The FAIR Guiding Principles: Implementation in Dataverse

Mercè Crosas, Ph.D.

Harvard University Research Data Officer, OVPR

Chief Data Science and Technology Officer, IQSS

Three Years Ago …

nature > scientific data > comment > article

SCIENTIFIC DATA

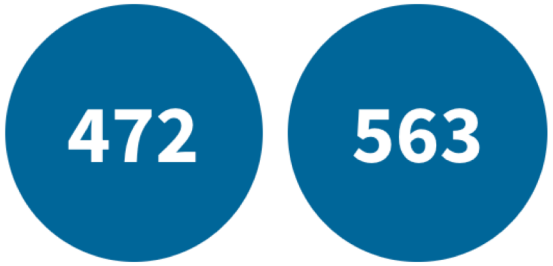
Comment | OPEN | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons ✉ - Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Download Citation ⬇

Total citations



Online attention

Altmetric score (what's this?)

- Tweeted by **1154**
- Blogged by **64**
- On **16** Facebook pages
- Mentioned in **6** Google+ posts
- Picked up by **81** news outlets

⊞ Show more

" A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles."

Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. 2016

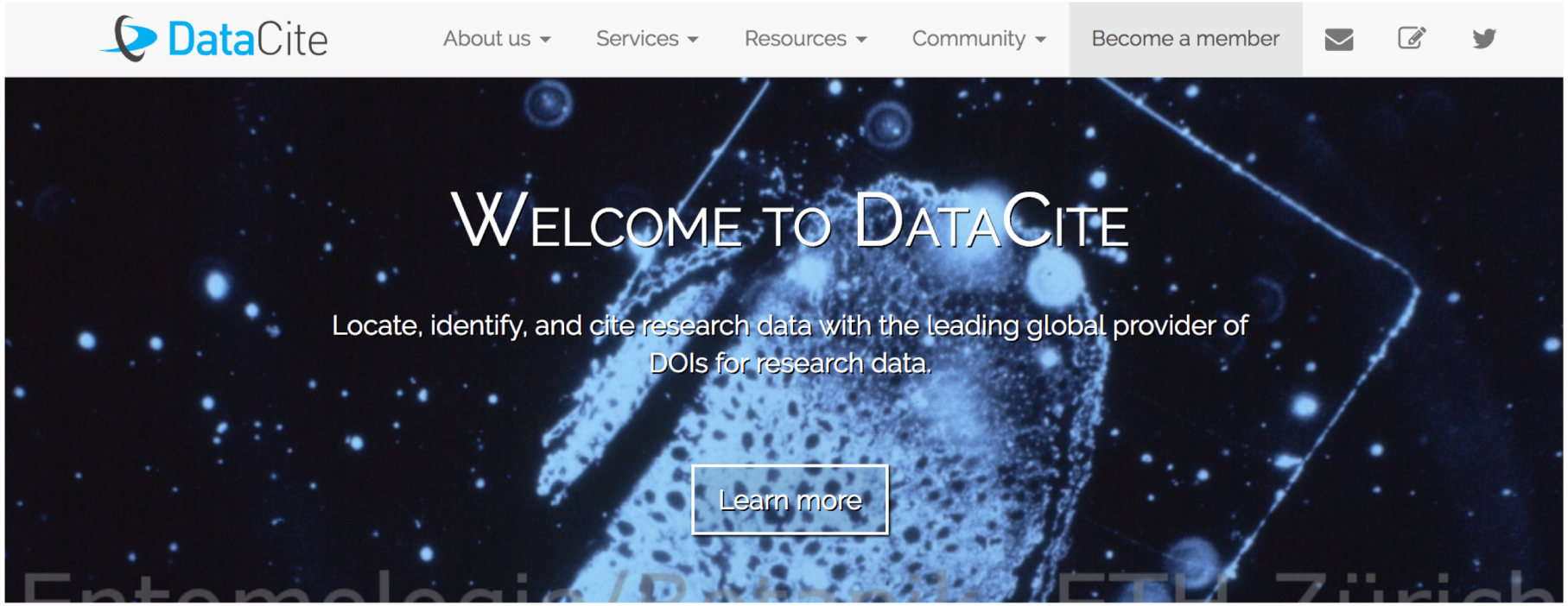
"The FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. "

Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. 2016

Findable

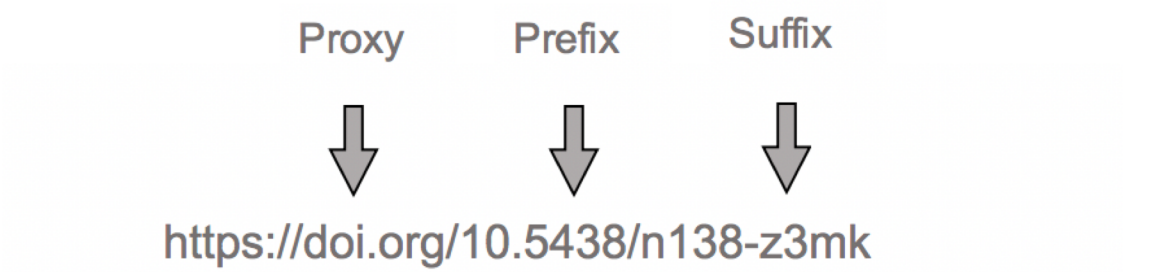
- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by the Reusable principles below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

DataCite helps Achieve Findable Principles



DataCite DOIs are Globally Unique and Persistent

- 14 Million public DOIs registered to DataCite; 5 Million for datasets
- DOIs include:
  - **Name:** Proxy + Prefix + Suffix
  - **Metadata:** description of the object
  - **URL:** resolves to a digital location, which contains object's details



DataCite Metadata

```
<!-- REQUIRED FIELDS -->
<xs:element name="identifier">...</xs:element>
<xs:element name="creators">...</xs:element>
<xs:element name="titles">...</xs:element>
<xs:element name="publisher">...</xs:element>
<xs:element name="publicationYear">...</xs:element>
<xs:element name="resourceType">...</xs:element>
```

- **Identifier:** A persistent identifier that identifies a resource. ("DOI")
- **Creators:** The main researchers involved working on the data, or the authors of the publication in priority order. May be a corporate/institutional or personal name.
- **Titles:** A name or title by which a resource is known.
- **Publisher:** The name of the entity that holds, archives, publishes prints, distributes, releases, issues, or produces the resource. (publish = making data available to the community)
- **Publication Year:** Year when the data is made publicly available. If an embargo period has been in effect, use the date when the embargo period ends.
- **Resource Type:** The type of a resource. ("dataset")

Citation Metadata across Standards

Citation Metadata	Dublin Core[a]	Schema.org[b]	DataCite[c]	DATS[d]
Dataset Identifier	identifier	@id*	identifier	identifier
Title	title	name	title	title
Creator**	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication Date	date	datePublished	publicationYear	date
Version	*not available*	version	version	version
Type	type	type	resourceTypeGeneral	type

Fenner, Crosas, et al., 2019. A data citation roadmap for scholarly data repositories. Nature Scientific Data, ***forthcoming***

Discovery Metadata across Standards

Discovery Metadata	Dublin Core	Schema.org	DataCite	DATS
Description	description	description	description	dataType dimension Material ...*
Keywords	subject	keywords	subject	keywords
License	license	license	rights	license
Related Dataset**	isPartOf isVersionOf references	isPartOf citation	relatedIdentifier	isPartOf
Related Publication***	bibliographicCitation	citation	relatedIdentifier	publication

Fenner, Crosas, et al., 2019. A data citation roadmap for scholarly data repositories. Nature Scientific Data, ***forthcoming***

Findable in Google Dataset Search

- Add metadata in schema.org to each page that describes a dataset
- Verify that the markup produces structured data that you expect in Structured Data Testing Tool
- If you have multiple pages, create a sitemap and add that sitemap to your Search Console

Accessible

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

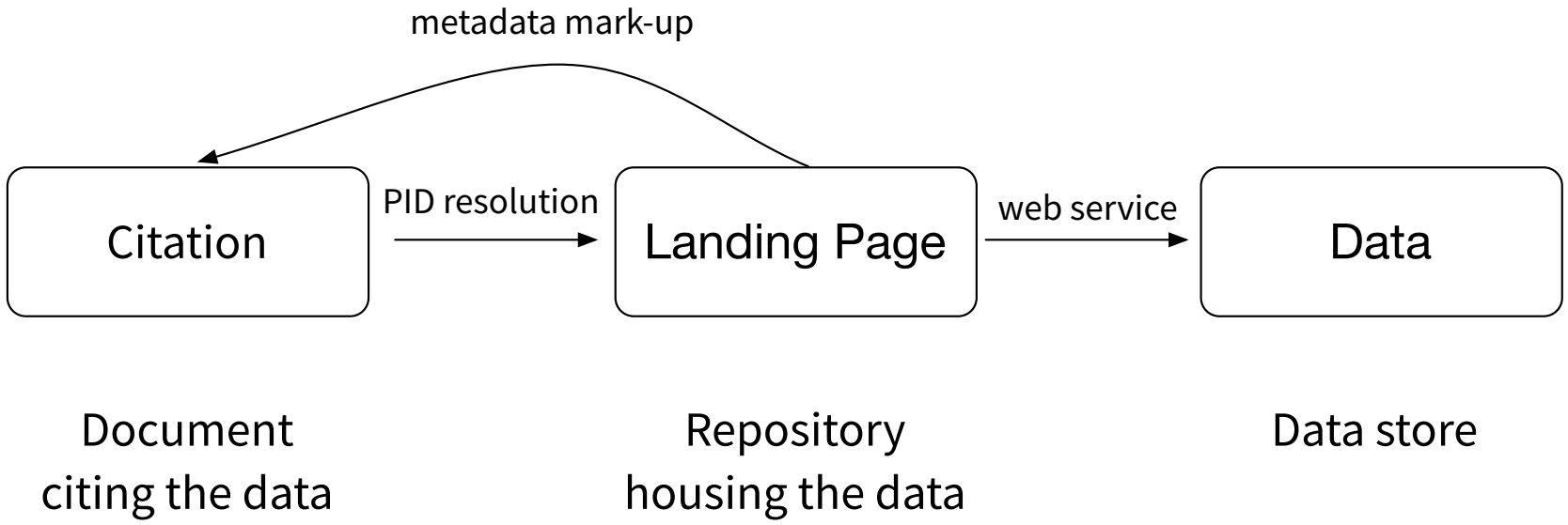


Fig. 1. Generic data citation - relationships of the citation reference, repository landing page and underlying data, taken with permission from[20]. The Document citing the data is linked to the landing page via persistent identifier (PID) resolution, the content is linked from the dataset landing page via web services.

Fenner, Crosas, et al., 2019. A data citation roadmap for scholarly data repositories. Nature Scientific Data, ***forthcoming***

Achieving Accessible Principles

- A dataset landing page should contain:
  - metadata
  - information on how to access the data via an open standard protocol
- Data might be restricted, and only accessible if granted permissions
- The landing page with descriptive metadata exists in perpetuity, even when data are not available in the repository.
  - In this case, the landing page should provide information about why the data are not available

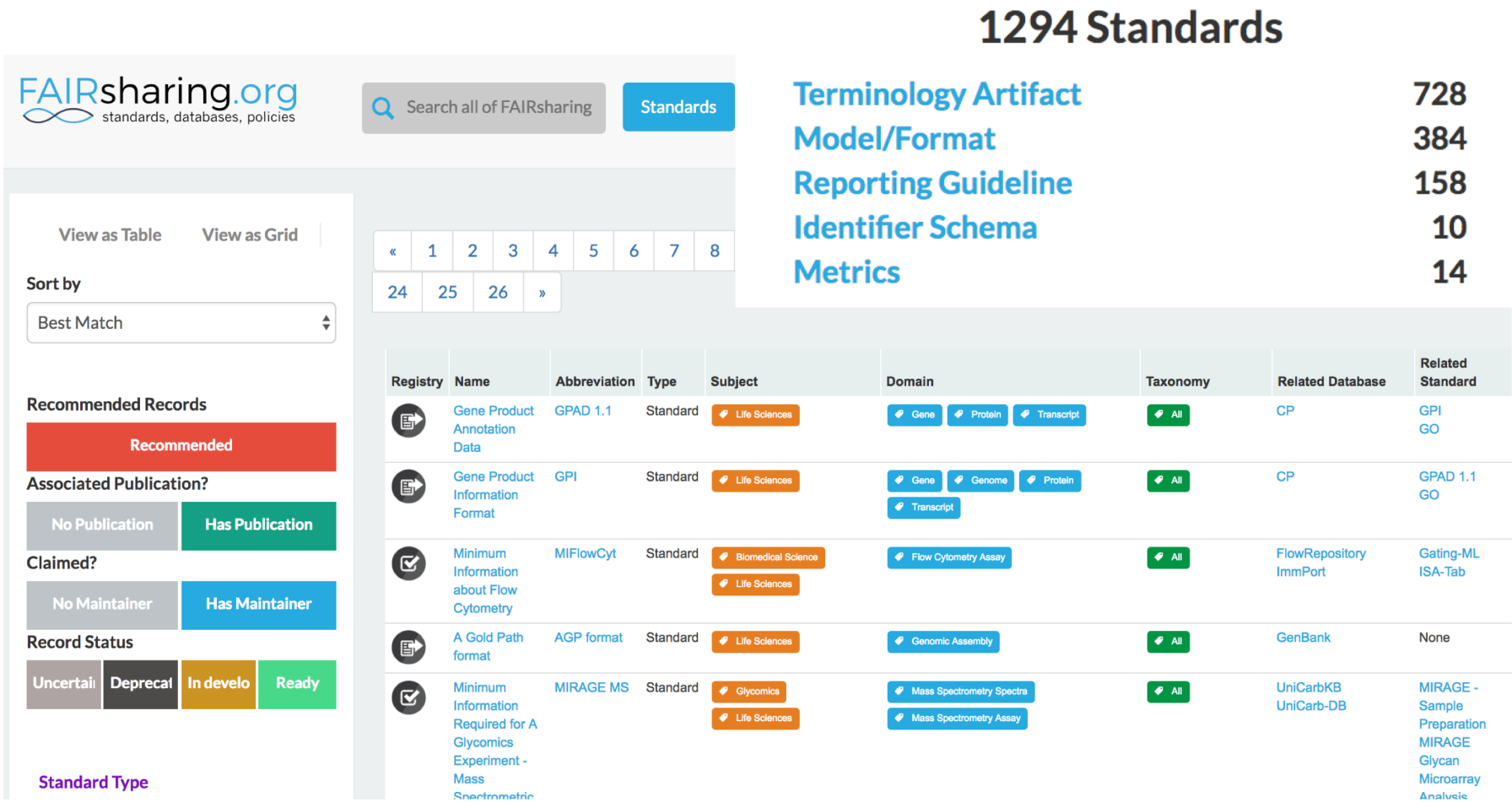
Interoperable

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

Achieving Interoperable Principles

- Interoperable Principles are the hardest to achieve!
- It's not sufficient to bring the data together via APIs
- Main goal is to facilitate Data Harmonization:
  - "**Data harmonization** is the process of bringing together your data of varying file formats, naming conventions, and columns, and transforming it into one cohesive data set." (Datorama)
  - "**Data harmonization** refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies." (ICPSR)
  - **"Retrospective data harmonization** involves pooling heterogeneous data from disparate data sets and transforming them into 1 common data set to use for new analyses. This can be done either by using variables that are common (e.g., age or sex) or similar (e.g., smoking status) across data sets or by deriving new variables from disparate questions with a common theme (e.g., total alcohol intake). " (Rolland et al. 2015, Towards rigorous data harmonization in cancer research: one approach" American Journal of Epidomology)

Metadata, Ontologies, and Vocabularies Standards help Achieve Interoperable Principles



Reusable

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

Metadata standards also help Achieve Reusable Principles

- High-level, Descriptive metadata standards:
  - Schema.org
  - DataCite
  - Dublin Core
- Domain-specific metadata and ontologies, such as:
  - **For social and health sciences**: Data Documentation Initiative (DDI)
  - **For life sciences:** DATS
  - **For molecular biology and genetics:** The Gene Ontology
- Provenance metadata:
  - W3C PROV

Data Documentation Initiative (DDI)

- Describes data that result from observational methods in the social, behavioral, economic, and health sciences
- Created in 1995. Widely used, in 80 countries.

DDI Lifecycle 3.2

DDI-Lifecycle is designed to document and manage data across the entire life cycle, from conceptualization to data publication and analysis and beyond. It encompasses all of the DDI-Codebook specification and extends it. Based on XML Schemas, DDI-Lifecycle is modular and extensible.

Implementation Guide: Best Practices for Usage of DDI 3.2 and Future Versions

Read more about DDI 3.2...

Online field level documentation

XSD Schema entry point

Download documentation

DDI Codebook 2.5

DDI-Codebook is a more light-weight version of the standard, intended primarily to document simple survey data. Originally DTD-based, DDI-C is now available as an XML Schema.

Read more about DDI 2.5...

Online field level documentation

XSD Schema entry point

Download documentation

DDI Codebook Metadata

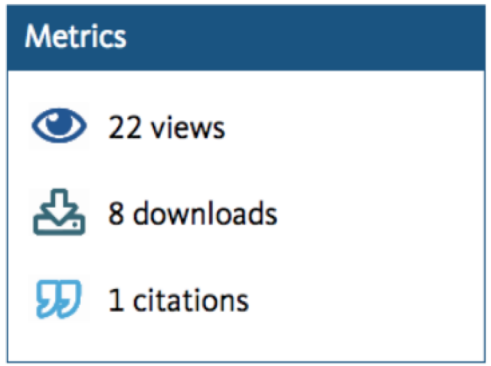
- Includes variable descriptions:
  - List of variables
  - Variable labels or names
  - Variable format, intervals, weights
  - Summary Statistics
- Includes usage licenses and terms
- Facilitates machine-actionable Interoperability and Reusability, in particular for tabular datasets

Reusable Data with Clear Licenses and Terms

- **CCo:**
  - " I hereby waive all copyright and related or neighboring rights together with all associated claims and causes of action with respect to this work to the extent possible under the law"
  - Citation of a dataset is expected as a scholarly norm, not by law
- **CC BY:**
  - "You must give appropriate credit, provide a link to the license, and indicate if changes were made. "
  - It's a license, not a waiver as CCo
- **Data Use Agreements (DUA):**
  - Used when data are restricted due to proprietary or privacy concerns.
  - Also referred to as a License Agreement, Confidentiality Agreement, Non-Disclosure Agreement, Memorandum of Understanding, Memorandum of Agreement.

Make Data Count: Reusable Data with Credit

- A new code of practice regarding how data usage and impact should be meaningfully measured and reported
- A data-level metrics aggregation and publication service
  - **Views and downloads:** from repository logs
  - **Data Citations:** from Crossref event data



University of California CDL California Digital Library

DataCite

DataONE

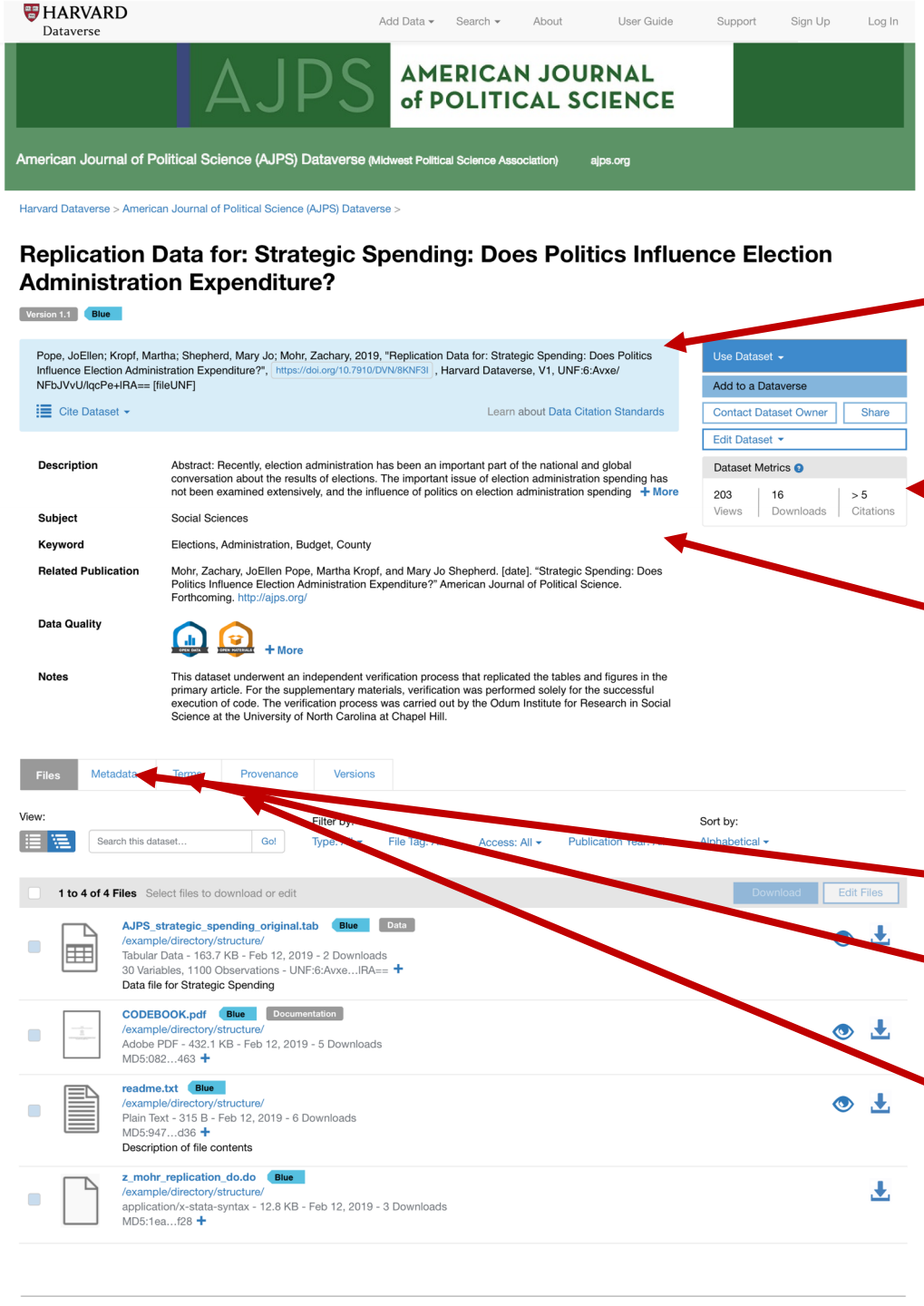
The Dataverse® Project

Open source research data repository software

A software, a community, many repositories

- Developed since 2006 at Harvard's Institute for Quantitative Social Science
- Over 80 contributors, most external to Harvard
- 12 releases a year
- 40 installations around the world
- 30,000 datasets deposited in Harvard Dataverse
- + 50,000 datasets harvested from other Repositories
- 250 new datasets added to Harvard Dataverse per month
- 6 million downloads from Harvard Dataverse

Implementation of FAIR Data Principles in Dataverse



Dataset Landing Page

Citation for entire dataset. DOI, with URL, and metadata registered to DataCite. (**Findable** and **Accessible**)

Make Data Count (**Reusable** metric)

Citation and discoverable metadata using DataCite, schema.org, Dublin Core, DDI standards (**Findable**, **Accessible** and **Reusable**)

More metadata, including domain-specific (**Reusable**)

Terms with usage license or Data Use Agreement (**Reusable**)

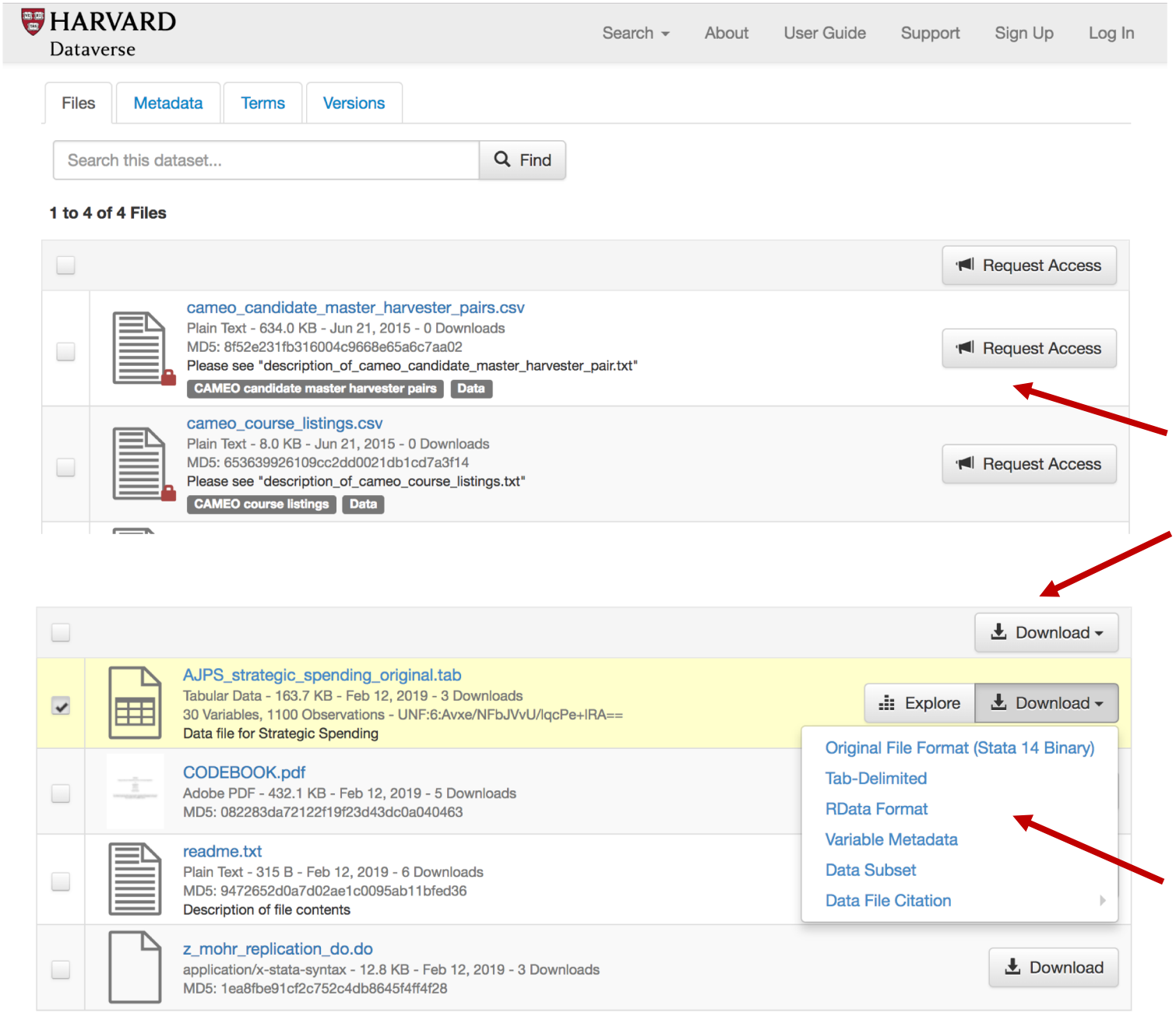
PROV metadata (**Reusable**)

Machine-actionable Dataset Landing Page


```
<title>Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure? - American Journal of Political Science (AJPS) Dataverse</title>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="DC.identifier" content="doi:10.7910/DVN/8KNF3I" />
<meta name="DC.type" content="Dataset" />
<meta name="DC.title" content="Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?" />
<meta name="DC.date" content="2019-03-11" />
<meta name="DC.publisher" content="Harvard Dataverse" />
<meta name="DC.description" content="Abstract: Recently, election administration has been an important part of the national and global conversation about the results of elections. The important issue of election administration spending has not been examined extensively, and the influence of politics on election administration spending levels has not been examined in the United States. While theories of voter turnout and policy preference suggest that politics should influence election administration spending levels in the counties that administer elections, to our knowledge, there has been no evidence produced to support a partisan election administration expenditure effect. This research finds that Republican county commissions in North Carolina spend significantly less on election administration once the county electorate is a sufficient Republican majority. The paper presents a novel model and method for estimating election administration spending and calls for additional research to examine the outcomes of these significant differences in spending on election administration." />
<meta name="DC.creator" content="Pope, JoEllen" />
<meta name="DC.creator" content="Kropf, Martha" />
<meta name="DC.creator" content="Shepherd, Mary Jo" />
<meta name="DC.creator" content="Mohr, Zachary" />
<meta name="DC.subject" content="Social Sciences" />
<script type="application/ld+json">
{"@context":"http://schema.org","@type":"Dataset","@id":"https://doi.org/10.7910/DVN/8KNF3I","identifier":"https://doi.org/10.7910/DVN/8KNF3I","name":"Replication Data for: Strategic Spending: Does Politics Influence Election Administration Expenditure?","creator":[{"name":"Pope, JoEllen","affiliation":"University of North Carolina at Charlotte"},{"name":"Kropf, Martha","affiliation":"University of North Carolina at Charlotte"},{"name":"Shepherd, Mary Jo","affiliation":"University of North Carolina at Charlotte"},{"name":"Mohr, Zachary","affiliation":"University of North Carolina at Charlotte"}],"author":[{"name":"Pope, JoEllen","affiliation":"University of North Carolina at Charlotte"},{"name":"Kropf, Martha","affiliation":"University of North Carolina at Charlotte"},{"name":"Shepherd, Mary Jo","affiliation":"University of North Carolina at Charlotte"},{"name":"Mohr, Zachary","affiliation":"University of North Carolina at Charlotte"}],"datePublished":"2019-02-12","dateModified":"2019-03-11","version":"1","description":["Abstract: Recently, election administration has been an important part of the national and global conversation about the results of elections. The important issue of election administration spending has not been examined extensively, and the influence of politics on election administration spending levels has not been examined in the United States. While theories of voter turnout and policy preference suggest that politics should influence election administration spending levels in the counties that administer elections, to our knowledge, there has been no evidence produced to support a partisan election administration expenditure effect. This research finds that Republican county commissions in North Carolina spend significantly less on election administration once the county electorate is a sufficient Republican majority. The paper presents a novel model and method for estimating election administration spending and calls for additional research to examine the outcomes of these significant differences in spending on election administration."],"keywords":["Social
```


Dublin Core meta-tags for citation metadata (**Findable** and **Accessible**)

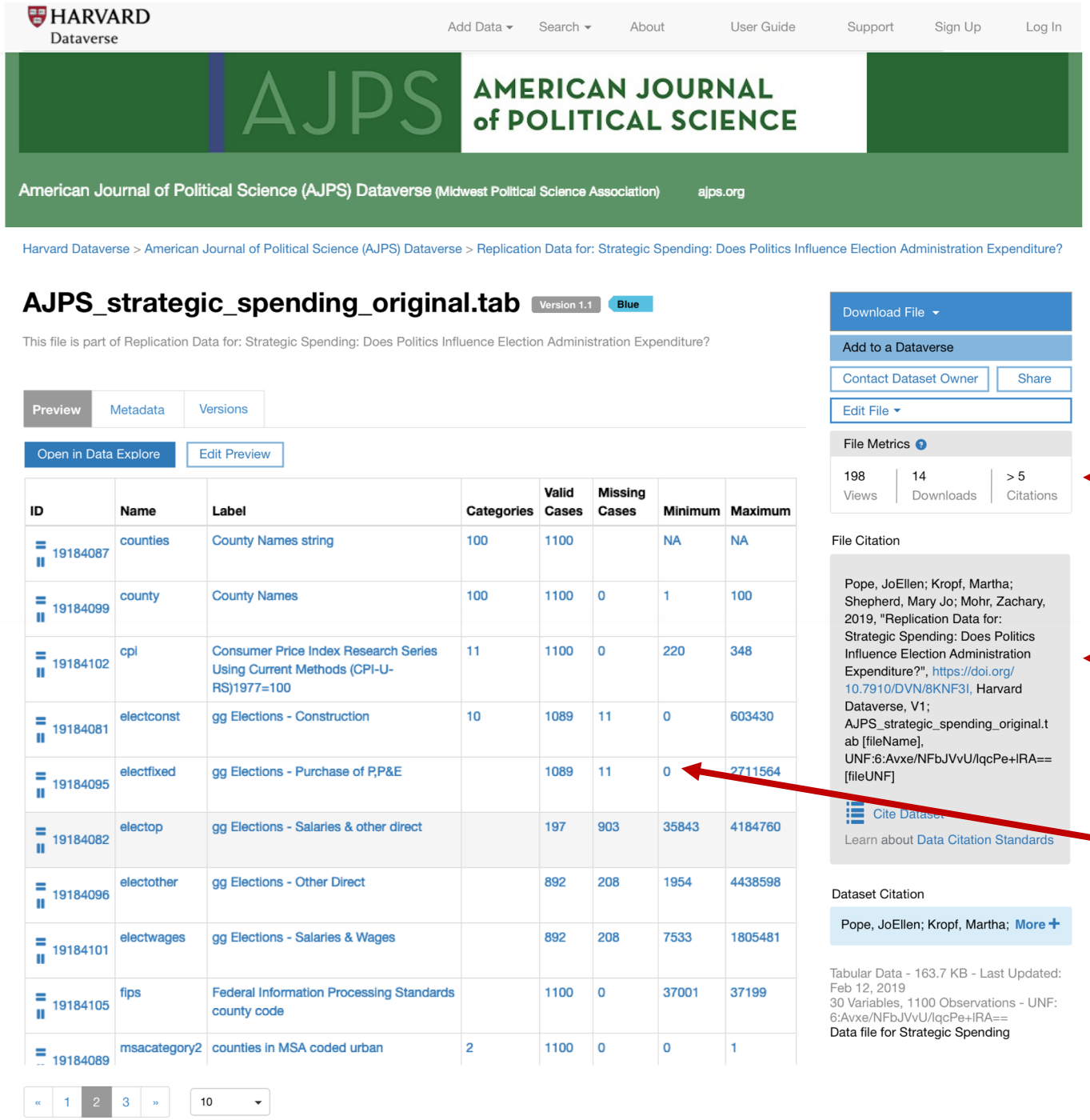
Schema.org JSON-LD (**Findable** in Google Dataset Search)



Restricted data files (Authentication and Authorization needed)
vs.
Open data files. (**Accessible**)

Landing Page with metadata and usage terms in both cases. (**Accessible**)

Domain-relevant standards. (**Reusable**)

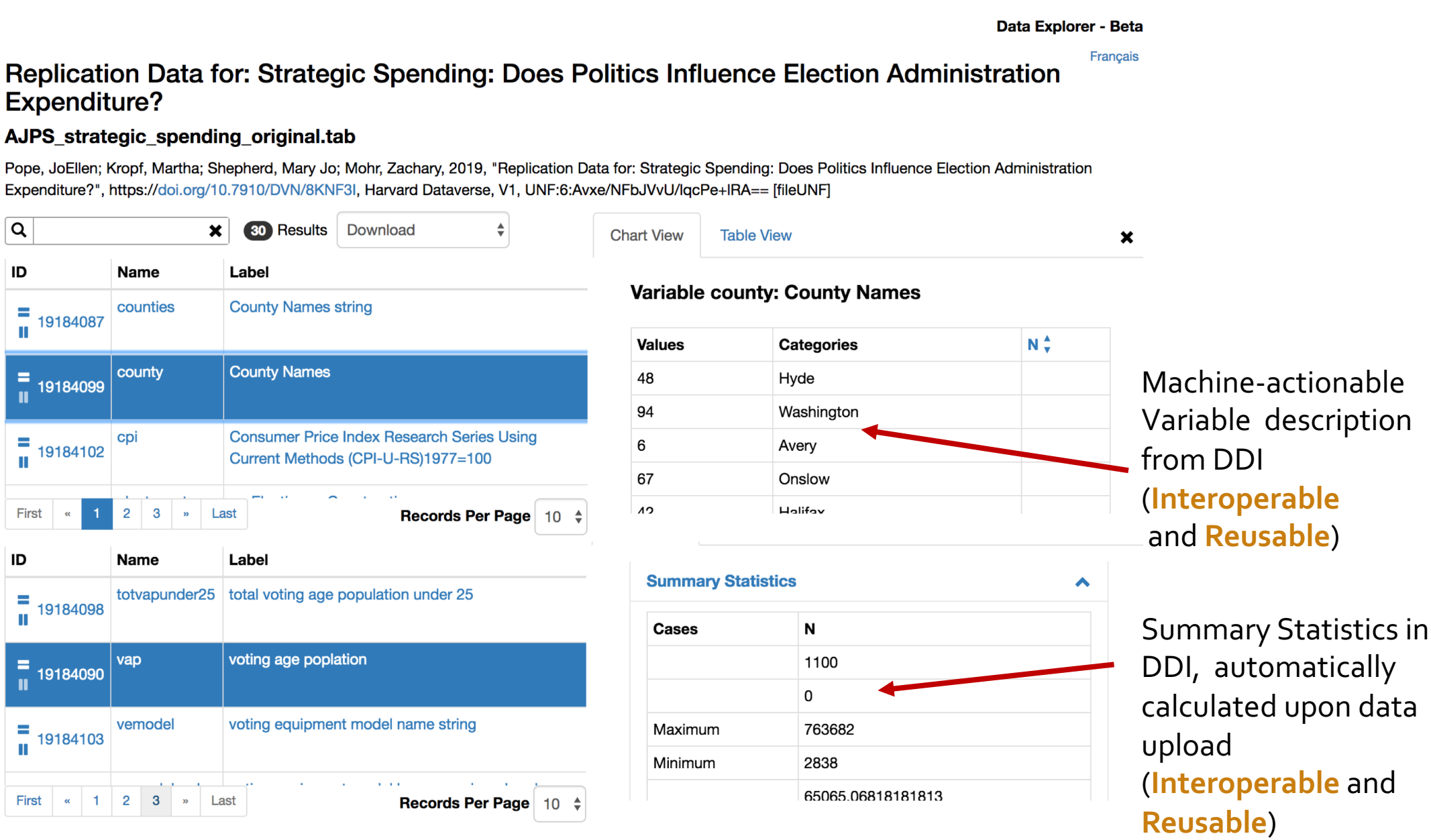


Data File Landing Page

Make Data Count s (**Reusable** metric)

Citation for data file, with DOI and URL for each file (**Findable** and **Accessible**)

Variable metadata for tabular data file, using DDI standard (**Interoperable** and **Reusable**)



Machine-actionable Variable description from DDI (**Interoperable** and **Reusable**)

Summary Statistics in DDI, automatically calculated upon data upload (**Interoperable** and **Reusable**)

Takeaways

- FAIR data (Findable, Accessible, Interoperable, Reusable):
  - Should be machine-readable and actionable
  - Is not equal to open data
  - Is an aspiration; it's never 100% FAIR
- When publishing restricted data, licenses and Data Use Agreements must be clearly set by data authors or providers
- A repository platform such as Dataverse can greatly facilitate making research data FAIR
- But, data authors must contribute using appropriate community metadata and vocabularies standards

Thank you

@mercecrosas

https://scholar.harvard.edu/mercecrosas