followed for sharing research data in sciences of all kinds

FAIR is a (simple) ideea

- **F**indable, **A**ccessible, **I**nteroperable, **R**e-usable (“The FAIR Guiding Principles”)*
- Promote data-sharing and reuse within and between domains
- DDI has been focused on data sharing and reuse for decades
- Important ideas whose time has come
  - Demand for more data (large projects, new technologies)
  - Broader acceptance of data-sharing as *important*

The key to FAIR data is *metadata*

* Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3. 160018. https://doi.org/10.1038/sdata.2016.18

DDI®
DATA DOCUMENTATION INITIATIVE

How does DDI help about FAIR

- Many people talk about FAIR but focus only on Findability and Accessibility
  - This isn't "FAIR", it's "FA"
- FAIR include the Interoperability and Reusability parts as well!
  - This has been the primary focus of DDI for a long time
  - The hard, expensive part…
- DDI is for people who want to be *serious* about FAIR!
  - DDI provides the rich metadata which is required
  - Reminder : DDI is written in XML

What is DDI?

- DDI = Data Documentation Initiative
  - An international metadata standard
  - Used primarily in the social and behavioural sciences, economics, health
- An open standard designed for data sharing and reuse
- A structure for describing data and its related information
- Describes data from surveys and other observation-based data collection methods
  - Currently moving towards covering new data types and data from new domains.

DDI Specifications

- DDI Codebook (aka DDI-C)
- DDI Lifecycle (aka DDI-L)
- DDI Cross Domain Integration (aka “DDI-CDI”) - to be released
- Each one suits different needs and types of data

DDI Codebook

- An XML description of a “codebook” (a data dictionary)
  - Rectangular files
  - No concept of metadata reuse
- Based on models in existing analysis tools
  - (Stata, SPSS, SAS, etc.)
- Included Dublin Core and descriptive “study-level” metadata
- Machine-readable
- Described data for a single study (one point in time)
- After-the-fact description to support archiving and reuse

DDI Codebook

Metadata not structured

Lost Metadata Manager

DDI Codebook structures

Metadata structured by DDI Codebook standard

DATA DOCUMENTATION INITIATIVE

Do you speak xml ? No? Then you need a tool.

```
<titl>Canadian Community Health Survey, 2012: Annual
Component </titl>
    <labl>Questionnaire (.pdf)</labl>

<dataDscr><notes>The variables in this study are
identical to earlier waves. </notes></dataDscr>

<titl>Canadian Gallup Poll, May 2000</titl>
    <dataChck>Quality checks were performed by Carleton
University Data Centre. </dataChck>
```

DATA DOCUMENTATION INITIATIVE

Want to know more ?

- Request Training from the DDI Alliance Training Group
- DDI Alliance, responsible for the standard, is a rich source of information (check the website!)
- DDI Training Group Zenodo Collection
- Attend CODATA / DDI Alliance webinars

The Dataverse project in brief

Dataverse in brief

Dataverse is an open source research data repository software, that is being developed at Harvard IQSS & supported by a consortium (GDCC). It has been initially dedicated to the SSH but is being widely used worldwide.

To date, 78 instances are using Dataverse, among which:

- Scholars Portal (Borealis)
- DataverseNO
- SODHA

Source: Dataverse Metrics

Dataverse at the CDSP

Our first need was a repository to host our data and our DDI metadata. Dataverse uses DDI-Codebook (DDI-C) / DDI 2.5

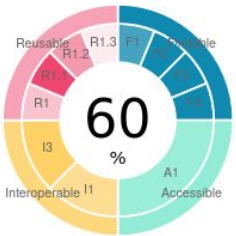
The CDSP has set up an institutional SSH research data repository: data.sciencespo, including two collections:

- CDSP databank: curated up to the level of the variable, pseudonymized,
  - national scope
- Sciences Po Self-deposit collection
  - institutional scope

What's in a data repository?

Dataverse and FAIR

Dataverse provides a framework to achieve data FAIRness (Findable, Accessible, Interoperable and Reusable). Customizing a Dataverse instance and curating a dataset improves data final reusability.



	Score earned:	Fair level:
Findable:	6.5 of 7	advanced
Accessible:	1 of 3	initial
Interoperable:	3 of 4	moderate
Reusable:	4 of 10	initial

Anusuriya Devaraju, & Robert Huber. (2020). F-UJI - An Automated FAIR Data Assessment Tool (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.4063720. Available at: https://www.f-uji.net , the test results shown here are based on preliminary data and code which still is under development.
Results shown are available here (resource evaluated 23-05-2022)

FAIR data

a persistent identifier

That takes you to <a landing page>

Where you find <quality metadata>

Including <a license>

And a one step <access to curated data>

Source: The Noun Project
Daily Ions, NIXX Design,
Deemak Daksina

What's in a data repository landing page?

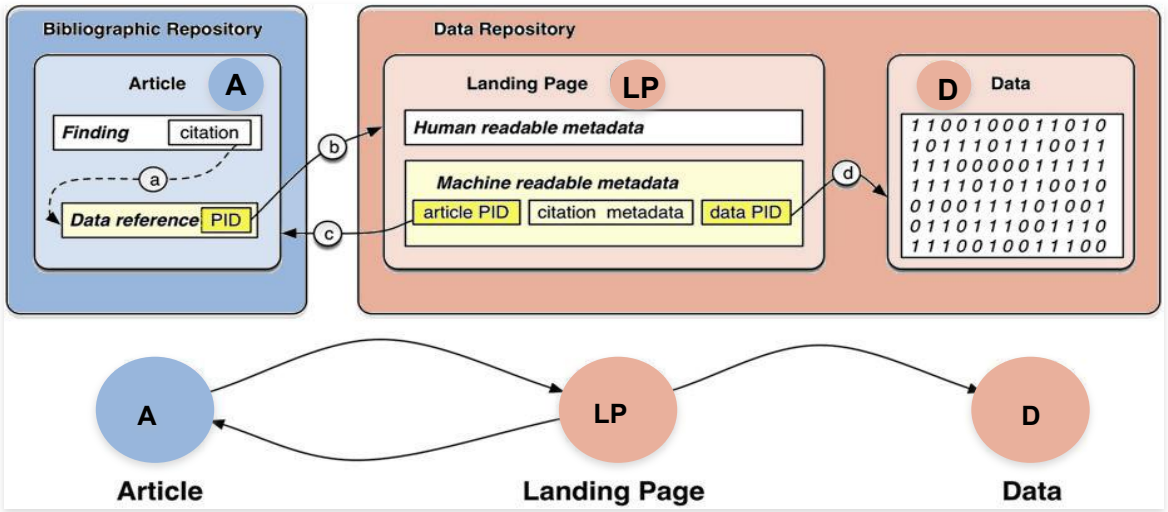
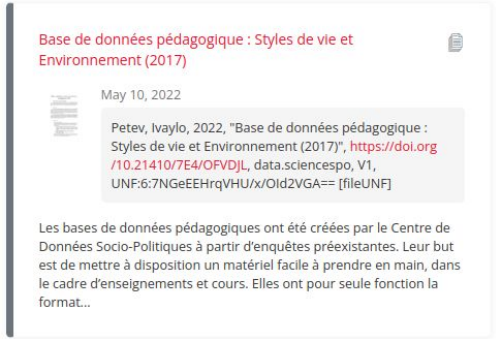


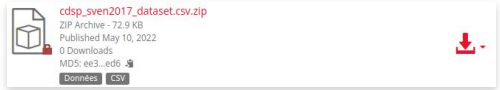
Figure from : Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone & Tim Clark, 2018, "A data citation roadmap for scientific publishers", www.nature.com. https://doi.org/10.1038/sdata.2018.259

Collections, Datasets, files

Information is arranged in Collections, Datasets and Files:

- Collections
  - including as many (sub-)collections as needed
- Datasets
  - are included in a collection and include
- Files





Anatomy of a dataset 1/3

Dataverse offers a comprehensive landing page for each dataset, a strong basis for FAIR data.

This is a public web page showing metadata.

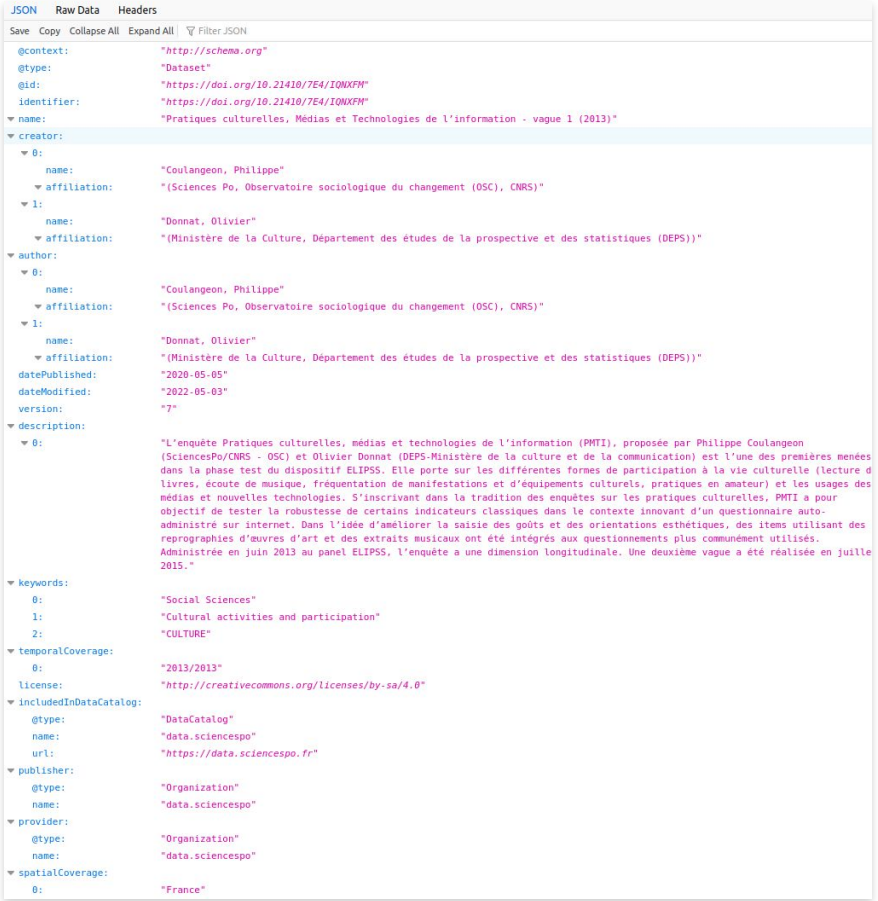
This page is meant for humans to read it.

But it serves also metadata for machines.

Anatomy of a dataset 2/3

Since version 4.8, Dataverse supports Schema.org embedded metadata in every dataset landing page.

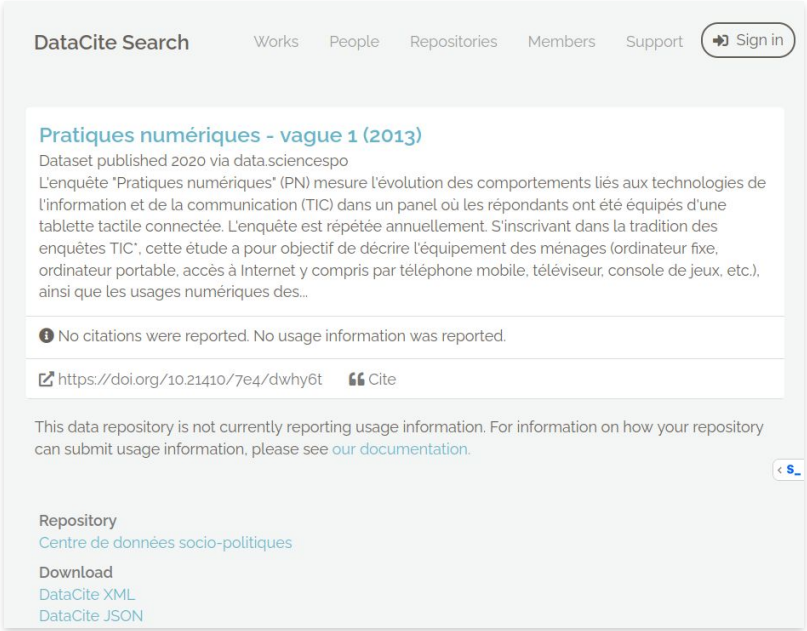
Schema.org metadata are meant for machines to support discoverability of datasets. Search engines rely on such metadata.



Anatomy of a dataset 3/3

When a dataset is created, Dataverse sends to the PID provider (Datacite for DOIs), a set of metadata and gets back a newly minted PID.

The PID provider secures the link between the PID and the dataset URL on the repository. It is the responsibility of the PID maintainer to maintain this link up to date.



Assess your needs

Who?	Resources and skills
For whom?	Data depositors, data curators, re-users
How?	Curation processes
What?	Data types, scientific field(s)

Who?

What are your available resources and skills? This should be your first question. In other terms: what are and will be your available IT resources (system admin, devOps) to setup and maintain a Dataverse instance over time.

You will first bear installation costs. Maintenance comes with a cost (upgrades and backups):

A repository is a commitment to sustainability

For whom?

What is the underlying organization?

Is internationalization needed?		See UI customization, limitations, metadata language
Do you need to signal metadata on other portal?	Will the repository be harvested? Will the repository harvest others?	See OAI-PMH
Who are the depositors, the curators?	Are they included in a directory (Univeristy Identity Provider)? A federated directory (like eduGAIN)?	See authentication

How? Research data management

How is Research Data Management organized?

Automation	Will you rely only on the web interface? Will you automate some tasks?	See Dataverse API
Curation	Are there dedicated team members for data and metadata curation? Are data providers self-depositing?	How should you structure the collections tree to facilitate RDM? See permissions
Workflows	Is there collaboration taking place before data publication?	Review, private URL Embargo, Workflows

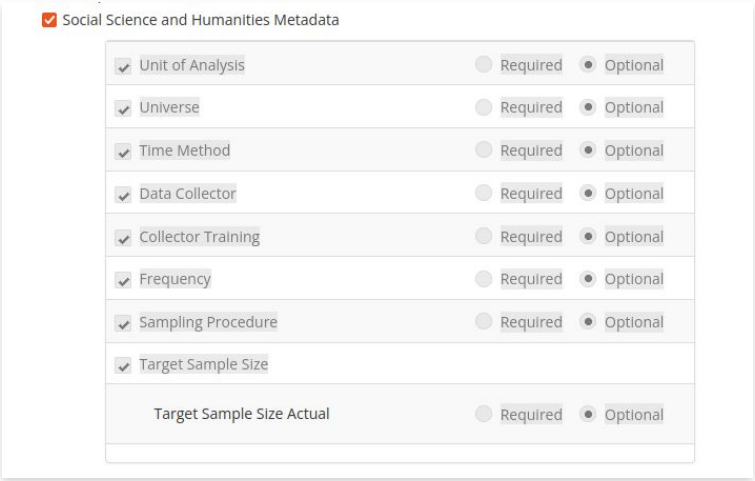
Which data?

Do you manage qualitative data?	See high volume and storage
Do you manage quantitative data?	See file ingest, integrations, previewers
Do you manage active data?	See collaboration features, dataset versioning
Is it a general purpose or a specialized repository?	See metadata blocks, metadata customization, templates

Main Dataverse features

Crowdsourced list of Dataverse features
https://dataverse.org/software-features

Dataset metadata



- **Persistent Identifiers (PID)**
  - DOI or Handle are supported by Dataverse
  - Mandatory at dataset level, optional at file level
- **Metadata**
  - Fields customization per collection (enable/disable blocks of metadata, select fields & configure as required/optional)
  - Controlled vocabularies
  - Templates

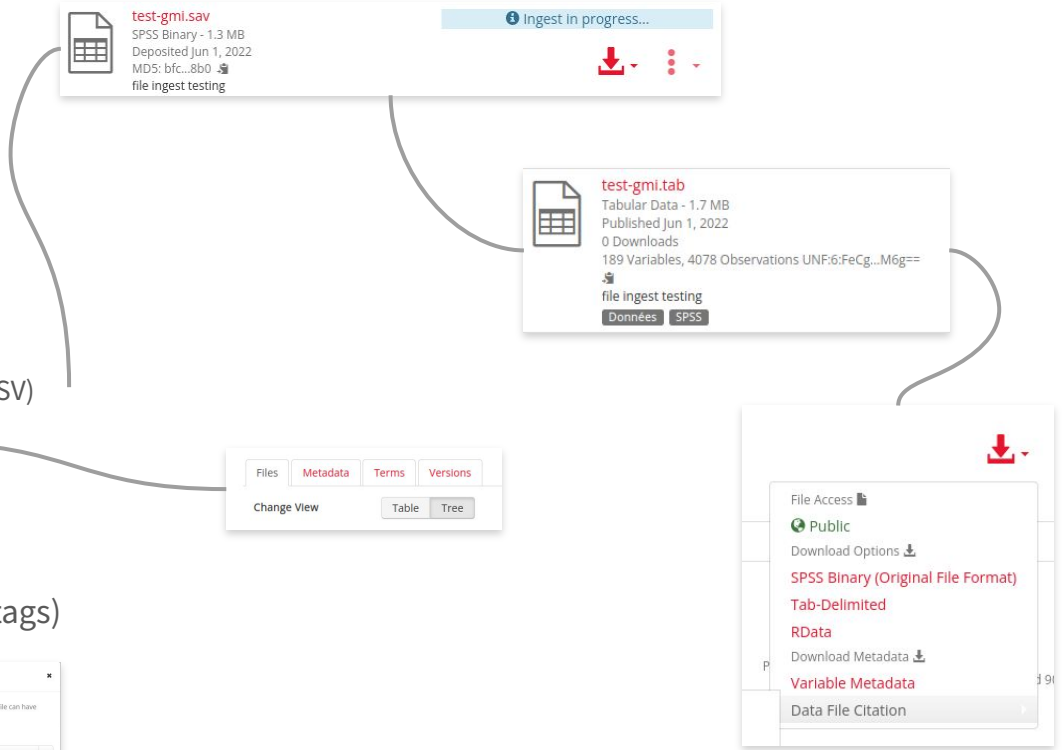
Improves metadata quality

- **Licenses**
  - From version 5.10 licenses are available as controlled vocabularies

Improves interoperability

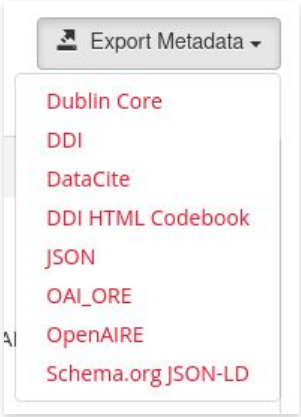
Dataset files

- File ingestion:
  - Tabular data (SPSS, Stata, R, Excel, C/TSV)
  - ZIP file ingestion & tree view
  - Other formats uploaded as is
- Access restrictions at file-level
- Metadata : Limited file level metadata (tags)



Metadata export formats

- DDI
- DDI HTML Codebook
- OpenAire / DataCite
- Dublin core
- OAI-ORE
- Schema.org JSON-LD
- JSON

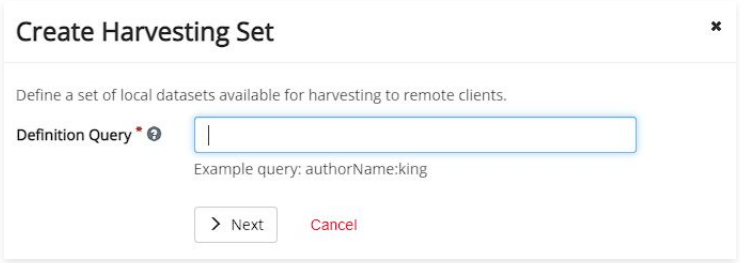


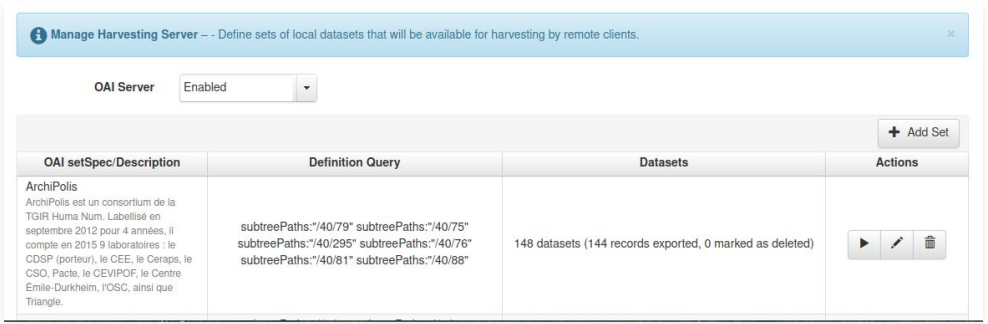
Publication workflow

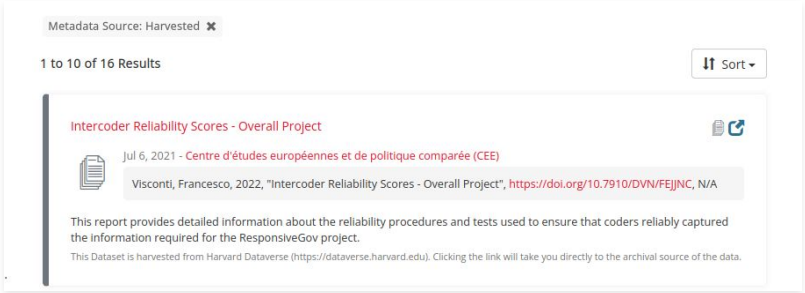
- Versioning
- Embargo (file level)
- Pre-reserved PID
  - PID available for a pre-print
- Private URL for review
- Advanced workflow customization, triggering actions
  - actions on pre/post publication
  - long term preservation

Harvesting metadata (OAI-PMH)

- Server: Enable harvesting for your Dataverse by creating sets (defined by a query)
- Client: Harvest other Dataverse instances



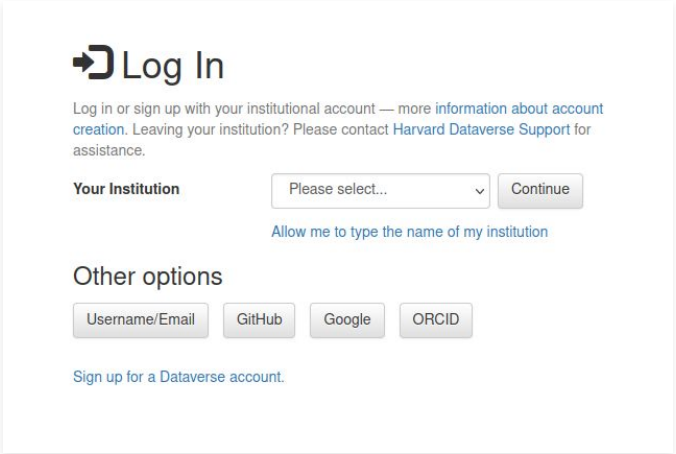




Authentication

You can configure a single source of authentication or a combination of several.

- Local account
  - Username/Password
- Institutional accounts (Shibboleth)
  - Home institution's Identity Provider (University SSO)
  - Federation of Identity Providers (EduGain)
- External accounts (OAuth2)
  - Google
  - ORCid
  - GitHub
  - Microsoft

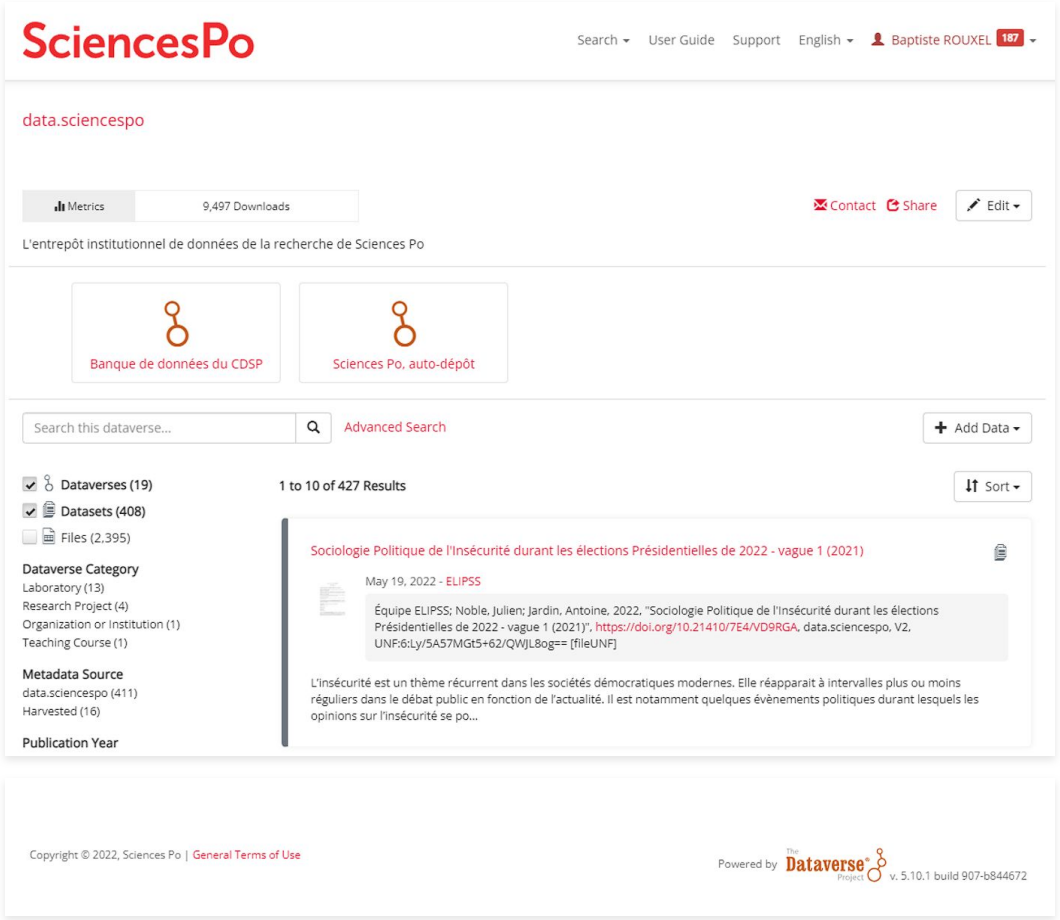


Permissions

- Customization of permissions at a collection, dataset or file level
- Manage permissions by
  - User
  - Group
  - Role (= list of predefined permissions: can publish, can edit, can download, …)
- If a file is restricted, users can send an access request

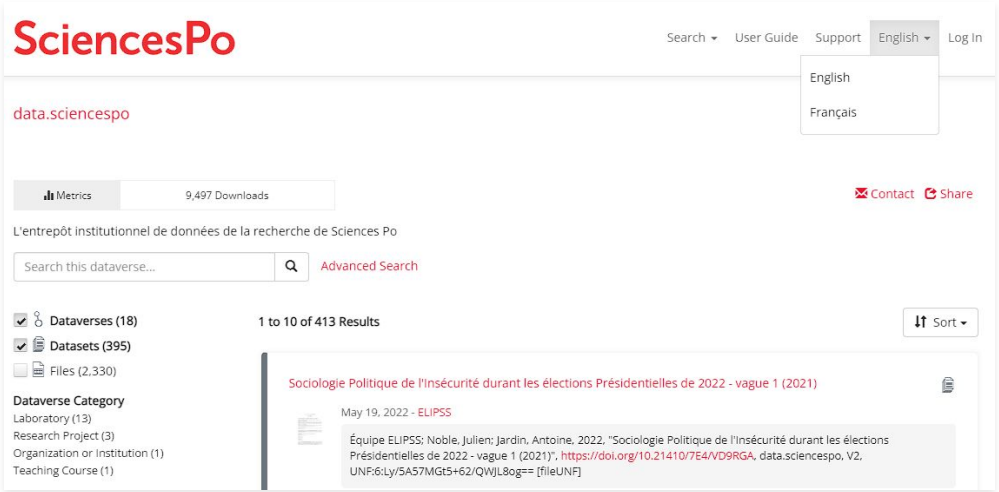
UI customization

- Homepage
- Stylesheet (CSS)
- Footer
- Carousel
- Search facets



Internationalization

- UI language switcher
- Does not translate the metadata



	Dataverse instance	Collection	Dataset	File
Branding	✓	✓		
Permissions	✓	✓	✓	✓
PID (DOI or Handle) PID granularity is set at the instance level			✓	✓
License			✓	
Search facets (metadata filters)	✓	✓		
Metadata blocks choice	✓	✓		
Metadata profiles (mandatory/optional metadata)	✓	✓		
Metadata language	✓	✓		
Metadata templates		✓		
Harvesting entrypoint (OAI-PMH)	✓	✓	✓	

Main settings and relevant scope for Dataverse

"How to set up and configure a Dataverse repository that suits your needs", A. Danciu, G. Michaud, B. Rouxel, IASSIST 2022

3 use cases & a Dataverse bingo

Use case #1

You are asked to set up a repository to host data produced within one astronomical research center. Researchers will (self) deposit data with minimum support. No curation team is involved. Data are meant to be public, except some unpublished data that will support pre-prints.

What choices would you suggest?

Use case #2

You are missioned to setup a generic purposes repository for a federation of universities. Data project are often created by research groups that span over more than one institution. Dedicated curation team are available in each partner institution. The repository metadata should be showcased on a national Open Science Portal.

Did we tell you this use case takes place in Belgium?

What choices would you suggest? How collection could be organized?

Use case #3

You are responsible for the national Open Science Portal.

Desirable improvements

Desirable improvements

- Current limitations
  - PID sources (one single shoulder / PID source)
  - Exporting metadata only possible after publication
- Possible improvements
  - OAI-PMH (lacks license, object type dataset)
  - DDI support (export / harvesting)
  - Internationalization (only UI and CVs, no multiple languages for other metadata)

Extending functionalities

- Out of the box
  - Advanced metadata customization
  - Long term preservation (BagIt)
- External tools & community tools
  - pyDataverse
  - Data previewers
  - Data explorer
  - Data curation tool
  - Dataverse Feed
- Other tools
  - DP Creator
  - Whole Tale

Advanced metadata customization

Customization can be done, but increases maintenance cost and can create errors.

- Adding new fields
  - It’s fine
- Adding controlled vocabularies
  - It’s fine but updates are tricky
- Editing existing fields
  - Could break functionalities in the current / upcoming version

Adding translations is needed if supporting multiple languages.
New fields are not supported in the DDI export (but controlled vocabularies are supported for original fields).

Custom metadata blocks at DaRUS (University of Stuttgart) Dataverse instance https://doi.org/10.18419/darus-2463

Long term preservation (BagIt)

- Dataverse supports BagIt
- Research Data Alliance (RDA) conformant
- Workflow can be configured to automatically archive data
- An infrastructure is needed to store the data

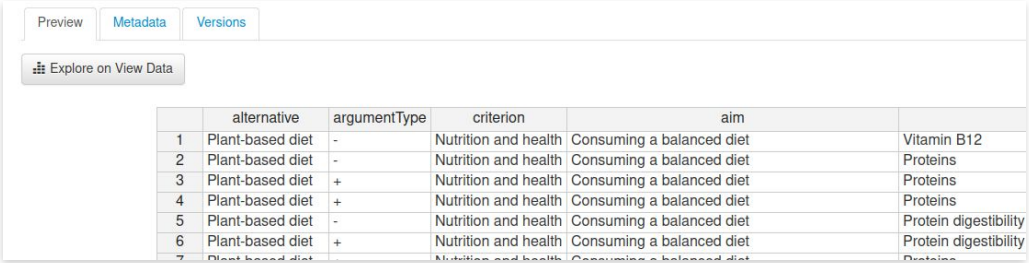
A RDM "swiss knife": pyDataverse

- Developed by Stefan Kasberger at AUSSDA
  - (Austrian Social Science Data Archive)
- Provides a python interface to make operations on Dataverse
  - (example: using the API, migrate data, …)
- Can also be used to explore data
  - (example: download data -> run any script)

https://pydataverse.readthedocs.io

Data previewers

- Developed by the Dataverse community
- Currently available:
  - PDF, text
  - Spreadsheet (example shown https://doi.org/10.15454/HOBUZH/RCK1NH)
  - Image, video, audio
  - HTML

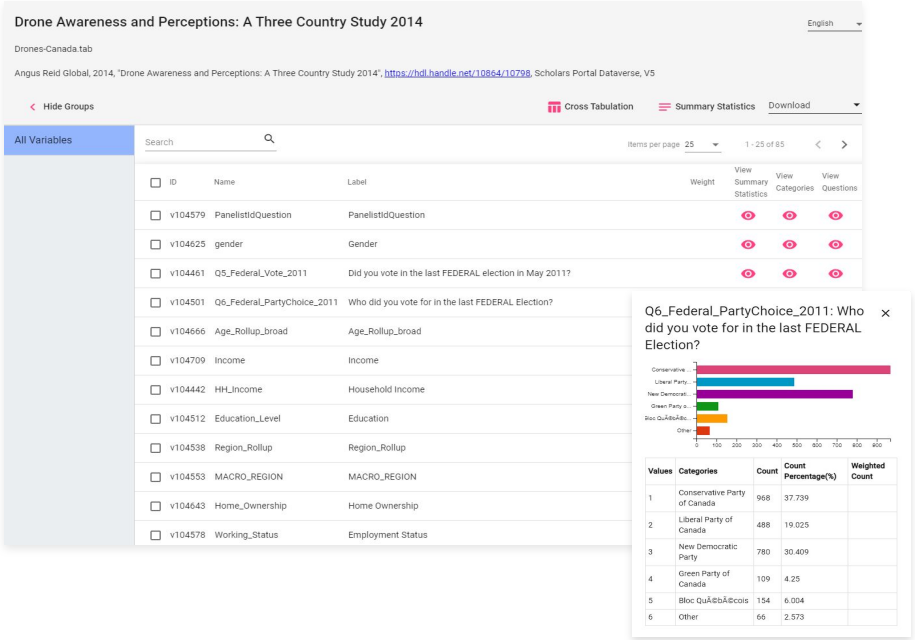


	alternative	argumentType	criterion	aim	
1	Plant-based diet	-	Nutrition and health	Consuming a balanced diet	Vitamin B12
2	Plant-based diet	-	Nutrition and health	Consuming a balanced diet	Proteins
3	Plant-based diet	+	Nutrition and health	Consuming a balanced diet	Proteins
4	Plant-based diet	+	Nutrition and health	Consuming a balanced diet	Proteins
5	Plant-based diet	-	Nutrition and health	Consuming a balanced diet	Protein digestibility
6	Plant-based diet	+	Nutrition and health	Consuming a balanced diet	Protein digestibility

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

Data explorer

- Developed by Scholars Portal
- Provides a web application to explore data from (ingested) tabular files



https://github.com/scholarsportal/dataverse-data-explorer-v2

Data curation tool

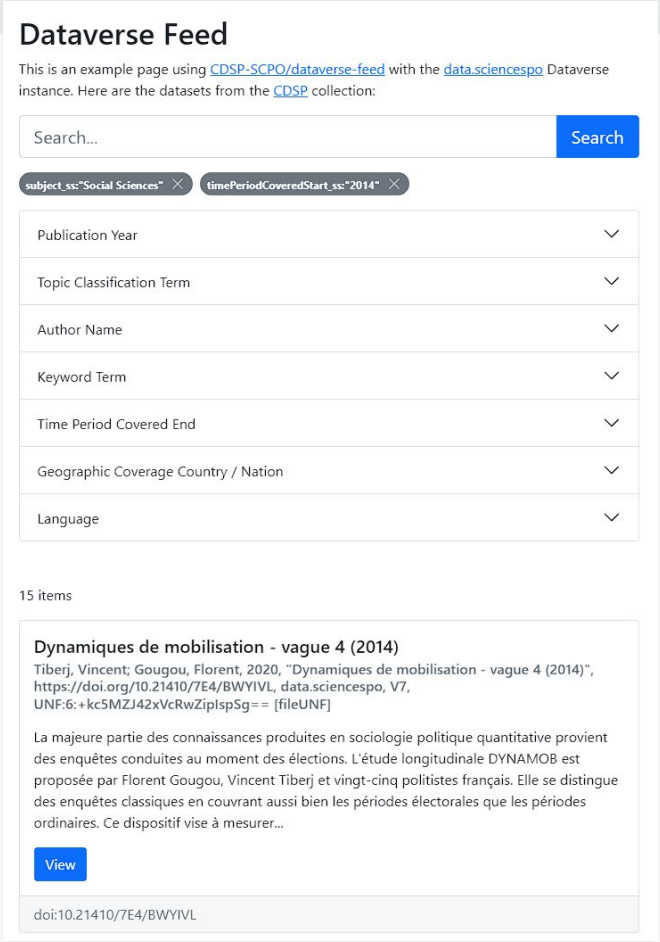
- Developed by Scholars Portal
- Provides a web application to help editing variable-level metadata for tabular files

https://github.com/scholarsportal/Dataverse-Data-Curation-Tool

Dataverse Feed

- Prototype
- Easily embed your datasets into an existing webpage
- Customizable (features, appearance)

https://github.com/CDSP-SCPO/dataverse-feed



Other tools

- Reproducible research tool: Whole Tale
  - NSF-funded Data Infrastructure Building Block (DIBBS)
- Sensitive data analysis: DP Creator (in development)
  - Leverages the open source tools from the OpenDP library (Harvard / Microsoft collaboration)
  - Produce differentially private statistics on Dataverse
- And more!

How to stay tuned / How to contribute to the Dataverse Project

Ways to contribute

- Bug reports / Feature ideas / Feedback
- Pull requests: source code changes proposals
- Sponsoring issues (GDCC) / Financial support

Dataverse Community

- Google group: official google group for Dataverse users interested in building a community around Dataverse Network software and expanding its functionality within and beyond social sciences.
- GitHub
- Slack / Matrix
- Events
  - Dataverse community calls (~2 per month)
  - Dataverse community meetings (1 per year)

Thank you

Alina.Danciu@sciencespo.fr Genevieve.Michaud@sciencespo.fr Baptiste.Rouxel@sciencespo.fr

Accreditation

The DDI and FAIR section (slides 3-15) is adapted from slides authored by the DDI Alliance Training Working Group

Based on content developed at the DDI Train-the-Trainers Dagstuhl workshop 2018

Picture credits - Codebook slide (13)

- Picture 1 (left) credit: woodleywonderworks, https://www.flickr.com/photos/wwworks/2472232245 (CC-BY) woodleywonderw
- Picture 2 (right) credit: Stephen Edmonds: https://www.flickr.com/photos/popcorncx/3516880947/ (CC-BY-SA)