Research Data Management with Dataverse

Haverford College Symposium, March 10, 2021

Mercè Crosas, Ph.D., Harvard University
University Research Data Management Officer, HUIT
Chief Data Science and Technology Officer, IQSS
scholar.harvard.edu/mercecrosas @mercecrosas

IQSS
The Institute for Quantitative Social Science

HARVARD UNIVERSITY

- **Global advances in Research Data Sharing**
- **The Landscape of Research Data and Computing Services in Research Institutions**
- **Findable, Accessible, Interoperable, Reusable (FAIR) data with Dataverse**

Recent Advances in Data Sharing

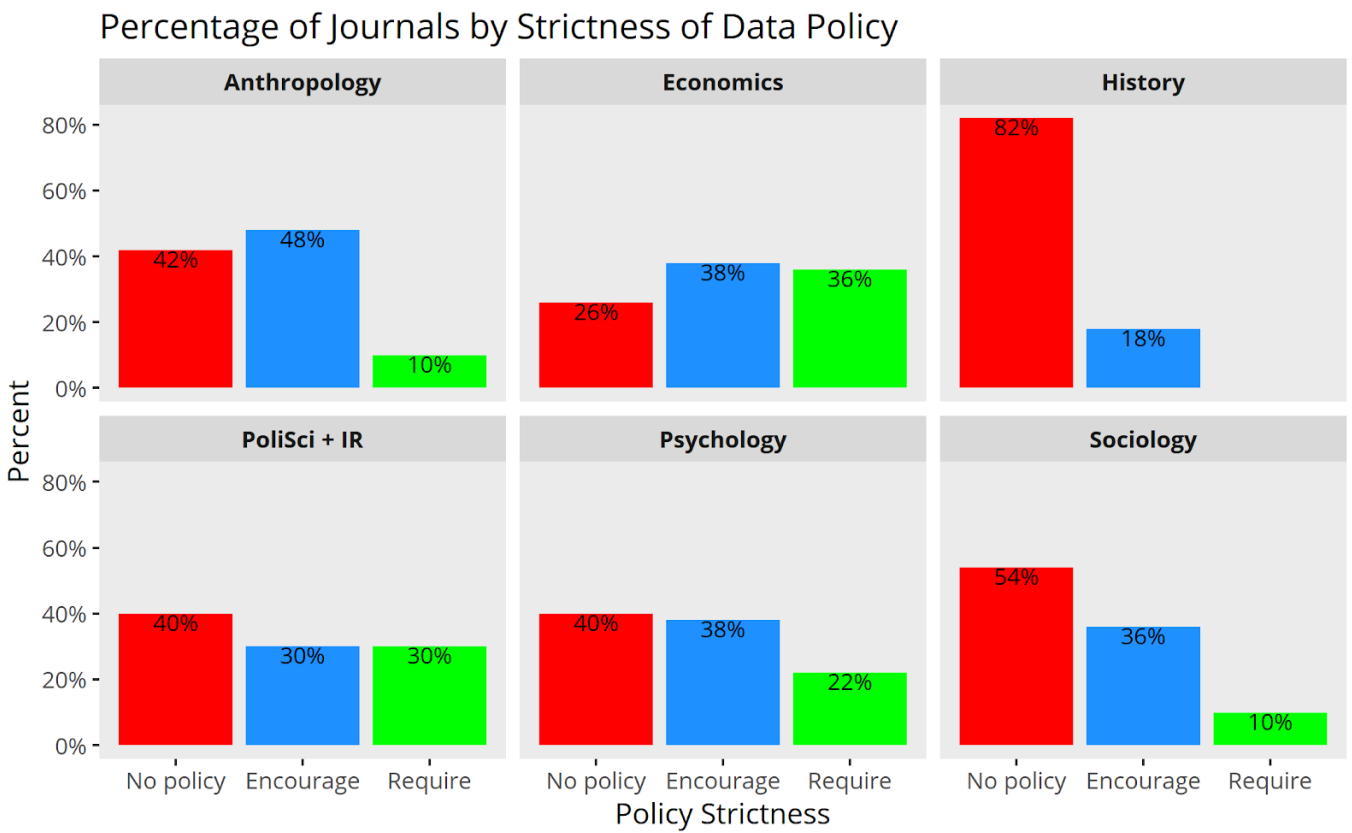
- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software powers > 60 repositories world-wide

Data Policies of top 50 journals in 6 disciplines

No Policy

Encourage data sharing

Require data sharing



Crosas, Gautier, Karcher, Kirilova, Otalora, Schwartz. Data Policies of Highly-Ranked Social Science Journals, *preprint,* https://osf.io/preprints/socarxiv/9h7ay

Recent Advances in Data Sharing

- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software powers > 60 repositories world-wide

Final NIH Policy for Data Management and Sharing

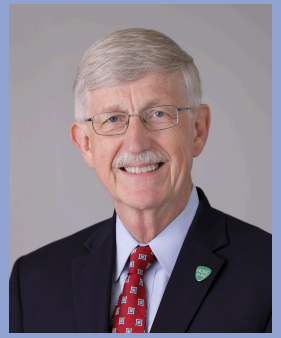
Notice Number:
NOT-OD-21-013

Key Dates

Release Date:
Effective Date:
October 29, 2020
January 25, 2023

Issued by

Office of The Director, National Institutes of Health (OD)



"This policy establishes the baseline expectation that data sharing is a fundamental component of the research process"

Francis S. Collins, M.D., Ph.D.
Director, National Institutes of Health

"[…] NIH encourages data management and sharing practices to be consistent with the **FAIR (Findable, Accessible, Interoperable, and Reusable)** data principles and reflective of practices within specific research communities."

Recent Advances in Data Sharing

- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software powers > 60 repositories world-wide

AGU ADVANCING EARTH AND SPACE SCIENCE

JOIN RENEW GIVE LOGIN

“Robust, verifiable, and reproducible science requires that evidence behind an assertion be accessible for evaluation. Researchers have a responsibility to collect, develop, and **share this evidence** in an ethical manner, that is **as open and transparent as possible**.”

Recent Advances in Data Sharing

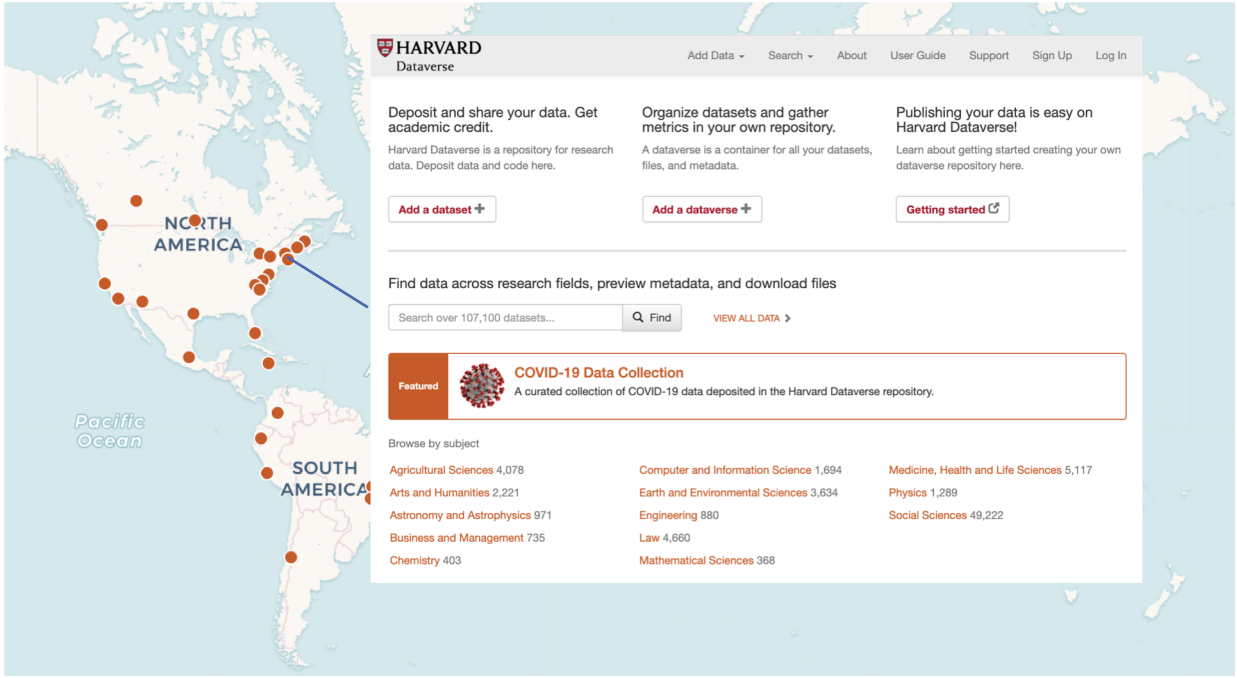
- **New data policies in journals**
  - *Example:* > 50% of top social science journals recommend or require sharing the data associated with the article
- **New data sharing mandates by funding entities**
  - *Example:* National Institutes of Health (NIH) recent release of Policy for Data Management and Sharing
- **Joint statements from scientific communities**
  - *Example:* American Geophysical Union (AGU) Position Statement on Data
- **Ubiquity of domain-specific and generalist data repositories**
  - *Example:* Dataverse software platform powers > 60 repositories worldwide

Federated FAIR data repositories worldwide

Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community and the Global Dataverse Community Consortium (https://dataverse.org)

- **Open-source**
- **67** installations
- **6** continents
- **8K** Dataverse collections
- **140K** datasets
- **900K** files
- **30M** file downloads
- **Metadata** shared across repositories

Harvard Dataverse repository



Open to all researchers across all disciplines

- Global advances in Research Data Sharing
- **The Landscape of Research Data and Computing Services in Research Institutions**
- Findable, Accessible, Interoperable, Reusable (FAIR) data with Dataverse

Review of Research Data Services in U.S. Universities

- Ithaka S+R report (Radecki & Springer, 2020, https://doi.org/10.18665/sr.314397):
  - Reviewed research data services from 120 U.S. Universities and Colleges
- A growing number of research data services distributed across various university units:

Within Libraries and IT (main providers)

- Consulting
- Training events
- Backend work (data architecture, metadata design)
- Front end work (web development, data visualizations)

Outside Libraries and IT

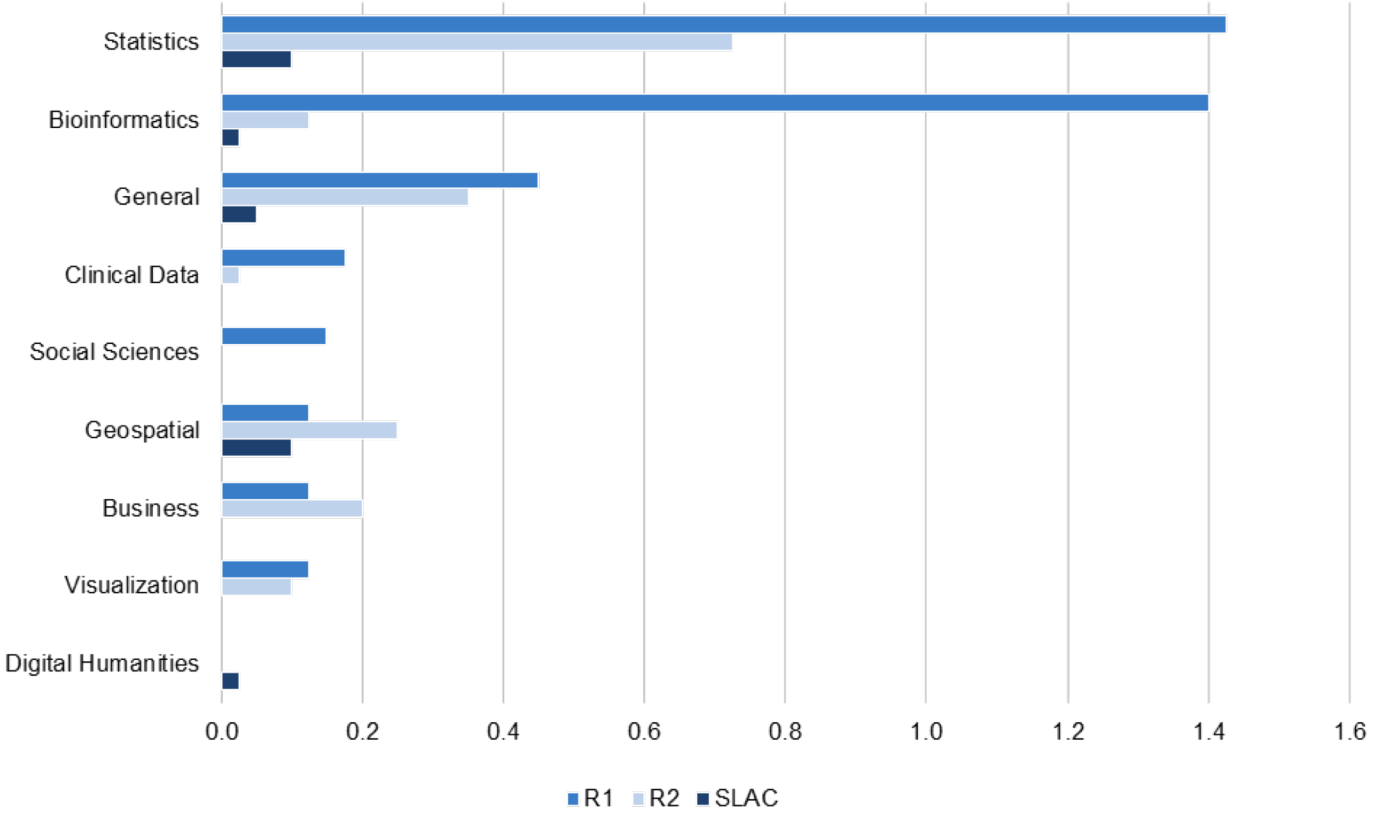
- Statistics
- Bioinformatics
- Geospatial
- Clinical data
- Business
- Social Science
- Visualizations

Profile of types of Library Data Services

Generalist consultation is the most common service offered by the libraries

	Consulting	Training Events	Front End Work	Back End Work	Total
General	35.9%	16.0%	3.2%	2.6%	**57.7%**
Geospatial	16.7%	9.0%	0.0%	0.0%	**25.6%**
Statistics	7.1%	1.3%	0.0%	0.0%	**8.3%**
Digital Humanities	2.6%	1.3%	0.0%	0.0%	**3.8%**
Social Sciences	0.6%	0.6%	0.0%	0.0%	**1.3%**
Health Sciences	0.6%	0.0%	0.0%	0.0%	**0.6%**
Other	1.3%	1.3%	0.0%	0.0%	**2.6%**
Total	**64.7%**	**29.5%**	**3.2%**	**2.6%**	**100%**

Average number of research data services per institution offered by centers and facilities, departments, and schools



Research Data and Computing Services at Harvard

- **Increase** of research data and computing services
  - Along with increase in data-centric and data science research
  - To support funders and journals requirements
  - Services distributes across units and schools
- **Collaboration** between Research, Library, and IT/Research Computing is key
- **Build a catalog** to learn what services are provided across units
  - Standardize the information
  - Unify services when possible
- **A research support hub** to find all service offerings in a common way *(to be launched in 2021)*
- **Find gaps and connect** services, tools, and teams
- **Foster a community** of research computing and data teams at the University (working groups, events)

Services offerings throughout the research lifecycle

Planning:
Access & Reuse
Plan & Design
(13 service offerings)

Active Research:
Collect & Create
Analyze & Collaborate
(20 service offerings)

Dissemination & Preservation:
Evaluate & Archive
Share & Disseminate
(5 service offerings)

https://researchsupport.harvard.edu/ (to be launched Spring 2021)

SERVICES

- Research Administration & Compliance
  - Data Safety & Regulated Data
  - Data Use Agreement Processing
  - eSupport - Committee on Microbiological Safety (eCOMS)
  - Human Subjects and Animal Research Resources
  - Pre- & Post-Award Resources
- Research Computing
- Research Data and Scholarship

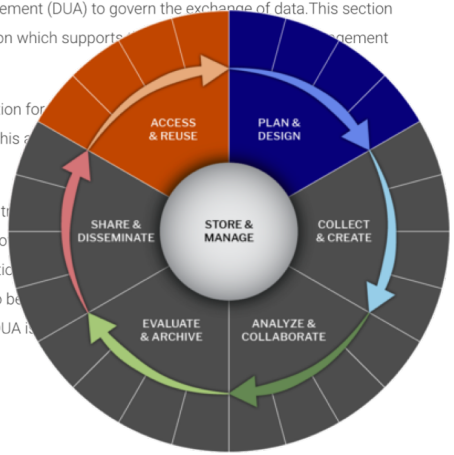
HOME / SERVICES / RESEARCH ADMINISTRATION & COMPLIANCE /

Data Use Agreement Processing

The transfer of data between organizations is common in the research community. When the data is confidential, proprietary, or otherwise considered sensitive or protected, the organization providing the data, whether that is Harvard or a third party, may require a Data Use Agreement (DUA) to govern the exchange of data.This section includes guidance on the DUA-Agreements Application which supports ...gement process for DUAs, and other related resources.

Data Use Agreement review and compliance application for ... from a third party, or providing data to a third party. This ... management process for Data Use Agreements.

A DUA is a binding contract governing access to and t... "Provider") to another party (a "Recipient"). DUAs are o... for Harvard data to be disclosed to another organizatio... and regulations governing the specific type of data to b... Provider and Recipient. If you are unsure whether a DUA i... research office.



Details by Provider

- HUIT Administrative Technology Services, Research Administration and Compliance
- Harvard Medical and Dental Schools
- Harvard T. H. Chan School of Public Health
- Harvard University Area

Example of Service offering in Planning Phase:

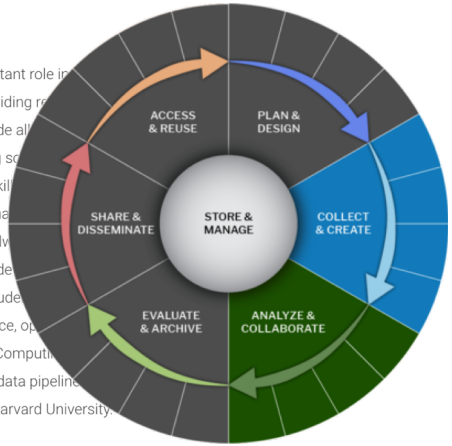
- DUA and Safety System to:
  - Track all DUAs for incoming and outgoing compliant data
  - Manage DUA while data are used for research
- Assistance with DUA negotiation
- Connect process with IRB and Security officers

SERVICES

- Research Administration & Compliance
- Research Computing
  - Cluster Computing
  - Data Science and Research Software Engineering Collaboration
  - Database
  - Research Computing Consulting and Facilitation
  - Research Data Storage
  - Virtual Instances
- Research Data and Scholarship

HOME / SERVICES / RESEARCH COMPUTING /

Data Science and Research Software Engineering Collaboration



Data Science and Software Engineering play an important role in
and analyze data, helping ensure reproducibility, and aiding r
data. The term software here is used broadly to include all
Researchers utilize software in their research by using sc
software. Data science also covers a wide range of skil
processing, and statistics that are typically beyond wha
the rapidly evolving nature of research, there are not alw
sources; therefore, the software or data pipelines are de
development was done with researchers (graduate stude
approach poses several issues in terms of maintenance, op
Scientist team can work closely with other Research Computi
optimize, and maintain software packages/tools and data pipelin
architectures to accelerate cutting-edge research at Harvard University.

Details by Provider

- Faculty of Arts and Sciences, Research Computing
- Institute for Quantitative Social Sciences
- Harvard Business School

Example of Service offering in Active Research phase:

Data Science Services offered by the Institute for Quantitative Social Science (IQSS)

- Focuses on **social science** support, but includes other scientific domains
- **Consulting:** short term
- **Collaboration:** longer project (fee)
- **Training materials** (in collaboration with Harvard Business School):
  - **Python:** Introduction, web scraping
  - **R:** Introduction, regressions models, graphics, data wrangling
  - **Stata:** Introduction, data management, regression models, graphics
  - **Other:** Introduction to programming, SAS introduction, data science tools

SERVICES

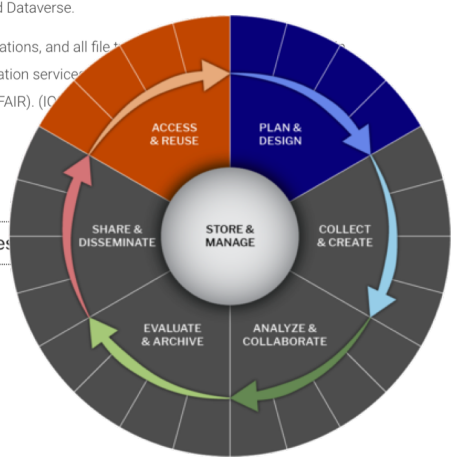
- Research Administration & Compliance
- Research Computing
- Research Data and Scholarship
  - Archiving Faculty Research Data and Archiving Data
  - Buying and Licensing Data
  - Copyright and Intellectual Property
  - Data Cleaning
  - Data Curation
  - Data Deposit
  - Data Handling
  - Data Retrieval
  - Data Security Support
  - Data Sharing and Publishing
  - Data Visualization
  - Dataset Creation
  - Finding Data
  - Geospatial Library, Data Analysis, Creation, Visualization

HOME / SERVICES / RESEARCH DATA AND SCHOLARSHIP /

Harvard Dataverse Curation

The Harvard Dataverse data curation team, staffed by member of IQSS and the Harvard Library (and separately, the Harvard Kennedy School Library), provides fee-based curation services and free consultations to researchers around the world who are depositing data into the Harvard Dataverse.

Research data replication datasets, data for related publications, and all file
the Harvard Dataverse. Through this engagement, the curation service
are discoverable, accessible, interoperable, and reusable (FAIR). (IC



Details by Provider

Institute for Quantitative Social Science

Audience

- All Affiliates
- All Faculty
- All Graduate Students
- All Undergraduate Students
- Public

Service Provider
Institute for Quantitative Social Sciences (IQSS)

Service Fee
Yes

Service Website
https://support.dataverse.harvard.edu/curation-services

Contact Information
support@dataverse.harvard.edu

Example of Service Offering in Dissemination phase:

Dataverse Curation services

- A **collaboration** between IQSS and the Harvard Library
- **Tiered service offerings:**
  - Free consultation (< 3 hours)
  - Extended consultation services
  - Dataverse collection set-up
  - administration and curation services
  - Custom services
- In 2021, new service for supporting “managed collections” interested in receiving **Core Trust Seal** certification.

- Global advances in Research Data Sharing
- The Landscape of Research Data and Computing Services in Research Institutions
- **Findable, Accessible, Interoperable, Reusable (FAIR) data with Dataverse**

Open Access | 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. Sv Velterop, Andra Waagmeester, Pet

-Show fewer authors

Scientific Data **3**, Article number: '

157k Accesses | **1991** Citations

Access & Citations

157k	1482	1991
Article Accesses	Web of Science	CrossRef

Online attention

1329 tweeters	112 blogs	21 Facebook pages
6 Google+ users	107 news outlets	1 Redditors
1 F1000	1 Video uploaders	4 Wikipedia page
3085 Mendeley	20 Citeulike	

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*

15 Principles

- Published in 2016
- 54 authors
- > 1990 citations
- > 157K accesses

Adoption by funding agencies, research communities

Emphasis on Machine-Actionability

" 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. "

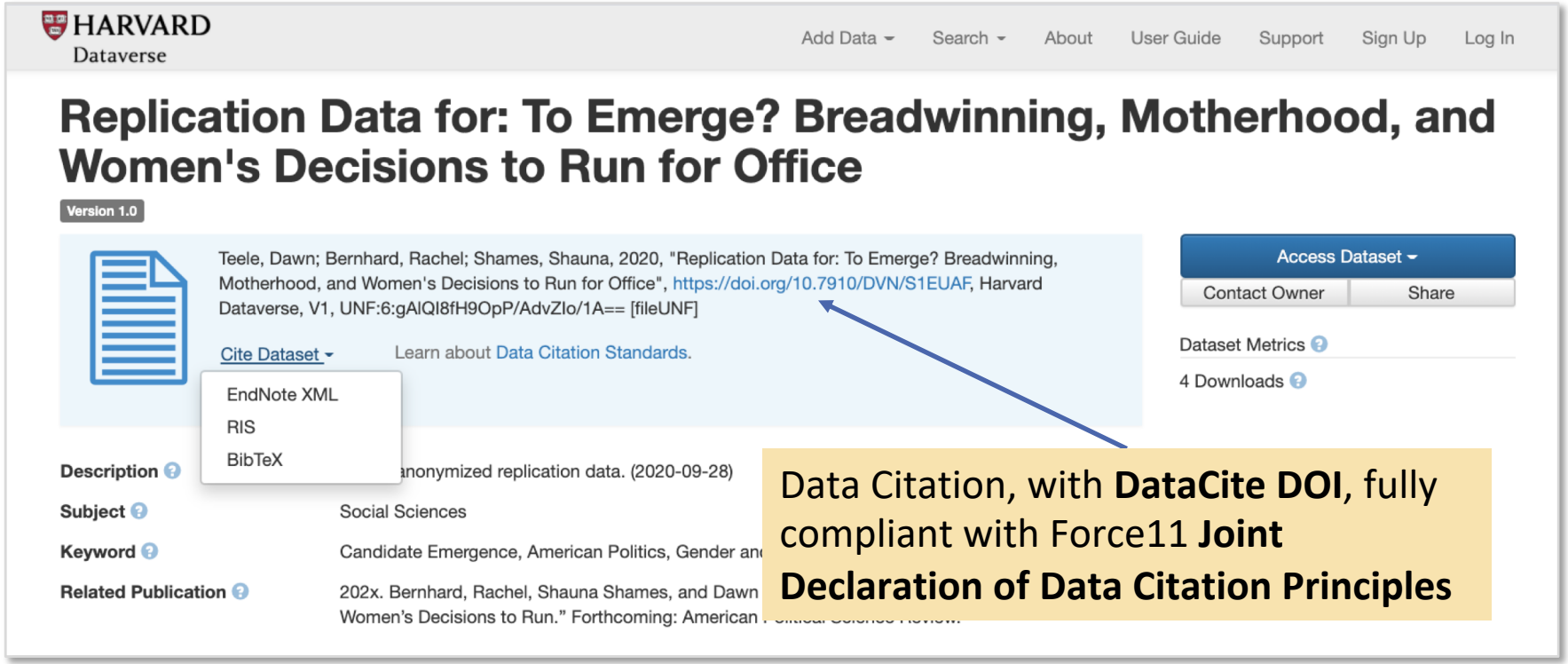
Wilkinson et al. 2016. Nature-Springer Scientific Data. *The FAIR Guiding Principles for Scientific Data Management and Stewardship.* doi: 10.1038/sdata.2016.18

To Be Findable (4 principles)

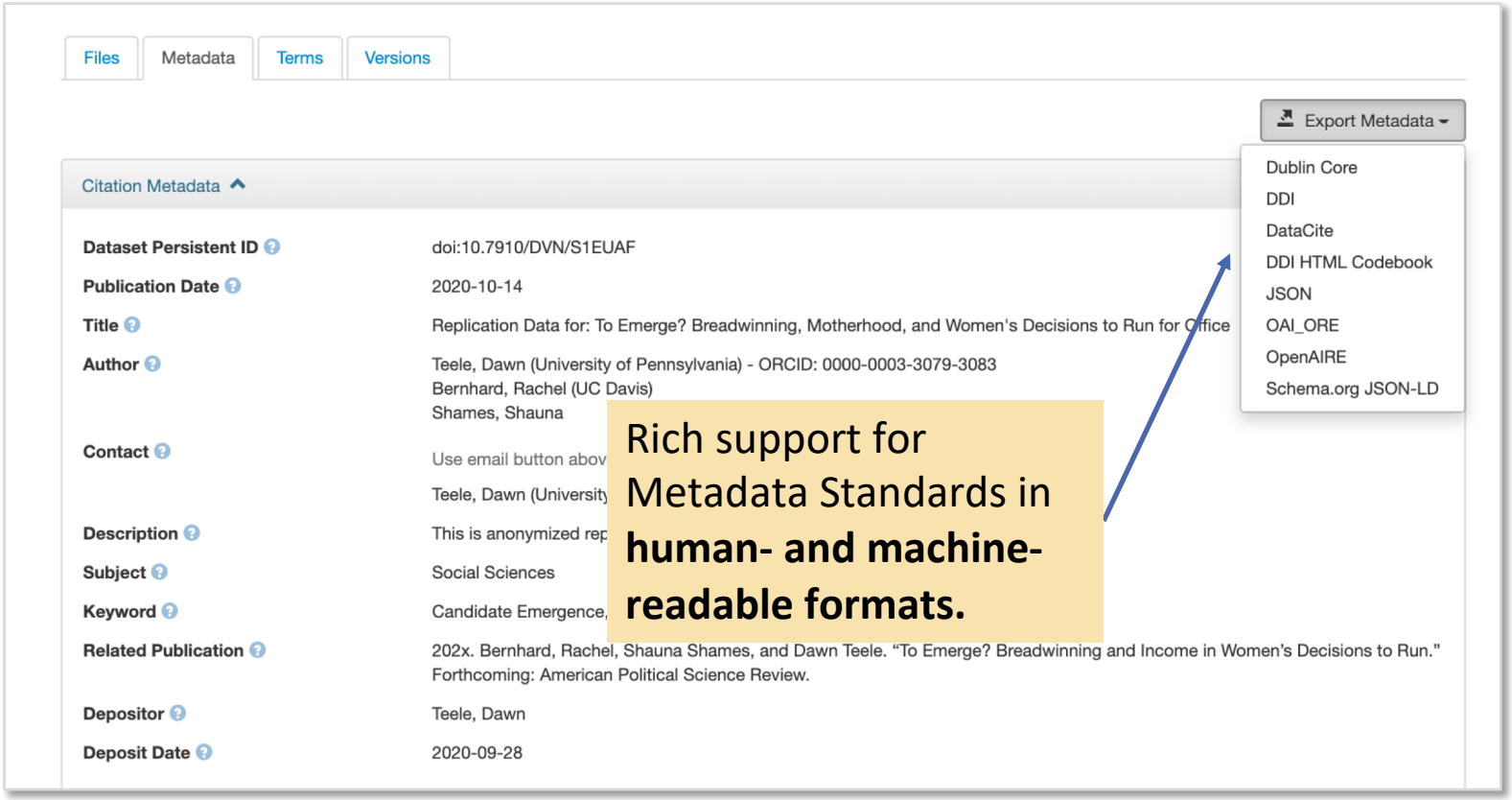
"Digital resources should be easy to find for both humans and computers. Extensive machine-actionable metadata are essential for automatic discovery of relevant datasets and services, and are therefore an essential component of the FAIRification process."

A. Jacobsen, et al. 2020. FAIR principles: Interpretations and implementation considerations. Data Intelligence 2(2020), 10-29. doi: 10.1162/dint_r_00024

Full, standard data citation automatically generated



Support for multiple metadata standards



Google

Last updated | Download format | Usage rights | Topic | Free | Saved datasets

100+ datasets found

H

Data from: The International Political Economy Data...

dataverse.harvard.edu
search.datacite.org

zip, txt +4

Updated Aug 7, 2020

H

CCES Common Content, 2018

dataverse.harvard.edu
search.datacite.org

pdf, tsv +1

Updated Nov 17, 2019

H

Data from: Mining texts to efficiently generate global dat...

dataverse.harvard.edu

txt +1

Updated Jul 8, 2015

Data from: The International Political Economy Data Resource

Related Article

Explore at Harvard Dataverse | Explore at search.datacite.org

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

zip, txt, application/gzip, text/tsv, type/x-r-syntax, docx

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

To Be Accessible (4 Principles)

"**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."

A. Jacobsen, et al. 2020.

Search

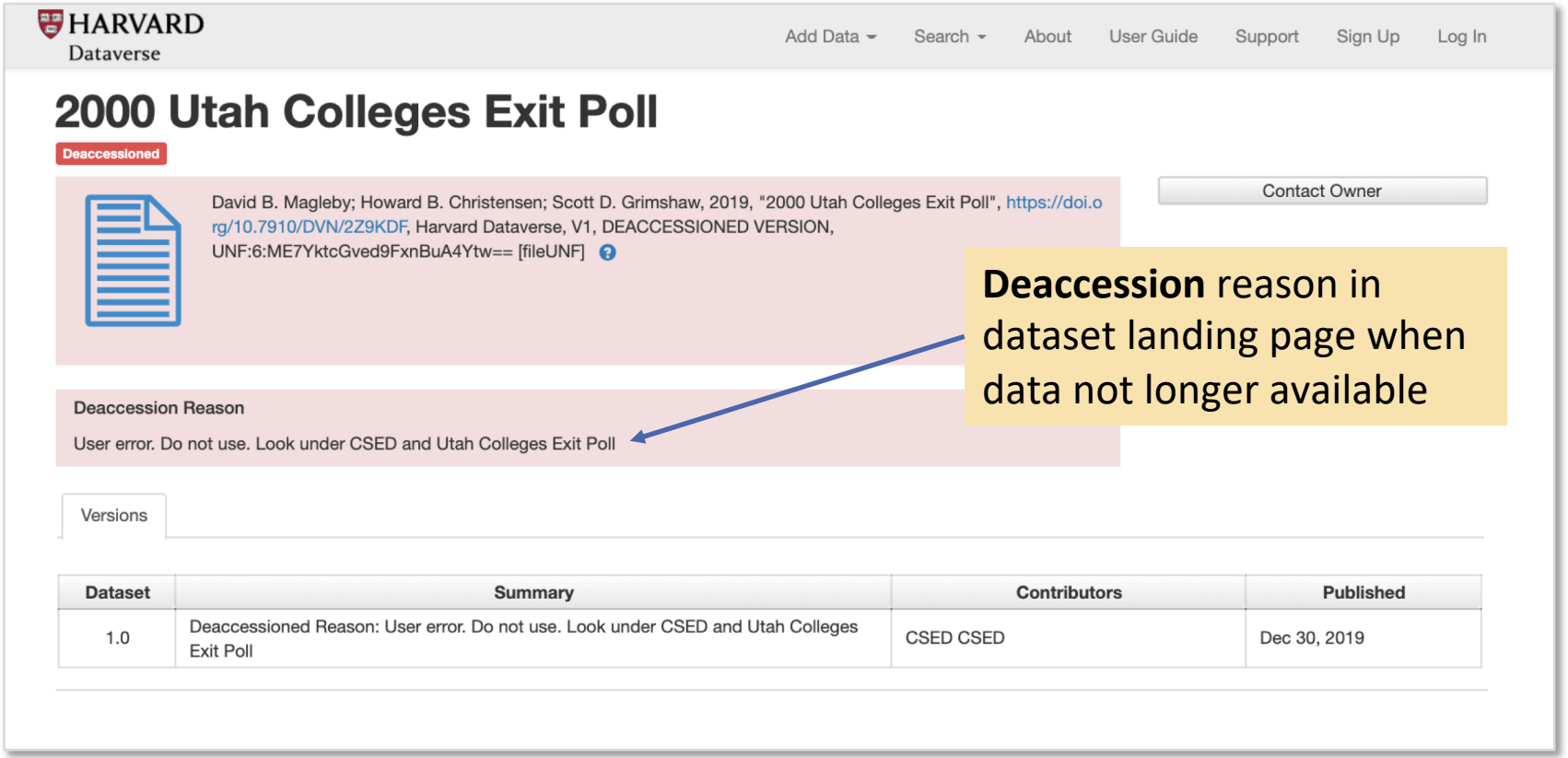
User Guide
Admin Guide
API Guide
Introduction
Getting Started with APIs
API Tokens and Authentication
Search API
Data Access API
Native API
Metrics API
SWORD API
Client Libraries
Building External Tools
Apps
Frequently Asked Questions
Installation Guide
Developer Guide
Style Guide

API Guide

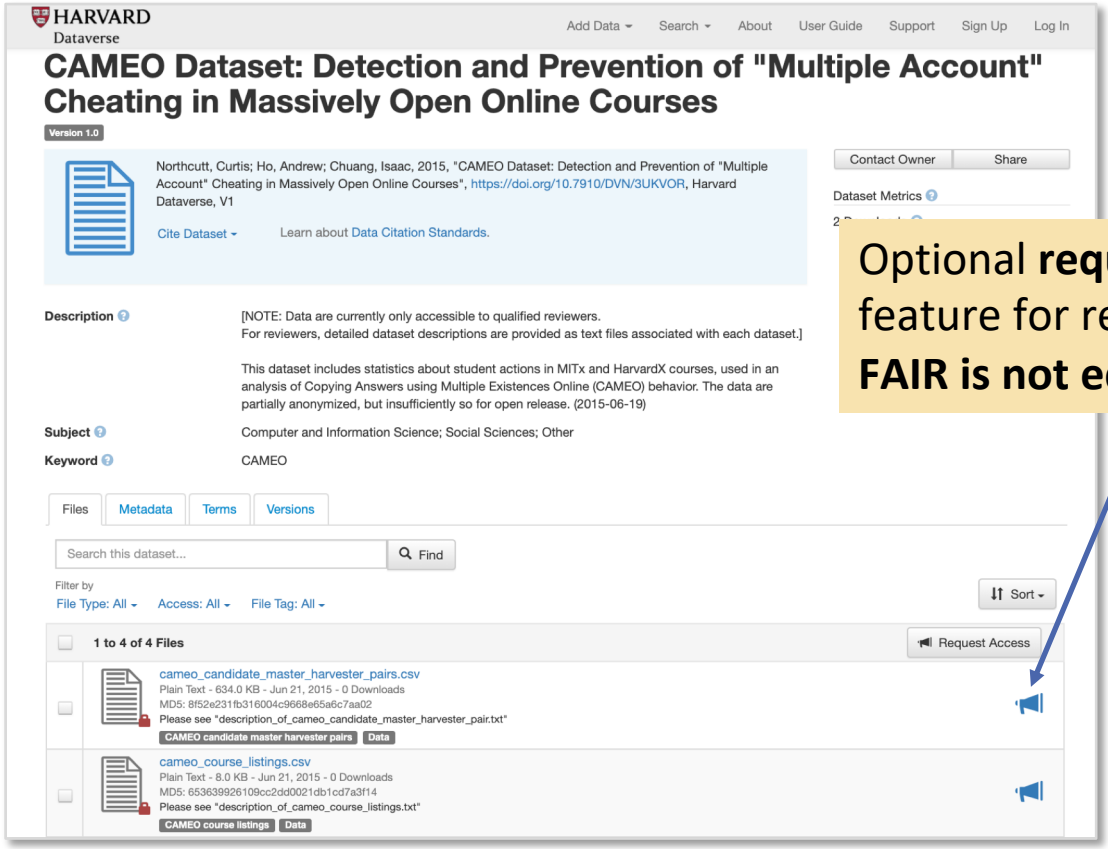
Contents: https://guides.dataverse.org

- 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

Metadata always available



Access terms available for restricted data



Optional **request access** feature for restricted data.

FAIR is not equal to Open.

To Be Interoperable (3 Principles)

"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."

A. Jacobsen, et al. 2020.

ClimateRegressionData_150327.tab

Version 1.0

File Citation

Albouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "ClimateRegressionData_150327.tab", *Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.*, https://doi.org/10.7910/DVN/QCE1XY/BNJLIA, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

Cite Data File

Learn about Data Citation Standards.

File Metrics

45 Downloads

This file is part of "Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.".

Dataset Citation

Albouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "Replication Data for: "Climate Amenities, Climate Chang
American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-24
https://doi.org/10.7910/DVN/QCE1XY, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

Cite Dataset

Learn about Data Citation Standards.

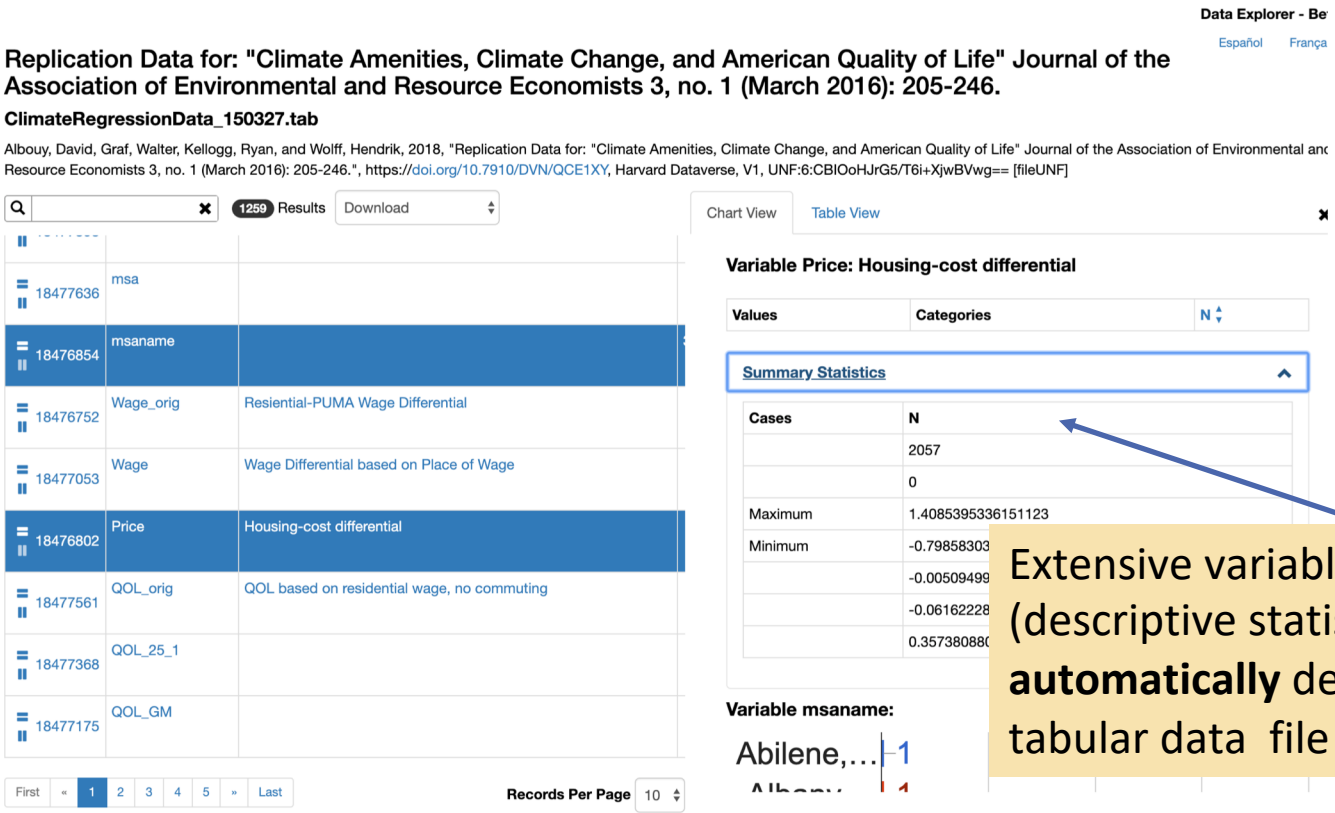
Variable metadata from tabular data file

Preview | Metadata | Versions

Open View Data

	statefip	PumaID	msa	msaname	Wage_orig	Wage	Price
1	1	100100.0	2650	Florence, AL	-0.14469655	-0.15300082	-0.36732796
2	1	100200.0	3440	Huntsville, AL	-0.06367312	-0.0687066	-0.21142627
3	1	100300.0	3440	Huntsville, AL	-0.06052007	-0.06744661	-0.3109654
4	1	100400.0	19999	Non-metro, AL	-0.16140184	-0.166009	-0.49454302
5	1	100500.0	19999	Non-metro, AL	-0.16811557	-0.15688014	-0.40440822
6	1	100600.0	2030	Decatur, AL	-0.07162431	-0.09517802	-0.34356594
7	1	100700.0	19999	Non-metro, AL	-0.21245104	-0.19640322	-0.6055518

DDI supports interoperability even at the variable level



Subject	Social Sciences
Topic Classification	mra 50 or fewer (Sampsize) http://www.murray.harvard.edu/vocabulary male (Gender) http://www.murray.harvard.edu/vocabulary 18-22 (Age) http://www.murray.harvard.edu/vocabulary White (Race) http://www.murray.harvard.edu/vocabulary student (SES) http://www.murray.harvard.edu/vocabulary 1 (Generations) http://www.murray.harvard.edu/vocabulary Special aspects of education (Education) http://authorities.loc.gov/ Mental health (Health) http://authorities.loc.gov/
Distributor	Murray Research Archive http://www.murray.harvard.edu
Distribution Date	1981
Time Period Covered	Start: 1930 ; End: 1940
Date of Collection	Start: 1930
Kind of Data	field study

Use **standard, global Controlled Vocabulary** from the Library of Congress

Australian National Political Attitudes, 1967: Supplemented with Treiman Prestige Scores (M023V1)

Version 2.0

Donald Treiman, 2012, "Australian National Political Attitudes, 1967: Supplemented with Treiman Prestige Scores (M023V1)", https://doi.org/10.7910/DVN/D1NDDL, Harvard Dataverse, V2

Cite Dataset

Learn about Data Citation Standards.

Dataset Metrics

2 Downloads

Related Material

McDonnell, Patrick, Leonard Blom, F. Lancaster Jones, and Paul Duncan-Jones, "Notes on the Australian Occupational Classification," Australian National University, Paper preared fro annual meeting of the Sociological Association of Australia and New Zealand, August 1976. Australia, Bureau of Census and Statistics, "Classification and Classified List of Occupations" (revised June 1961), Government Printer, Canberra, 1961. Australia, Bureau of census and Statistics, "Index of Occupations" (revised June 1961), Government Printer, Canberra, 1961. Broom, Leonard, F. Lancaster Jones and Jerzy Zubrzycki, "A Occupational Classification of the Australian Workforce," THE AUSTRALIAN AND NEW ZEALAND JOURNAL OF SOCIOLOGY Vol. 1, No. 2 (October, 1965), p.1-2.

Related Datasets

Aitkin, Donald, Michael Kahan, and Donald E. Stokes. AUSTRALIAN NATIONAL POLITICAL ATTITUDES, 1967. Conducted by Donald Aitkin and Michael Kahan, Australian National University, and Donald E. Stokes, University of Michigan. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 1975. doi:10.3886/ICPSR07282.v1; Aitkin, Donald, Michael Kahan, and Donald E. Stokes. Australian National Political Attitudes, 1969. ICPSR07393-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2001. doi:10.3886/ICPSR07393.v1

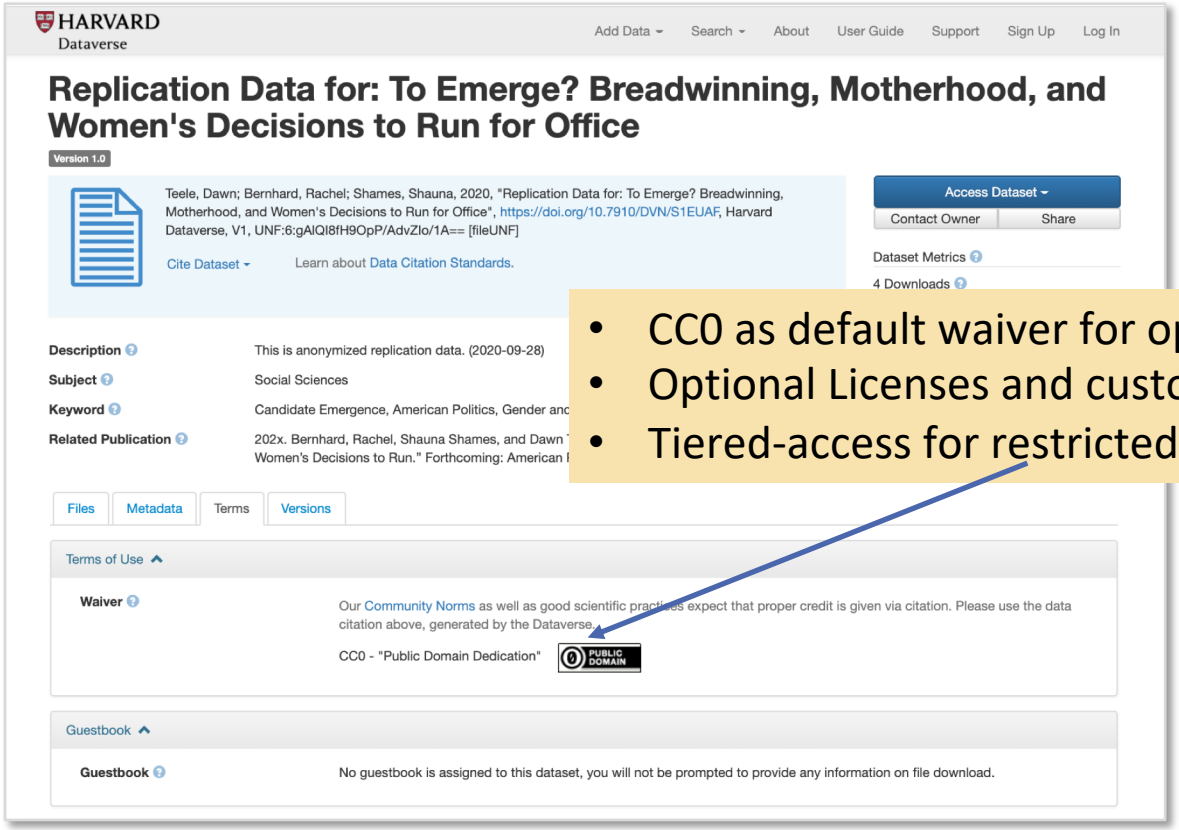
Reference in metadata to **related datasets** or other research objects

To Be Reusable (4 Principles)

"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."

A. Jacobsen, et al. 2020.

Licenses, Terms, and Tiered-Access to Data



- CC0 as default waiver for open data;
- Optional Licenses and custom Terms;
- Tiered-access for restricted data

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Sign Up | Log In

Files | Metadata | Terms | Versions

Versioning to track provenance and help with reuse

View Differences

	Dataset	Summary	Contributors	Published
	8.0	**Files (Added: 2; Removed: 2);** View Details	Sergio Petralia	Jan 19, 2019
	7.2	**Terms of Use/Access Changed** View Details	Sergio Petralia	Dec 19, 2017
	7.1	**Citation Metadata:** Author (3 Changed); Contact (1 Changed); **Additional Citation Metadata:** (5 Changed); View Details	Sergio Petralia	Oct 4, 2017
	7.0	**Files (Added: 1; Removed: 1; Replaced: 1);** View Details	Sergio Petralia	Sep 12, 2017
	6.1	**Citation Metadata:** Notes (Changed); View Details	Sergio Petralia	Jan 10, 2017
	6.0	**Files (Added: 2; Removed: 2);** View Details	Sergio Petralia	Jan 10, 2017
	5.0	**Files (Added: 1; Removed: 1);** View Details	Sergio Petralia	Oct 3, 2016
	4.0	**Citation Metadata:** Description (1 Changed); **Additional Citation Metadata:** (1 Added, 8 Changed); **Files (Added: 2; Removed: 2; Changed File Metadata: 1);** View Details	Sergio Petralia	Sep 29, 2016
	3.3	**Additional Citation Metadata:** (5 Changed); View Details	Sergio Petralia	Sep 7, 2016
	3.2	**Additional Citation Metadata:** (2 Added); View Details	Sergio Petralia	Sep 7, 2016
	3.1	**Additional Citation Metadata:** (5 Added); View Details	Sergio Petralia	Sep 7, 2016
	3.0	**Files (Added: 1; Removed: 1);** View Details	Sergio Petralia	Sep 7, 2016
	2.1	**Citation Metadata:** Notes (Changed); View Details	Sergio Petralia	Sep 1, 2016
	2.0	**Files (Added: 1);** View Details	Sergio Petralia	Sep 1, 2016
	1.1	**Additional Citation Metadata:** (2 Added); View Details	Sergio Petralia	Sep 1, 2016
	1.0	This is the first published version.	Sergio Petralia	Aug 26, 2016

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

Summary of FAIR 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**

Q & A

Thank you!

- **Data Sharing** is becoming a key part of transparent and reproducible research
- **Research institutions** should provide **data management and sharing services** to support the increase of demands on data-centric research
- **Dataverse repositories** can be part of the solution, providing support for FAIR data and giving credit to data authors

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

IQSS
The Institute for Quantitative Social Science

HARVARD UNIVERSITY