

DATAVERSE

A SOFTWARE PLATFORM

A COMMUNITY

A NETWORK OF REPOSITORIES

Mercè Crosas, Ph.D. @mercecrosas
Harvard Research Data Officer
IQSS Chief Data Science and Technology Officer

Role of Generalist Repositories to Enhance Data Discoverability and Reuse
NIH, Feb. 11-12

DATAVERSE SOFTWARE PLATFORM USED WORLDWIDE

Network of Dataverses



- 54 installations
- In 6 continents
- 6000 dataverses
- 130K datasets
- Metadata is shared across the Network of Dataverses

A GROWING AND ENGAGED COMMUNITY

“The
community
is the one
thing that
makes this
work!”

Sherry Lake,
Scholarly Repository
Librarian, University
of Virginia



- **Core development** at Harvard
- > 100 code contributors
- Regular, popular Community meetings, calls, discussions
- **Fast adoption (e.g., in Europe)** w/ multiple language support
- Governance and assistance from the **Global Dataverse Community Consortium**

Dataverse repositories come in various types and sizes.

The **Harvard Dataverse** is a generalist repository open to all, where NIH funded data may be published.

NIH funded data may also be published in **other US Dataverse repositories**, including domain-specific and institutional repositories.

EXAMPLES

HARVARD DATAVERSE: Free and Open to all

The screenshot shows the Harvard Dataverse website. At the top left is the Harvard Dataverse logo. The navigation bar includes links for 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In'. The main content area is split into two columns. The left column is titled 'Deposit and share your data. Get academic credit.' and contains the text 'Harvard Dataverse is a repository for research data. Deposit data and code here.' Below this, it shows '94,119 datasets' and '10,637,539 downloads'. At the bottom of this column is a button labeled 'Add a dataset +'. The right column is titled 'Organize datasets and gather metrics in your own repository.' and contains the text 'A dataverse is a container for all your datasets, files, and metadata.' Below this, it shows '3,620 dataverses'. At the bottom of this column is a button labeled 'Add a dataverse +'. Below the main content area, there is a search bar with the text 'Search over 94,100 datasets...' and a 'Find' button. At the bottom, there is a 'Browse by subject' section with a grid of subject categories and their corresponding dataset counts.

HARVARD
Dataverse

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

Deposit and share your data. Get academic credit.
Harvard Dataverse is a repository for research data. Deposit data and code here.

94,119 datasets 10,637,539 downloads

Add a dataset +

Organize datasets and gather metrics in your own repository.
A dataverse is a container for all your datasets, files, and metadata.

3,620 dataverses

Add a dataverse +

Find data across research fields, preview metadata, and download files

Search over 94,100 datasets... **Q Find**

Browse by subject

Agricultural Sciences 1,317	Computer and Information Science 1,053	Medicine, Health and Life Sciences 3,274
Arts and Humanities 871	Earth and Environmental Sciences 2,020	Physics 897
Astronomy and Astrophysics 738	Engineering 464	Social Sciences 39,719
Business and Management 489	Law 296	
Chemistry 235	Mathematical Sciences 226	

- **90K** discoverable datasets
- **> 3000** in biomedicine
- **> 80** journal dataverses
- Datasets up to TB
- Files up to 10 GB
- Offers new fee-based **data curation services**

SBGRID DATA: A Repository for Structural Biology primary data

The screenshot displays the SBGrid Data Bank interface. At the top, there are navigation links for 'SBGrid Databank' and 'Publication Guidelines'. The main header includes the SBGrid logo, 'Data Bank', and a 'SBDB/DV testing' badge. A search bar is present with a 'Find' button and a link to 'Advanced Search'. On the left, there are filters for 'Dataverses (78)', 'Datasets (432)', and 'Files (432)'. Below these are sections for 'Dataverse Category' (Laboratory (78)), 'Publication Year' (2015 (210), 2016 (153), 2017 (100), 2018 (47)), and 'Data Type' (X-Ray Diffraction (409), Micro-Electron Diffraction (13), Structural Model (7), XFEL Diffraction (2), Lattice Light-Sheet Microscopy (1)). At the bottom left, there is a 'Beamline and Collection Facility' section for 'APS 24-ID-C (16)'. The main content area shows '1 to 10 of 510 Results' with a 'Sort' button. Three results are visible, each with a thumbnail, title, date, author, and a link to the dataset. The first result is for 'X-Ray Diffraction data from Npl4 zinc finger and MPN domains (Chaetomium thermophilum), source of 6CDD structure' by Bodnar, Nicholas; Rapoport, Tom, 2018. The second is for 'X-Ray Diffraction data from complex of engineered human chemokine CX3CL1 with viral US28 and two nanobodies, source of 5WB2 structure' by Jude, Kevin M; Tsutsumi, Naotaka; Garcia, K. Christopher, 2018. The third is for 'X-Ray Diffraction data from single chain complex of viral US28 with stabilizing nanobody nb7, source of 5WB1 structure' by Jude, Kevin M; Burg, John S; Garcia, K. Christopher, 2018.


- Helmsley grant to PIs Sliz and Crosas to expand Dataverse for SBGrid Data and other biomedical repositories
- Added support for **large datasets, in-place computation, ORCID auth**
- Leverages Dataverse community for longer term repository sustainability

Beta: not in production yet

QUALITATIVE DATA REPOSITORY:

Dedicated to qualitative and multi-method research


Search ▾ About ▾ Deposit ▾ Discover ▾ Guidance and Resources ▾ Publications ▾ Initiatives ▾ Log In Register



Data for: Carbon captured: How business and labor control climate politics

Version 1.0

Mildenberger, Matto. 2020. "Data for: Carbon captured: How business and labor control climate politics". Qualitative Data Repository. <https://doi.org/10.5064/F6GYLSON>. QDR Main Collection. V1


 [Learn about Data Citation Standards.](#)

Description

This data project has been published in parallel to Mildenberger, M. 2020. *Carbon Captured: How Business and Labor Control Climate Politics* (Cambridge, MA: MIT Press). In this book, I advance a new theory to explain cross-national differences in the timing and content of climate reforms across advanced industrial economies.

This book seeks to make the sources of its inferences as transparent as possible. In doing so it draws from best practices outlined by the American Political Science Association in its 2012 Council Statement on Openness in Political Science and in a 2013 Guideline for Data Access and Research Transparency for Qualitative Research in Political Science. This data project and its associated Transparency Appendix (TRAX) are key components of this effort. Where permitted by copyright, the data project provides digital copies of grey literatures, policy documents, and media reports referenced in the book or in the book's

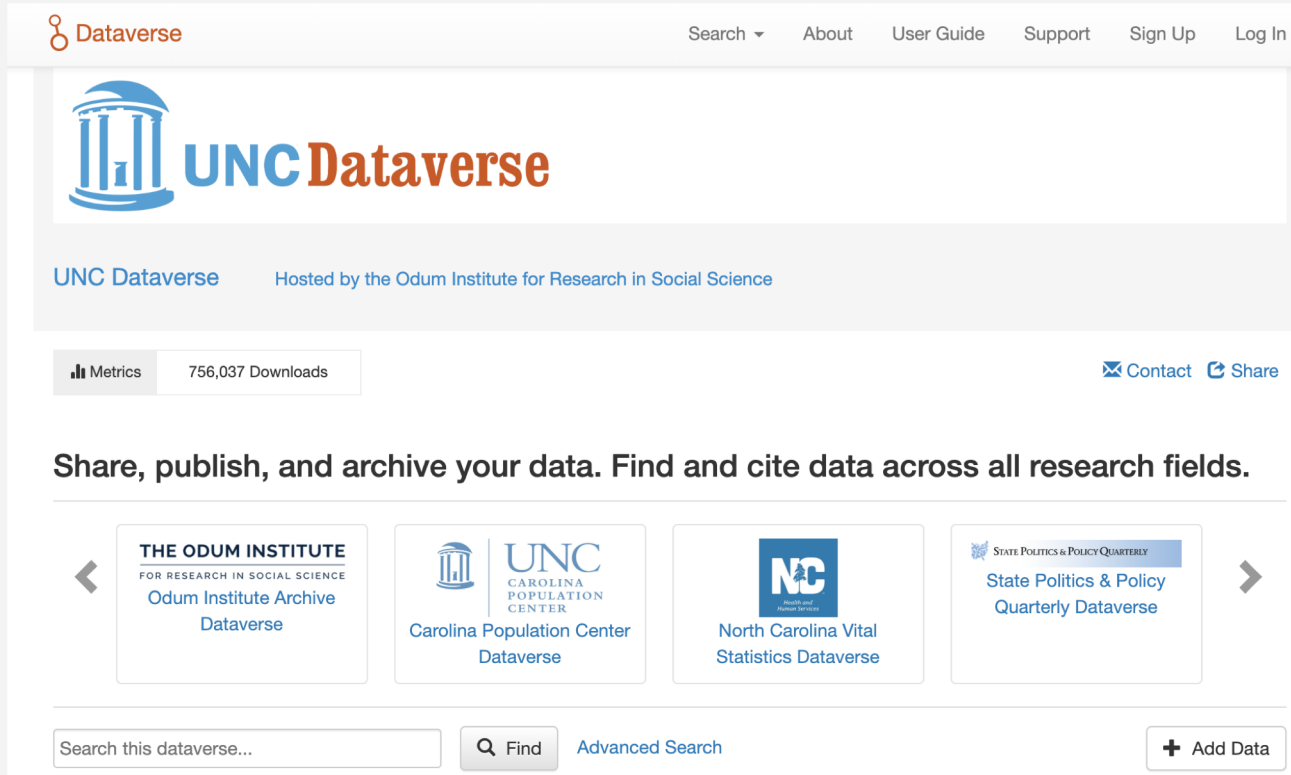
Data Project Metrics

1 Download 



UNC DATAVERSE:

Supports repositories for others through partnerships

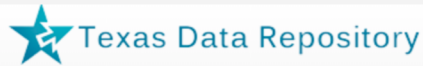


The screenshot shows the UNC Dataverse website interface. At the top, there is a navigation bar with the Dataverse logo and links for Search, About, User Guide, Support, Sign Up, and Log In. Below the navigation bar is a large banner featuring the UNC Dataverse logo and the text "UNC Dataverse Hosted by the Odum Institute for Research in Social Science". Underneath the banner, there is a metrics section showing "756,037 Downloads" and links for "Contact" and "Share". A central message reads "Share, publish, and archive your data. Find and cite data across all research fields." Below this message is a carousel of four partner logos: The Odum Institute for Research in Social Science Odum Institute Archive Dataverse, UNC Carolina Population Center Dataverse, North Carolina Vital Statistics Dataverse, and State Politics & Policy Quarterly State Politics & Policy Quarterly Dataverse. At the bottom, there is a search bar with the placeholder text "Search this dataverse...", a "Find" button, a link to "Advanced Search", and an "Add Data" button.



In process

TEXAS DATA REPOSITORY: A Consortium of 11 Universities



Search ▾ About User Guide Support Log In

Metrics 29,314 Downloads

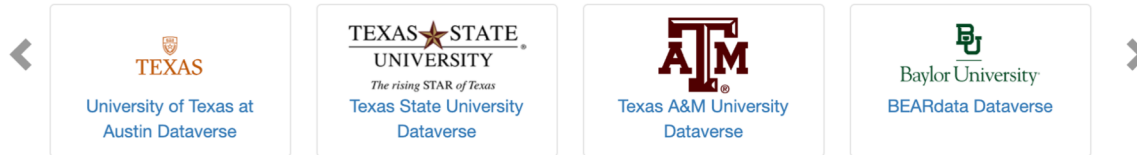
Contact Share

Share, publish, and manage your data. Find and cite data across all research fields.

Welcome to the Texas Data Repository Dataverse, a research data management system for Texas Digital Library (TDL) member institutions. To add, share, and publish your data or work on a project, select your local institutional repository from the institutions below. To find datasets from across Texas institutional dataverses, start here.

LEARN MORE

- [Go to the user guide.](#)
- [Contact a local university librarian for help.](#)



Search this dataverse...

Find

[Advanced Search](#)



UVA DATAVERSE and JOHNS HOPKINS DATAVERSE: Institutional repositories serving their community

UVA DATAVERSE

Search ▾ About User Guide Support Log In

Add your dataset to LibraData

LibraData is a place for UVA researchers to share data publicly, and is part of the Libra Scholarly Repository suite of services which includes works of UVA scholarship such as articles, books, theses, and data.

Step One: Start with the [LibraData Checklist](#)

Step Two: Select the dataverse in which you want to add a dataset. You can create a dataset at the main level of LibraData or in a sub-dataverse.

Select a dataverse... ▾

Step Three: [Add Data](#)

Search Datasets

Search below for datasets in LibraData and all of UVA Dataverse

Search ALL datasets... [Search](#)

More Information:

- [FAQs about LibraData](#)
- [Policies & Community Sharing Norms](#)
- [Need help? Send us an email!](#) ✉

JOHNS HOPKINS LIBRARIES

Johns Hopkins Data Services

Dataverse

Search ▾ User Guide Support Log In

Johns Hopkins University Data Archive (JHU Data Archive)

[Metrics](#) 5,117 Downloads [Contact](#) [Share](#)

A repository where Johns Hopkins researchers can share, archive, and get citations for their data.

Search this dataverse... [Find](#) [Advanced Search](#)

[Dataverses \(35\)](#)

[Datasets \(79\)](#)

[Files \(707\)](#)

Dataverse Category

- Research Project (16)
- Research Group (5)
- Laboratory (1)

Publication Year

2019 (28)

1 to 10 of 114 Results [Sort ▾](#)

Data associated with the 2020 PLoS One publication entitled "Actor feedback and rigorous monitoring: Essential quality assurance tools for testing behavioral interventions with simulation"

Jan 13, 2020

[Turnbull, Alison E.; Li, Xintong; Basyal, Pragyasree S.; Taply, Melissa L.; Singh, Arun L., 2020, "Data associated with the 2020 PLoS One publication entitled "Actor feedback and rigorous monitoring: Essential quality assurance tools for testing behavioral interventions with simulation", <https://doi.org/10.7281/1/6BHBC4>, Johns Hopkins University Data Archive, V1](#)

Evaluation of a quality assurance program designed to ensure that 3 actors performing in simulated medical encounters, provided standardized performances.

FAIR AND CORE TRUST SEAL

Dataverse substantially facilitates supporting FAIR data principles and Core Trust Seal certification.

Best archival practices, dedicated support, and data reviews and curation contribute to **improve data quality and trustworthiness.**

FM [AID*]	Question	Dataverse Q'aire	Dataverse Optimized
Identifier type	1	DOI	DOI
F1A	2		
F1B	Not tested in Q'aire		
F2A	4A		
F2A	4B		
F3	5B		
F4	6A		
F4	6B		
A1.1	7A		
A1.2	8A		
A1.2	8B	N/A	N/A
A2	9		
I1	10		
I2	11		
I3	12		
R1.1	13		
R1.2	14A		

DATAVERSE FAIR SUMMARY

FAIR Test:

www.biorxiv.org/content/10.1101/418376v2.full

- Strong support for Findable, Accessible, and Reusable principles
- (F3 was fixed after test)
- Interoperable principles more difficult to follow; often requires input from data authors or curators
- Emphasis in machine-readability
- FAIR is not a “standard”, it’s a path


FINDABILITY:

Full, Standard Data Citation Automatically Generated

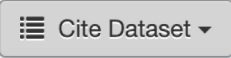
The screenshot shows the Harvard Dataverse interface. At the top left is the Harvard logo and 'Dataverse' text. The top navigation bar includes 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In'. The main content area features a document icon, the dataset title 'REPLICATION DATA FOR: Bootstrap Methods for Inference in the Parks Model', and 'Version 3.0'. A light blue box highlights the citation text: 'Moundigbaye, Mantobaye; Messemer, Clarisse; Parks, Richard W.; Reed, W. Robert, 2017, "REPLICATION DATA FOR: Bootstrap Methods for Inference in the Parks Model", <https://doi.org/10.7910/DVN/94EU5T>, Harvard Dataverse, V3'. Below this is a 'Cite Dataset' button with a dropdown menu showing 'EndNote XML', 'RIS', and 'BibTeX'. To the right, a 'Dataset Metrics' box shows '126 Downloads'. A light orange box highlights the text: 'Data Citation, with DataCite DOI, fully compliant with Joint Declaration of Data Citation Principles'. Below this, the dataset description and subject/keyword information are visible.

HARVARD
Dataverse

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

 **REPLICATION DATA FOR: Bootstrap Methods for Inference in the Parks Model**
Version 3.0

Moundigbaye, Mantobaye; Messemer, Clarisse; Parks, Richard W.; Reed, W. Robert, 2017, "REPLICATION DATA FOR: Bootstrap Methods for Inference in the Parks Model", <https://doi.org/10.7910/DVN/94EU5T>, Harvard Dataverse, V3

 Cite Dataset ▾

- EndNote XML
- RIS
- BibTeX

Dataset Metrics ?

126 Downloads ?

Learn about [Data Citation Standards](#).

Data Citation, with DataCite DOI, fully compliant with Joint Declaration of Data Citation Principles

This dataset contains all the data used to estimate the "REPLICATION DATA FOR: Bootstrap Methods for Inference in the Parks Model". Please cite this dataset when you use the software, and the files consist of:

Business and Management; Social Sciences

Subject ?

Keyword ?

Parks model, PCSE estimator, SUR (Seemingly Unrelated Regression), Panel data, Bootstrap, Cross-sectional dependence

FINDABILITY AND REUSABILITY: Support Many Metadata Standards

The screenshot displays the Harvard Dataverse website interface. At the top, the Harvard logo and 'Dataverse' name are on the left, and navigation links for 'Add Data', 'Search', 'About', 'User Guide', 'Support', 'Sign Up', and 'Log In' are on the right. Below the navigation is a tabbed interface with 'Files', 'Metadata', 'Terms', and 'Versions' tabs. The 'Metadata' tab is active, showing 'Citation Metadata' for a dataset. The dataset information includes: Dataset Persistent ID (doi:10.7910/DVN/94EU5T), Publication Date (2017-10-26), Title (REPLICATION DATA FOR: Bootstrap Methods for Inference in the Parks Model), Author (M... Me... Pa... Re... 459-8174), Contact (Us... Re...), Description (This dataset contains all the materials needed to reproduce the results in "Bootstrap Methods for Inference in the Parks Model". Please read the README document first. The results were obtained using SAS/IML software, and the files consist of SAS data sets and SAS programs. (2019-06-06)), and Subject (Business and Management; Social Sciences). An orange callout box with the text 'Rich support for Metadata Standards in human- and machine-readable formats.' is overlaid on the author information. An orange arrow points from this box to an 'Export Metadata' dropdown menu that is open, showing a list of supported standards: Dublin Core, DDI, DataCite, DDI HTML Codebook, JSON, OAI_ORE, OpenAIRE, and Schema.org JSON-LD.

**Rich support for
Metadata Standards in
human- and machine-
readable formats.**

Export Metadata ▾

- Dublin Core
- DDI
- DataCite
- DDI HTML Codebook
- JSON
- OAI_ORE
- OpenAIRE
- Schema.org JSON-LD


ACCESSIBILITY:

Metadata Always Available

HARVARD
Dataverse

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

Harvard Dataverse > 2000 Utah Colleges Exit Poll

 **2000 Utah Colleges Exit Poll**
Deaccessioned

Deaccession reason in dataset landing page when data not longer available

[✉ Contact](#)

David B. Magleby; Howard B. Christensen; Scott D. Grimshaw, 2019, "2000 Utah Colleges Exit Poll", <https://doi.org/10.7910/DVN/2Z9KDF>, Harvard Dataverse, V1, DEACCESSIONED VERSION, UNF:6:ME7YktcGved9FxnBuA4Ytw== [fileUNF] ?

Deaccession Reason
User error. Do not use. Look under CSED and Utah Colleges Exit Poll

Versions

Dataset	Summary	Contributors	Published
1.0	Deaccessioned Reason: User error. Do not use. Look under CSED and Utah Colleges Exit Poll	CSED CSED	Dec 30, 2019

INTEROPERABLE:

Rich Metadata Derived from each Variable in data file

Data Explorer - Be
Español Franca

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.
ClimateRegressionData_150327.tab

Abouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "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>, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

1259 Results Download

ID	Variable Name	Description
18477636	msa	
18476854	msaname	
18476752	Wage_orig	Residential-PUMA Wage Differential
18477053	Wage	Wage Differential based on Place of Wage
18476802	Price	Housing-cost differential
18477561	QOL_orig	QOL based on residential wage, no commuting
18477368	QOL_25_1	
18477175	QOL_GM	

Chart View Table View

Variable Price: Housing-cost differential

Values Categories N

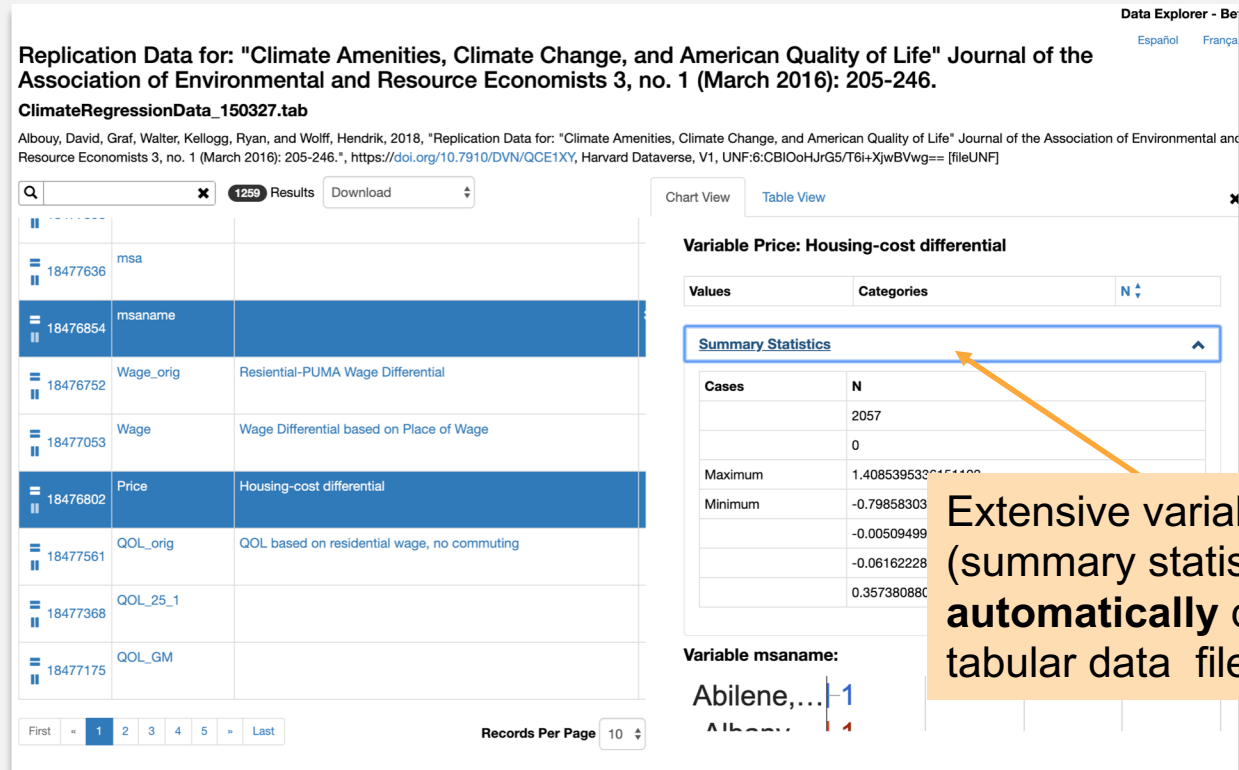
Summary Statistics

Cases	N
	2057
	0
Maximum	1.4085395320151100
Minimum	-0.79858303
	-0.00509499
	-0.06162228
	0.357380880

Variable msaname:
Abilene,... | 1
Albany... | 1

Records Per Page 10

First 1 2 3 4 5 Last



Extensive variable metadata
(summary statistics)
automatically derived from
tabular data file

INTEROPERABLE: Facilitating Use of Standard Vocabularies

The screenshot shows the Harvard Dataverse metadata form. The top navigation bar includes the Harvard Dataverse logo, 'Add Data', 'Search', 'About', 'User Guide', 'Support', and a user profile for 'Merce Crosas'. The form is divided into several sections:

- Text:** A large text input field with a '+' button to its right.
- Date:** A date input field with the placeholder 'YYYY-MM-DD'.
- Subject:** A list of subject categories with checkboxes: Agricultural Sciences, Arts and Humanities, Astronomy and Astrophysics, Business and Management, and Chemistry.
- Keyword:** A section with two input fields: 'Term' and 'Vocabulary', and a '+' button to the right.
- Vocabulary URL:** An input field with the placeholder 'Enter full URL, starting with http/'.
- Topic Classification:** A section with two input fields: 'Term' and 'Vocabulary', and a '+' button to the right.
- Vocabulary URL:** An input field with the placeholder 'Enter full URL, starting with http/'.

Two orange arrows originate from an orange callout box on the right and point to the 'Vocabulary' input fields in the 'Keyword' and 'Topic Classification' sections.

FAIR Controlled Vocabularies can be added in **metadata template**

REUSABLE: Licenses, Terms, and Tier-access to Data

The screenshot shows the Harvard Dataverse interface for a dataset. At the top, the Harvard Dataverse logo is on the left, and navigation links (Add Data, Search, About, User Guide, Support, Sign Up, Log In) are on the right. The dataset title is "Replication Data for: Bean Counters: The Effect of Soy Tariffs on Change in Republican Vote Share between the 2016 and 2018 Elections". Below the title, there is a "Version 1.0" badge. A citation box contains the following text: "Chyzh, Olga; R. Urbatsch, 2020, "Replication Data for: Bean Counters: The Effect of Soy Tariffs on Change in Republican Vote Share between the 2016 and 2018 Elections", <https://doi.org/10.7910/DVN/CV7GYN>, Harvard Dataverse, V1, UNF:6:ZhfyGSrIUw/89MqilqKAmQ== [fileUNF]". A "Cite Dataset" button is present. To the right, a "Dataset Metrics" box shows "0 Downloads". Below the citation, there is a "Description" section with the text: "How do trade wars affect voting for the Presidential electoral support base, provide a unique opportunity to study support. If trade-related considerations were a factor in the Trump-initiated tariffs immediately preceded the 2016 and 2018 congressional election, the data shows a robust inverse relationship between the 2016 and 2018 congressional election". There is also a "Subject" section with "Social Sciences" and a "Keyword" section with "trade wars, tariffs, China, vote share, agriculture". Below these sections are tabs for "Files", "Metadata", "Terms", and "Versions". The "Terms of Use" section is expanded, showing a "Waiver" section with the text: "Our [Community Norms](#) as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse." Below this text is the CC0 - "Public Domain Dedication" license logo, which is a black rectangle with a white circle containing a zero and the words "PUBLIC DOMAIN" in white capital letters. An orange callout box with a list of three items is overlaid on the right side of the page, with an orange arrow pointing from the callout box to the CC0 logo. The callout box contains the following text: "• CC0 as default waiver for open data; • Other Licenses and custom Terms; • Tier-access for restricted data".

Replication Data for: Bean Counters: The Effect of Soy Tariffs on Change in Republican Vote Share between the 2016 and 2018 Elections

Version 1.0

Chyzh, Olga; R. Urbatsch, 2020, "Replication Data for: Bean Counters: The Effect of Soy Tariffs on Change in Republican Vote Share between the 2016 and 2018 Elections", <https://doi.org/10.7910/DVN/CV7GYN>, Harvard Dataverse, V1, UNF:6:ZhfyGSrIUw/89MqilqKAmQ== [fileUNF]

Cite Dataset

Learn about [Data Citation Standards](#).

Description

How do trade wars affect voting for the Presidential electoral support base, provide a unique opportunity to study support. If trade-related considerations were a factor in the Trump-initiated tariffs immediately preceded the 2016 and 2018 congressional election, the data shows a robust inverse relationship between the 2016 and 2018 congressional election

Subject

Social Sciences

Keyword

trade wars, tariffs, China, vote share, agriculture

Files Metadata Terms Versions

Terms of Use

Waiver

Our [Community Norms](#) as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

CC0 - "Public Domain Dedication" 

- CC0 as default waiver for open data;
- Other Licenses and custom Terms;
- Tier-access for restricted data

BEYOND FAIR, CURRENT DATAVERSE EFFORTS

- Responsible FAIR for **sensitive data**:
DataTags, Differential Privacy tools,
Remote Trusted Storages (**larger data**)
- Data curation for **data quality**:
New data curation tools and services
- Capsules for **reproducibility**:
Integration with computational
platforms, capsules and research objects



HARVARD
Dataverse



Global Dataverse
Community Consortium

Thanks!