



**HAL**  
open science

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

Alina Danciu, Geneviève Michaud, Baptiste Rouxel, Tom Villette

### ► To cite this version:

Alina Danciu, Geneviève Michaud, Baptiste Rouxel, Tom Villette. Workshop: How to set up and configure a Dataverse repository that suits your needs. École thématique. France. 2022. hal-03906342

**HAL Id: hal-03906342**

**<https://sciencespo.hal.science/hal-03906342>**

Submitted on 19 Dec 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

---

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

IASSIST 2022

07-06-2022

[DOI:10.5281/zenodo.6625972](https://doi.org/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**

---

---

## Quiz : 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

---

## Quiz : 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

---

## Quiz : 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

---

## Quiz : 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) idea

---

- Findable, Accessible, Interoperable, Re-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>

---

# 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

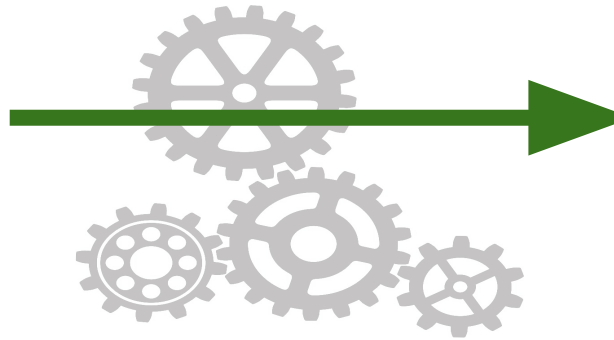
# DDI Codebook

**Metadata not structured**



Lost Metadata Manager

*DDI Codebook structures*



**Metadata structured by DDI Codebook standard**



---

## 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>
```

---

## 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**

---

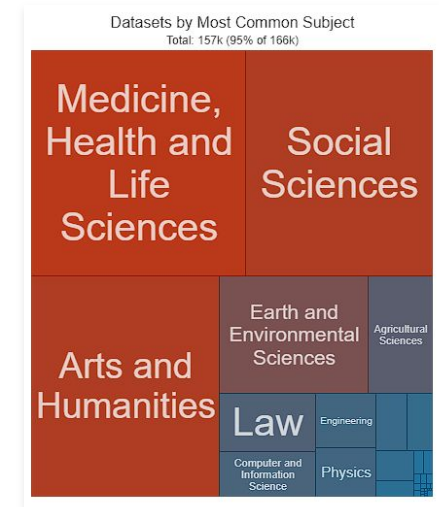
---

# 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 CDSPP

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

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

- [CDSPP databank](#): curated up to the level of the variable, pseudonymized,
  - national scope
- [Sciences Po](#) Self-deposit collection
  - institutional scope



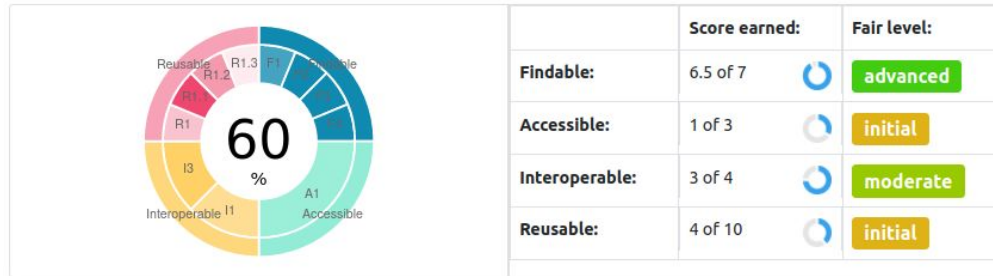
# **What's in a data repository?**

---

---

# Dataverse and FAIR

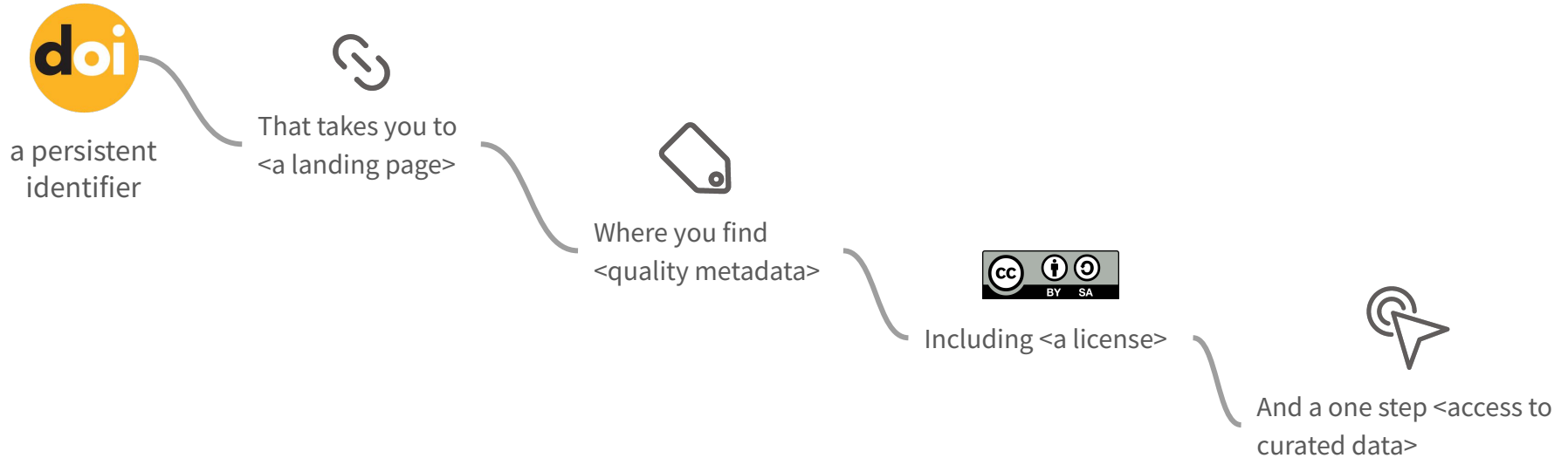
Dataverse provides a framework to achieve data FAIRness (Findable, Accessible, Interoperable and Reusable). Customizing a Dataverse instance and curating a dataset improves data final reusability.



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-ujii.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



---

# What's in a data repository landing page?

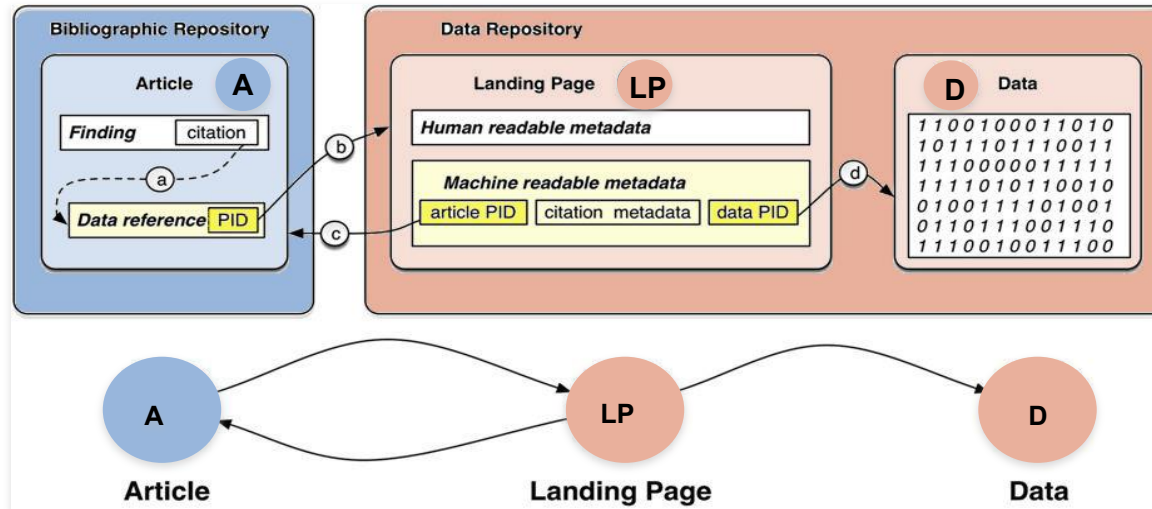




Figure from : Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone & Tim Clark, 2018, "A data citation roadmap for scientific publishers", www.nature.com, <https://doi.org/10.1038/sdata.2018.259>

# 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

**données qualitatives** (Sciences Po, Centre de données socio-politiques (CDSP), CNRS) 

Mar 29, 2021





**Base de données pédagogique : Styles de vie et Environnement (2017)** 

May 10, 2022

 Petev, Ivaylo, 2022, "Base de données pédagogique : Styles de vie et Environnement (2017)", <https://doi.org/10.21410/7E4/OFVDJL>, data.sciencespo, V1, UNF:6:7NGeEHrqVHU/x/Old2VGA== [fileUNF]

Les bases de données pédagogiques ont été créées par le Centre de Données Socio-Politiques à partir d'enquêtes préexistantes. Leur but est de mettre à disposition un matériel facile à prendre en main, dans le cadre d'enseignements et cours. Elles ont pour seule fonction la format...

 **cdsp\_sven2017\_dataset.csv.zip**  
ZIP Archive - 72.9 KB  
Published May 10, 2022  
0 Downloads  
MDS:res\_e06 

**Données** **CSV** 





# 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**.

The screenshot shows the landing page for a dataset on the SciencesPo Dataverse platform. The page is titled "Pratiques culturelles, Médias et Technologies de l'information - vague 1 (2013)" and is version 7.0. The main content area includes a citation for Coulangeon, Philippe; Donnat, Olivier, 2020, "Pratiques culturelles, Médias et Technologies de l'information - vague 1 (2013)", with a DOI link. There are buttons for "Cite Dataset" and "Learn about Data Citation Standards". The description section provides a detailed overview of the survey, mentioning its purpose to test the robustness of certain indicators in the context of an innovative questionnaire administered on the internet. The subject is "Social Sciences" and the keyword is "CULTURE". The license is "CC BY-SA 4.0". Navigation tabs for "Files", "Metadata", "Terms", and "Versions" are visible. At the bottom, there are expandable sections for "Citation Metadata", "Geospatial Metadata", and "Social Science and Humanities Metadata".



## Anatomy of a dataset 2/3

Since version 4.8, Dataverse supports [Schema.org](https://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.

```
JSON Raw Data Headers
Save Copy Collapse All Expand All Filter JSON

@context: "http://schema.org"
@type: "Dataset"
@id: "https://doi.org/10.21410/7E4/I0NKFM"
identifier: "https://doi.org/10.21410/7E4/I0NKFM"
creator:
  0:
    name: "Coulangeon, Philippe"
    affiliation: "(Sciences Po, Observatoire sociologique du changement (OSC), CNRS)"
  1:
    name: "Donnat, Olivier"
    affiliation: "(Ministère de la Culture, Département des études de la prospective et des statistiques (DEPS))"
author:
  0:
    name: "Coulangeon, Philippe"
    affiliation: "(Sciences Po, Observatoire sociologique du changement (OSC), CNRS)"
  1:
    name: "Donnat, Olivier"
    affiliation: "(Ministère de la Culture, Département des études de la prospective et des statistiques (DEPS))"
datePublished: "2020-05-05"
dateModified: "2022-05-03"
version: "7"
description:
  0: "L'enquête Pratiques culturelles, médias et technologies de l'information (PMTI), proposée par Philippe Coulangeon (SciencesPo/CNRS - OSC) et Olivier Donnat (DEPS-Ministère de la culture et de la communication) est l'une des premières menées dans la phase test du dispositif ELIPSS. Elle porte sur les différentes formes de participation à la vie culturelle (lecture de livres, écoute de musique, fréquentation de manifestations et d'équipements culturels, pratiques en amateur) et les usages des médias et nouvelles technologies. S'inscrivant dans la tradition des enquêtes sur les pratiques culturelles, PMTI a pour objectif de tester la robustesse de certains indicateurs classiques dans le contexte innovant d'un questionnaire auto-administré sur internet. Dans l'idée d'améliorer la saisie des goûts et des orientations esthétiques, des items utilisant des reprographes d'œuvres d'art et des extraits musicaux ont été intégrés aux questionnements plus communément utilisés. Administrée en juin 2013 au panel ELIPSS, l'enquête a une dimension longitudinale. Une deuxième vague a été réalisée en juillet 2015."
keywords:
  0: "Social Sciences"
  1: "Cultural activities and participation"
  2: "CULTURE"
temporalCoverage:
  0: "2013/2013"
license: "http://creativecommons.org/licenses/by-sa/4.0"
includedInDataCatalog:
  @type: "DataCatalog"
  name: "data.sciencespo"
  url: "https://data.sciencespo.fr"
publisher:
  @type: "Organization"
  name: "data.sciencespo"
provider:
  @type: "Organization"
  name: "data.sciencespo"
spatialCoverage:
  0: "France"
```



## 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.

The screenshot shows the DataCite Search interface. At the top, there are navigation links for 'Works', 'People', 'Repositories', 'Members', and 'Support', along with a 'Sign in' button. The main content area displays the dataset title 'Pratiques numériques - vague 1 (2013)' in blue. Below the title, it states 'Dataset published 2020 via data.sciencespo' and provides a detailed description in French: 'L'enquête "Pratiques numériques" (PN) mesure l'évolution des comportements liés aux technologies de l'information et de la communication (TIC) dans un panel où les répondants ont été équipés d'une tablette tactile connectée. L'enquête est répétée annuellement. S'inscrivant dans la tradition des enquêtes TIC', cette étude a pour objectif de décrire l'équipement des ménages (ordinateur fixe, ordinateur portable, accès à Internet y compris par téléphone mobile, téléviseur, console de jeux, etc.), ainsi que les usages numériques des...'. A status bar indicates 'No citations were reported. No usage information was reported.' Below this, a link to the DOI 'https://doi.org/10.21410/7e4/dwhy6t' is shown with a 'Cite' button. A message at the bottom states: 'This data repository is not currently reporting usage information. For information on how your repository can submit usage information, please see our documentation.' At the bottom right, there is a small 'S' icon. The repository name 'Centre de données socio-politiques' is listed under the 'Repository' section, and download options for 'DataCite XML' and 'DataCite JSON' are provided under the 'Download' section.

---

## Assess your needs

<b>Who?</b>	Resources and skills
<b>For whom?</b>	Data depositors, data curators, re-users
<b>How?</b>	Curation processes
<b>What?</b>	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 <b>internationalization</b> needed?		See <a href="#">UI customization</a> , <a href="#">limitations</a> , metadata language
Do you need to <b>signal</b> metadata on other portal?	Will the repository be harvested? Will the repository harvest others?	See <a href="#">OAI-PMH</a>
<b>Who</b> are the <b>depositors</b> , the <b>curators</b> ?	Are they included in a directory (Univeristy Identity Provider)? A federated directory (like eduGAIN)?	See <a href="#">authentication</a>

---

# How? Research data management

How is **Research Data Management** organized?

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

---

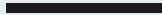
## Which data?

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



# Main Dataverse features

[Crowdsourced list of Dataverse features](https://dataverse.org/software-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
- **Licenses**
  - From version 5.10 licenses are available as controlled vocabularies

Improves metadata quality

Improves interoperability

Social Science and Humanities Metadata		
<input checked="" type="checkbox"/> Unit of Analysis	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Universe	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Time Method	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Data Collector	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Collector Training	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Frequency	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Sampling Procedure	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
<input checked="" type="checkbox"/> Target Sample Size	<input type="radio"/> Required	<input checked="" type="radio"/> Optional
Target Sample Size Actual	<input type="radio"/> Required	<input checked="" type="radio"/> Optional

# 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)

test-gmi.sav  
SPSS Binary - 1.3 MB  
Deposited Jun 1, 2022  
MDS: bfc...8b0  
file ingest testing

Ingest in progress...

Download icon, More options icon

test-gmi.tab  
Tabular Data - 1.7 MB  
Published Jun 1, 2022  
0 Downloads  
189 Variables, 4078 Observations UNF:6:FeCg...M6g==  
file ingest testing

Données SPSS

Files Metadata Terms Versions

Change View Table Tree

File Access

Public

Download Options

SPSS Binary (Original File Format)

Tab-Delimited

RData

Download Metadata

Variable Metadata

Data File Citation

Edit Tags

Select existing file tags or create new tags to describe your files. Each file can have more than one tag.

Selected Tags: Custom, SPSS

File Tags: Select...

Custom File Tag: Documentation, Questionnaire, Donnees, SPSS, Data, Code

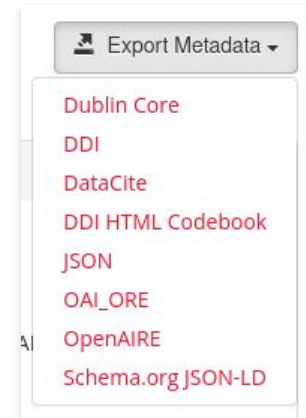
Tabular Data Tags

Save Changes Cancel

---

# 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

### Create Harvesting Set

Define a set of local datasets available for harvesting to remote clients.

Definition Query \*

  
Example query: authorName:king

> Next Cancel

**Manage Harvesting Server** -- Define sets of local datasets that will be available for harvesting by remote clients.

OAI Server Enabled

+ Add Set

OAI setSpec/Description	Definition Query	Datasets	Actions
ArchiPolis ArchiPolis est un consortium de la TIGIR Huma Num. Labellisé en septembre 2012 pour 4 années, il compte en 2015 9 laboratoires : le CDSR (porteur), le CEE, le Ceraps, le CSO, Pacte, le CEVIPOF, le Centre Emile-Durkheim, ROSC, ainsi que Triangle.	subtreePaths:"/40/79" subtreePaths:"/40/75" subtreePaths:"/40/295" subtreePaths:"/40/76" subtreePaths:"/40/81" subtreePaths:"/40/88"	148 datasets (144 records exported, 0 marked as deleted)	

Metadata Source: Harvested

1 to 10 of 16 Results Sort

**Intercoder Reliability Scores - Overall Project**

Jul 6, 2021 - Centre d'études européennes et de politique comparée (CEE)

Visconti, Francesco, 2022, "Intercoder Reliability Scores - Overall Project", <https://doi.org/10.7910/DVN/FEJINC>, N/A

This report provides detailed information about the reliability procedures and tests used to ensure that coders reliably captured the information required for the ResponsiveGov project.

This Dataset is harvested from Harvard Dataverse (<https://dataverse.harvard.edu>). Clicking the link will take you directly to the archival source of the data.

---

# 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

The screenshot shows the 'Log In' page of the Dataverse system. At the top, there is a 'Log In' header with a right-pointing arrow icon. Below the header, a paragraph of text explains that users can log in or sign up with their institutional account, with a link to 'more information about account creation'. It also provides contact information for 'Harvard Dataverse Support' for assistance. The main form area is titled 'Your Institution' and contains a dropdown menu with the text 'Please select...' and a 'Continue' button. Below the dropdown, there is a link that says 'Allow me to type the name of my institution'. Underneath this, there is a section titled 'Other options' which contains four buttons: 'Username/Email', 'GitHub', 'Google', and 'ORCID'. At the bottom of the form, there is a link that says 'Sign up for a Dataverse account.'

---

# 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

The screenshot displays the SciencesPo data portal interface. At the top, the logo "SciencesPo" is in red, followed by navigation links for "Search", "User Guide", "Support", "English", and a user profile for "Baptiste ROUXEL" with 187 items. Below the header, the URL "data.sciencespo" is shown. A "Metrics" section indicates "9,497 Downloads". Action buttons for "Contact", "Share", and "Edit" are present. The main content area features two data sources: "Banque de données du CDSP" and "Sciences Po, auto-dépôt". A search bar contains the text "Search this dataverse..." and "Advanced Search". A sidebar on the left lists filters: "Dataverses (19)", "Datasets (408)", and "Files (2,395)". Under "Dataverse Category", there are "Laboratory (13)", "Research Project (4)", "Organization or Institution (1)", and "Teaching Course (1)". Under "Metadata Source", there are "data.sciencespo (411)" and "Harvested (16)". The "Publication Year" filter is also visible. The main search results area shows "1 to 10 of 427 Results" and a "Sort" dropdown. The first result is "Sociologie Politique de l'Insécurité durant les élections Présidentielles de 2022 - vague 1 (2021)", dated "May 19, 2022 - ELIPSS". The abstract text reads: "Équipe ELIPSS: Noble, Julien; Jardin, Antoine. 2022. 'Sociologie Politique de l'Insécurité durant les élections Présidentielles de 2022 - vague 1 (2021)', <https://doi.org/10.21410/7E4VD9RGA>, data.sciencespo, V2. UNF:6:Ly/5A57MGt5+6Z/QWJL8og== [fileUNF]". A brief description follows: "L'insécurité est un thème récurrent dans les sociétés démocratiques modernes. Elle réapparaît à intervalles plus ou moins réguliers dans le débat public en fonction de l'actualité. Il est notamment quelques événements politiques durant lesquels les opinions sur l'insécurité se po...". The footer contains the copyright notice "Copyright © 2022, Sciences Po | General Terms of Use" and the Dataverse Project logo with version information "v. 5.10.1 build 907-b844672".

# Internationalization

- UI language switcher
- Does not translate the metadata

The screenshot displays the SciencesPo data.sciencepo website. At the top left is the SciencesPo logo. On the top right, there are navigation links for Search, User Guide, Support, English (with a dropdown menu), and Log In. The dropdown menu for English shows options for English and Français. Below the header, the URL data.sciencepo is shown. A Metrics bar indicates 9,497 Downloads. There are links for Contact and Share. The main content area features a search bar with the placeholder text "Search this dataverse..." and a search icon, followed by an "Advanced Search" link. Below the search bar, there are filters for Dataverses (18), Datasets (395), and Files (2,330). A "Dataverse Category" section lists Laboratory (13), Research Project (3), Organization or Institution (1), and Teaching Course (1). The search results show "1 to 10 of 413 Results" and a "Sort" button. The first result is titled "Sociologie Politique de l'Insécurité durant les élections Présidentielles de 2022 - vague 1 (2021)" and includes a date of May 19, 2022, and the ELPSS team.

	Dataverse instance	Collection	Dataset	File
<b><u>Branding</u></b>	✓	✓		
<b><u>Permissions</u></b>	✓	✓	✓	✓
<b><u>PID (DOI or Handle)</u></b> PID granularity is set at the instance level			✓	✓
<b><u>License</u></b>			✓	
<b><u>Search facets</u></b> (metadata filters)	✓	✓		
<b><u>Metadata blocks choice</u></b>	✓	✓		
<b><u>Metadata profiles</u></b> (mandatory/optional metadata)	✓	✓		
<b><u>Metadata language</u></b>	✓	✓		
<b><u>Metadata templates</u></b>		✓		
<b><u>Harvesting endpoint</u></b> (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).



A screenshot of a web interface showing a navigation bar with four tabs: "Files", "Metadata", "Terms", and "Versions". The "Metadata" tab is selected. In the top right corner, there is a button labeled "Export Metadata" with a download icon. Below the navigation bar, there are four expandable metadata sections, each with a dropdown arrow:

- Citation Metadata
- Process Metadata
- Engineering Metadata
- Metadata for Research Software

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/RC K1NH>)
  - Image, video, audio
  - HTML



DEPOSIT\_tab\_arguments\_animal-or-plant-based-diets\_anonymized.tab  
Tabular Data - 42.2 KB - Jan 10, 2020 - 41  
Downloads  
16 Variables, 145 Observations - UNF:6:Wrc5h-XrkFmINNLPI+Vyg==  
145 arguments from the web and grey literature

Preview Metadata Versions



	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
7	Plant-based diet	-	Nutrition and health	Consuming a balanced diet	Proteins

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

# Data explorer

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

The screenshot displays the 'Data Explorer' interface for the dataset 'Drone Awareness and Perceptions: A Three Country Study 2014'. The main panel shows a list of variables with columns for ID, Name, Label, Weight, and View options (View, Summary Statistics, View Categories, View Questions). A search bar is located at the top of the variable list.

An inset window titled 'Q6\_Federal\_PartyChoice\_2011: Who did you vote for in the last FEDERAL Election?' shows a horizontal bar chart and a summary table. The bar chart displays the distribution of votes across six categories: Conservative Party of Canada, Liberal Party of Canada, New Democratic Party, Green Party of Canada, Bloc Québécois, and Other. The summary table below the chart provides the following data:

Values	Categories	Count	Count Percentage(%)	Weighted Count
1	Conservative Party of Canada	968	37.739	
2	Liberal Party of Canada	488	19.025	
3	New Democratic Party	780	30.409	
4	Green Party of Canada	109	4.25	
5	Bloc Québécois	154	6.004	
6	Other	66	2.573	

<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>

## Dataverse Feed

This is an example page using [CDSP-SCPO/dataverse-feed](#) with the [data.sciencespo](#) Dataverse instance. Here are the datasets from the [CDSP](#) collection:

subject\_ss:"Social Sciences" × timePeriodCoveredStart\_ss:"2014" ×

Publication Year	▼
Topic Classification Term	▼
Author Name	▼
Keyword Term	▼
Time Period Covered End	▼
Geographic Coverage Country / Nation	▼
Language	▼

15 items

**Dynamiques de mobilisation - vague 4 (2014)**  
Tiberj, Vincent; Gougou, Florent, 2020, "Dynamiques de mobilisation - vague 4 (2014)", <https://doi.org/10.21410/7E4/BWYIVL>, data.sciencespo, V7, UNF:6:+kc5MZJ42xVcRwZiplsp5g== [fileUNF]

La majeure partie des connaissances produites en sociologie politique quantitative provient des enquêtes conduites au moment des élections. L'étude longitudinale DYNAMOB est proposée par Florent Gougou, Vincent Tiberj et vingt-cinq politistes français. Elle se distingue des enquêtes classiques en couvrant aussi bien les périodes électorales que les périodes ordinaires. Ce dispositif vise à mesurer...

doi:10.21410/7E4/BWYIVL

---

## 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

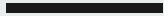
- Bug reports / Feature ideas / Feedback
- Pull requests: source code changes proposals
- Sponsoring issues (GDCC) / Financial support

## 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](mailto:Alina.Danciu@sciencespo.fr) [Genevieve.Michaud@sciencespo.fr](mailto:Genevieve.Michaud@sciencespo.fr) [Baptiste.Rouxel@sciencespo.fr](mailto: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)