La gestió FAIR de les dades de la recerca amb Dataverse

Cicle de Conferències sobre la gestió de dades de recerca
Consorci de Serveis Universitaris de Catalunya
24 de Novembre, 2020

Mercè Crosas, Ph.D., Harvard University
scholar.harvard.edu/mercecrosas @mercecrosas

Photo credit: http://lovedevinandkelly.com/sightseeing/
FAIR everywhere
The FAIR Guiding Principles for scientific data management and stewardship


Access & Citations

157k Article Accesses
1482 Web of Science
1991 CrossRef

Scientific Data 3, Article number: 1
157k Accesses | 1991 Citations

Online attention

1609
1329 tweeters
6 Google+ users
1 F1000
3085 Mendeley
112 blogs
107 news outlets
1 Video uploaders
21 Facebook pages
21 Redditors
4 Wikipedia page

This article is in the 99th percentile (ranked 59th) of the 266,984 tracked articles of a similar age in all journals and the 1st percentile (ranked 1st) of the 1 tracked articles of a similar age in Scientific Data

Published in 2016
54 authors
> 1990 citations
> 157K accesses
Adoption by funding agencies, research communities
Final Report and Action Plan from the European Commission Expert Group on FAIR Data

TURNING FAIR INTO REALITY

2018

Six Recommendations for Implementation of FAIR Practice

By FAIR in Practice Task Force of the European Open Science Cloud FAIR Working Group

Independent Expert Report

EOSC Executive Board
WG FAIR
October 2020
The National Institutes of Health (NIH) Policy for Data Management and Sharing (herein referred to as the DMS Policy) reinforces NIH’s longstanding commitment to making the results and outputs of NIH-funded research available to the public through effective and efficient data management and data sharing practices. Data sharing enables researchers to rigorously test the validity of research findings, strengthen analyses through combined datasets, reuse hard-to-generate data, and explore new frontiers of discovery. In addition, NIH emphasizes the importance of good data management practices, which provide the foundation for effective data sharing and improve the reproducibility and reliability of research findings. NIH encourages data management and data sharing practices consistent with the FAIR data principles.

Final Supplemental Information: The final Supplemental Information asks researchers to describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools. This wording change is meant to highlight the importance of using a PID or other standard indexing tool so the data are findable, which is a key component of the FAIR (Findable, Accessible, Interoperable, and Reusable) Principles. PIDs are also listed as a desirable characteristic of data repositories in the Supplemental Information to the NIH Policy for Data Management and Sharing: Selecting a Repository for Data Resulting from NIH-Supported Research.

Coalition for Publishing Data in Earth and Space Science Commitment Statement

To enable these principles

Repositories will strive to:

- Ensure that research outputs (e.g., data, software, technology, and physical samples) curated by repositories are open and FAIR, have essential documentation, and include human-readable and machine-readable metadata (e.g., on landing pages) in standard formats that are exposed and publicly discoverable.
- Ingest and expose data to promote interoperability and reuse.
- Ensure that unique, persistent identifiers are used for authors (e.g., ORCID), research objects (e.g., Digital Object Identifier), and physical samples (e.g., IGSN).
- Create associations among the research outputs that they manage and other related entities.
- Ensure that data and software have licenses that are as open as possible, and as protected as necessary.
- Support peer-review of related manuscripts by enabling access to the research outputs prior to publication.
- Gain third-party validation of trustworthy and sustainable practices and capabilities.

15 FAIR PRINCIPLES

FINDABLE
- 4 principles
- 2 sub-principles

ACCESSIBLE
- 2 principles
- 2 sub-principles

INTEROPERABLE
- 3 principles
- 3 sub-principles

REUSABLE
- 1 principle
- 3 sub-principles
“The FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.”

About Dataverse
Federated FAIR data repositories worldwide

- Open-source
- 63 installations
- 6 continents
- 7K Dataverse collections
- 135K datasets
- 800K files
- 28M file downloads

Developed at Harvard’s Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community (https://dataverse.org)
Organization of a Dataverse Repository

Dataverse collection
- Collection of datasets
- Own administration
- Own branding (& can be embedded in your site)

Dataset
- Citation
- Metadata
- Versioning
- Terms/permissions
- Collection of Files

File
- Citation
- Preview/Explore
- Metadata
- Versioning
- Permissions

FAIR Principles @mercecrosas
Key Dataverse Features

- Data Citation with **DOI for datasets and files** with credit to data authors
- Link from data to related article
- **Standard** schemas and custom metadata
- Access controls (open vs guestbook vs restricted) with **licenses and terms of use**
- Versioning and provenance
- **Descriptive Statistics** generated from **variables** in tabular data files
- Conversion to multiple formats of tabular data files
- Flexible upload of large data files (>> 5GB): Web UI, API, Standalone Client
- Integration with external tools through extensive **API**
- Data usage metrics with **Make Data Count**
Implementation of FAIR principles in Dataverse
“Digital resources should be easy to find for both humans and computers. **Extensive machine-actionable metadata** are essential for automatic discovery of relevant datasets and services, and are therefore an essential component of the FAIRification process.”

Principle 1: To Be Findable

PRINCIPLE F1

(meta)data* are assigned a globally unique and persistent identifier

*(meta)data refers to data and metadata

DATAVERSE IMPLEMENTATION

- Support for DataCite DOIs; or Handles from Handle.net
- Always at the dataset level
- Optionally at the file level
Full, standard data citation automatically generated

Data Citation, with DataCite DOI, fully compliant with Force11 Joint Declaration of Data Citation Principles
**Principle 1: To Be Findable**

**PRINCIPLE F2**

Data are described with rich metadata

**DATAVERSE IMPLEMENTATION**

- Metadata standards in human- and machine-readable formats: Dublin Core; Documentation Data Initiative (DDI); DataCite; Schema.org
- Optional custom metadata
Support for multiple metadata standards

**PRINCIPLE F3**

.metadata clearly and explicitly include the identifier of the data it describes

**DATAVERSE IMPLEMENTATION**

- ID is in the metadata tab of the Dataset landing page
- ID is in the metadata tab of the File landing page
- ID is included in exported metadata files
This XML file does not appear to have any style information associated with it. The document tree is shown below.
Principle 1: To Be Findable

**PRINCIPLE F4**

(meta)data are registered or indexed in a searchable resource

**DATAVERSE IMPLEMENTATION**

- DataCite metadata is registered and indexed by DataCite Search
- Schema.org metadata is indexed by Google Dataset Search
Data from: The International Political Economy Data Resource

Explore at Harvard Dataverse  Explore at search.datacite.org

6 scholarly articles cite this dataset (View in Google Scholar)

Unique identifier
https://doi.org/10.7910/DVN/X093TV

Dataset updated Aug 7, 2020

Dataset provided by
Harvard Dataverse

License
CC0 1.0 Universal Public Domain Dedication
License information was derived automatically

Description
Quantitative scholars in international relations often draw repeatedly on the same sources of country-year data across a diverse range of projects. The IPE Data Resource seeks to provide a public good to the field by standardizing and merging together variables from 89 IPE data sources into a single dataset, increasing
“Protocols for retrieving digital resources should be made explicit, for both humans and machines, including well-defined mechanisms to obtain authorization for access to protected data.”

Principle 2: To Be Accessible

PRINCIPLE A1

(meta)data are retrievable by their identifier using a standardized communications protocol

Sub-Principle A1.1: the protocol is open, free and universally implementable

Sub-Principle A1.2: the protocol allows for an authentication and authorization procedure, where necessary

DATAVERSE IMPLEMENTATION

- Support for HTTP (W3C), Rsync over ssh (GNU General Public license)
- RESTful API
- Authentication API Tokens
- Authorization service
API Guide

Contents:

- Introduction
  - What is an API?
  - Types of Dataverse API Users
    - API Users Within a Single Installation of Dataverse
      - Users of Integrations and Apps
      - Power Users
      - Support Teams and Superusers
      - Sysadmins
      - In House Developers
    - API Users Across the Dataverse Project
      - Developers of Integrations, External Tools, and Apps
      - Developers of Dataverse API Client Libraries
      - Developers of Dataverse Itself
  - How This Guide is Organized
    - Getting Started
    - API Tokens and Authentication
    - Lists of Dataverse APIs
    - Client Libraries
    - Examples
    - Frequently Asked Questions
  - Getting Help
    - Getting Started with APIs
      - Servers You Can Test With
      - Getting an API Token
      - curl Examples and Environment Variables
      - Depositing Data

https://guides.dataverse.org


**Principle 2: To Be Accessible**

**PRINCIPLE A2**

metadata are accessible, even when the data are no longer available

**DATAVERSE IMPLEMENTATION**

- A deaccessioned dataset (data not available) is still findable and citable
- Metadata includes why the data are not available
- Metadata always accessible for restricted data files
Deaccession reason in dataset landing page when data not longer available

**Deaccession** reason in dataset landing page when data not longer available
Access terms available for restricted data

Optional request access feature for restricted data. FAIR is not equal to Open.
Principle 3: To Be Interoperable

“When two or more digital resources are related to the same topic or entity, it should be possible for machines to merge the information into a richer, unified view of that entity. Similarly, when a digital entity is capable of being processed by an online service, a machine should be capable of automatically detecting this compliance and facilitating the interaction between the data and that tool.”

(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

**DATAVERSE IMPLEMENTATION**

- Linked data support with JSON-LD for Schema.org
- Data Documentation Initiative, DDI, (XML) as a rich schema to support extensive variable metadata
- DDI is unique providing detailed information for each variable
# ClimateRegressionData_150327.tab

**File Citation**


**Dataset Citation**


Learn about Data Citation Standards.

## Preview | Metadata | Versions

### Open View Data

<table>
<thead>
<tr>
<th>statefp</th>
<th>PumaID</th>
<th>msa</th>
<th>mssize</th>
<th>mssize</th>
<th>Wage_orig</th>
<th>Wage</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 1</td>
<td>100100.0</td>
<td>2650</td>
<td>Florence, AL</td>
<td></td>
<td>-0.14469655</td>
<td>-0.1530082</td>
<td>-0.36732796</td>
</tr>
<tr>
<td>2 1</td>
<td>100200.0</td>
<td>3440</td>
<td>Huntsville, AL</td>
<td></td>
<td>-0.06367312</td>
<td>-0.0687066</td>
<td>-0.21142627</td>
</tr>
<tr>
<td>3 1</td>
<td>100300.0</td>
<td>3440</td>
<td>Huntsville, AL</td>
<td></td>
<td>-0.06052007</td>
<td>-0.06744661</td>
<td>-0.3109954</td>
</tr>
<tr>
<td>4 1</td>
<td>100400.0</td>
<td>19999</td>
<td>Non-metro, AL</td>
<td></td>
<td>-0.16140184</td>
<td>-0.16609</td>
<td>-0.49454302</td>
</tr>
<tr>
<td>5 1</td>
<td>100500.0</td>
<td>19999</td>
<td>Non-metro, AL</td>
<td></td>
<td>-0.16811557</td>
<td>-0.15680014</td>
<td>-0.40404822</td>
</tr>
<tr>
<td>6 1</td>
<td>100600.0</td>
<td>2030</td>
<td>Decatur, AL</td>
<td></td>
<td>-0.07162341</td>
<td>-0.09517802</td>
<td>-0.34365948</td>
</tr>
<tr>
<td>7 1</td>
<td>100700.0</td>
<td>19999</td>
<td>Non-metro, AL</td>
<td></td>
<td>-0.21453401</td>
<td>-0.19643022</td>
<td>-0.6055818</td>
</tr>
</tbody>
</table>

Variable metadata from tabular data file
Extensive variable metadata (descriptive statistics) automatically derived from tabular data file in DDI format
Principle 3: To Be Interoperable

PRINCIPLE I2

(meta)data use vocabularies that follow FAIR principles

DATAVERSE IMPLEMENTATION

- FAIR controlled vocabularies and data models used in well-curated datasets
- Metadata template can help
- But, controlled vocabularies and ontologies not supported by default
<table>
<thead>
<tr>
<th>Subject</th>
<th>Social Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic Classification</td>
<td>mra</td>
</tr>
<tr>
<td>50 or fewer (Sampsise)</td>
<td><a href="http://www.murray.harvard.edu/vocabulary">http://www.murray.harvard.edu/vocabulary</a></td>
</tr>
<tr>
<td>male (Gender)</td>
<td><a href="http://www.murray.harvard.edu/vocabulary">http://www.murray.harvard.edu/vocabulary</a></td>
</tr>
<tr>
<td>18-22 (Age)</td>
<td><a href="http://www.murray.harvard.edu/vocabulary">http://www.murray.harvard.edu/vocabulary</a></td>
</tr>
<tr>
<td>White (Race)</td>
<td><a href="http://www.murray.harvard.edu/vocabulary">http://www.murray.harvard.edu/vocabulary</a></td>
</tr>
<tr>
<td>student (SES)</td>
<td><a href="http://www.murray.harvard.edu/vocabulary">http://www.murray.harvard.edu/vocabulary</a></td>
</tr>
<tr>
<td>1 (Generations)</td>
<td><a href="http://www.murray.harvard.edu/vocabulary">http://www.murray.harvard.edu/vocabulary</a></td>
</tr>
<tr>
<td>Special aspects of education (Education)</td>
<td><a href="http://authorities.loc.gov/">http://authorities.loc.gov/</a></td>
</tr>
<tr>
<td>Mental health (Health)</td>
<td><a href="http://authorities.loc.gov/">http://authorities.loc.gov/</a></td>
</tr>
<tr>
<td>Distributor</td>
<td>Murray Research Archive <a href="http://www.murray.harvard.edu">http://www.murray.harvard.edu</a></td>
</tr>
<tr>
<td>Distribution Date</td>
<td>1981</td>
</tr>
<tr>
<td>Time Period Covered</td>
<td>Start: 1930 ; End: 1940</td>
</tr>
<tr>
<td>Date of Collection</td>
<td>Start: 1930</td>
</tr>
<tr>
<td>Kind of Data</td>
<td>field study</td>
</tr>
</tbody>
</table>

Use standard, global Controlled Vocabulary from the Library of Congress.
Enable use of standard vocabularies

FAIR Controlled Vocabularies can be added in a metadata template
PRINCIPLE I3

(meta)data include qualified references to other (meta)data

DATAVERSE IMPLEMENTATION

- DDI schema supports references to other data
- Not yet supported: related objects in exported DataCite metadata
Reference in metadata to related datasets or other research objects
“Digital resources are sufficiently well described for both humans and computers, such that a machine is capable of deciding: if a digital resource should be reused; if a digital resource can be reused, and under what conditions; and who to credit if it is reused.”

Principle 4: To Be Reusable

**PRINCIPLE R1**

(meta)data are richly described with a plurality of accurate and relevant attributes

Sub-Principle R1.1:
(meta)data are released with a clear and accessible data usage license

**DATAVERSE IMPLEMENTATION**

Included in Metadata:
- Data use license/waiver
- Data access and use terms
Licenses, Terms, and Tiered-Access to Data

- CC0 as default waiver for open data;
- Optional Licenses and custom Terms;
- Tiered-access for restricted data
Sub-Principle R1.2: (meta)data are associated with detailed provenance

DATAVERSE IMPLEMENTATION

- Full data citation metadata with credit to data authors, providers, distributors
- Versions with changes documented automatically
- W3C PROV support
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Summary</th>
<th>Contributors</th>
<th>Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.0</td>
<td>Files (Added: 2; Removed: 2); View Details</td>
<td>Sergio Petralia</td>
<td>Jan 19, 2019</td>
</tr>
<tr>
<td>7.2</td>
<td>Terms of Use/Access Changed View Details</td>
<td>Sergio Petralia</td>
<td>Dec 19, 2017</td>
</tr>
<tr>
<td>7.1</td>
<td>Citation Metadata: Author (3 Changed); Contact (1 Changed); Additional Citation Metadata: (5 Changed); View Details</td>
<td>Sergio Petralia</td>
<td>Oct 4, 2017</td>
</tr>
<tr>
<td>7.0</td>
<td>Files (Added: 1; Removed: 1; Replaced: 1); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 12, 2017</td>
</tr>
<tr>
<td>6.1</td>
<td>Citation Metadata: Notes (Changed); View Details</td>
<td>Sergio Petralia</td>
<td>Jan 10, 2017</td>
</tr>
<tr>
<td>6.0</td>
<td>Files (Added: 2; Removed: 2); View Details</td>
<td>Sergio Petralia</td>
<td>Jan 10, 2017</td>
</tr>
<tr>
<td>5.0</td>
<td>Files (Added: 1; Removed: 1); View Details</td>
<td>Sergio Petralia</td>
<td>Oct 3, 2016</td>
</tr>
<tr>
<td>4.0</td>
<td>Citation Metadata: Description (1 Changed); Additional Citation Metadata: (1 Added, 8 Changed); Files (Added: 2; Removed: 2; Changed File Metadata: 1); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 29, 2016</td>
</tr>
<tr>
<td>3.3</td>
<td>Additional Citation Metadata: (5 Changed); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 7, 2016</td>
</tr>
<tr>
<td>3.2</td>
<td>Additional Citation Metadata: (2 Added); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 7, 2016</td>
</tr>
<tr>
<td>3.1</td>
<td>Additional Citation Metadata: (5 Added); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 7, 2016</td>
</tr>
<tr>
<td>3.0</td>
<td>Files (Added: 1; Removed: 1); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 7, 2016</td>
</tr>
<tr>
<td>2.1</td>
<td>Citation Metadata: Notes (Changed); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 1, 2016</td>
</tr>
<tr>
<td>2.0</td>
<td>Files (Added: 1); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 1, 2016</td>
</tr>
<tr>
<td>1.1</td>
<td>Additional Citation Metadata: (2 Added); View Details</td>
<td>Sergio Petralia</td>
<td>Sep 1, 2016</td>
</tr>
<tr>
<td>1.0</td>
<td>This is the first published version.</td>
<td>Sergio Petralia</td>
<td>Aug 26, 2016</td>
</tr>
</tbody>
</table>
## Principle 4: To Be Reusable

### Sub-Principle R1.3: (meta)data meet domain-relevant community standards

<table>
<thead>
<tr>
<th>PRINCIPLE R1</th>
<th>DATAVERSE IMPLEMENTATION</th>
</tr>
</thead>
</table>
| **Sub-Principle R1.3:** (meta)data meet domain-relevant community standards | • DDI for social science data  
• FITS for astronomy data  
• Metadata blocks for other community standards  
• File format conversion to reusable formats (tabular) |
Metadata extracted from Astronomy FITS files
Conclusions
<table>
<thead>
<tr>
<th>FAIR AUTOMATED TEST</th>
</tr>
</thead>
</table>

- **Dataverse scored well in Findable, Accessible, and Reusable principles**
- **Since this FAIR test (in 2018), all Findable principles have been implemented in Dataverse**
- **Interoperable principles support is now under improvement**

<table>
<thead>
<tr>
<th>Identifier type</th>
<th>Question</th>
<th>Dataverse Q'aire</th>
<th>Dataverse Optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F1A</strong></td>
<td>2</td>
<td>DOI</td>
<td>DOI</td>
</tr>
<tr>
<td><strong>F1B</strong></td>
<td>Not tested in Q'aire</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F2A</strong></td>
<td>4A</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F2A</strong></td>
<td>4B</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F3</strong></td>
<td>5B</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F4</strong></td>
<td>6A</td>
<td></td>
<td>FIXED</td>
</tr>
<tr>
<td><strong>F4</strong></td>
<td>6B</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A1.1</strong></td>
<td>7A</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A1.2</strong></td>
<td>8A</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>A1.2</strong></td>
<td>8B</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>A2</strong></td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>I1</strong></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>I2</strong></td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>I3</strong></td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>R1.1</strong></td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>R1.2</strong></td>
<td>14A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Improving Interoperable Principles: Controlled Vocabularies

User selects from the available list of ontologies for a specific field, for example, *Keyword* (configurable)

Contribution by Vyacheslav Tikhonov under review
After a vocabulary for Keyword is selected, suggested terms will appear (information obtained through external API, per language and vocabulary).

Contribution by Vyacheslav Tikhonov under review
Improving Interoperable Principles: Controlled Vocabularies

The Vocabulary Term URI is filled automatically upon Term selection.

Contribution by Vyacheslav Tikhonov under review
The Dataverse software:

- Provides strong support for FAIR principles
- Is adding standard vocabularies and ontologies to improve interoperability

Software support is not enough and there is always room for improvement:

- FAIR principles are not a standard, they are guidelines
- Support for FAIR is a process, it can always be improved
- FAIR data enables the use of automated tools and AI algorithms on the data
- Software support is not everything, data curation remains important
- Beyond FAIR, we need to consider data quality and responsible data sharing
Gràcies

Merce Crosas, Ph.D., Harvard University
scholar.harvard.edu/mercecrosas @mercedrosas