FAIR and Responsible Data Sharing with Dataverse

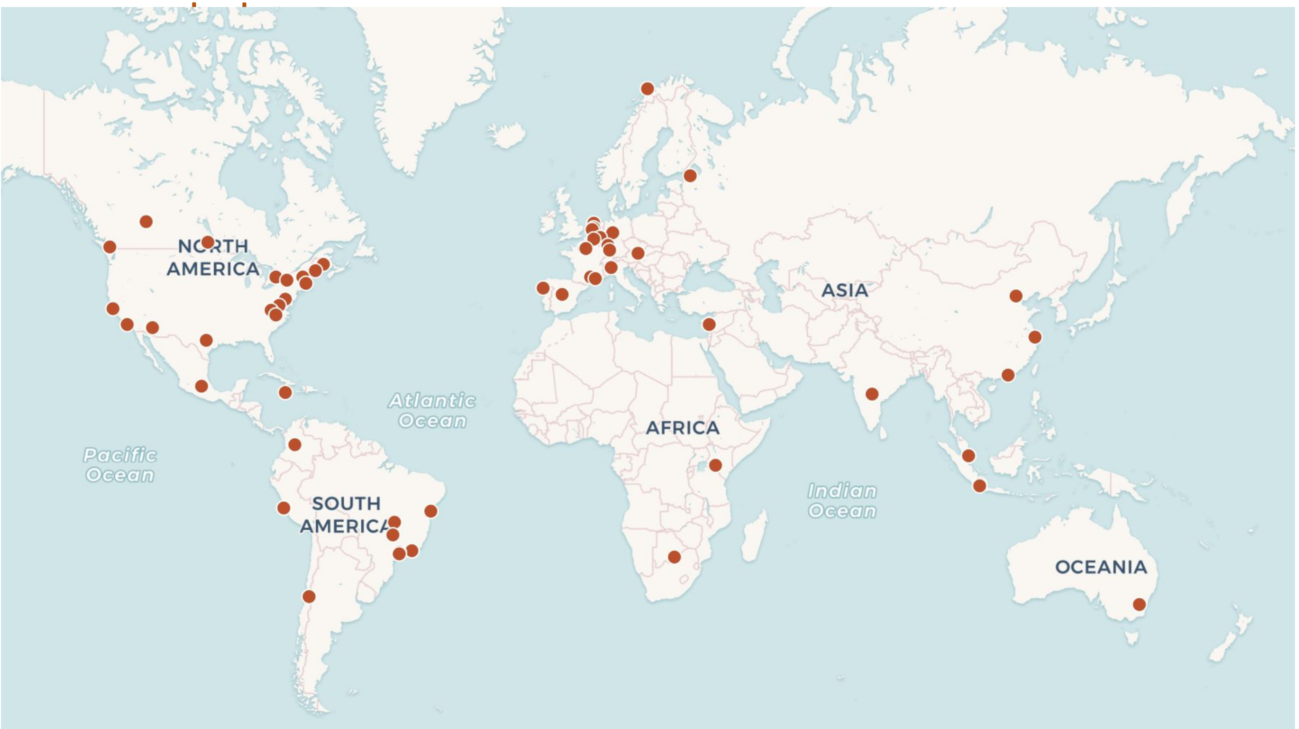
Beilstein Open Science Symposium

October 27, 2020

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

Dataverse Software Platform Used Worldwide

A Network of Dataverse repositories openly sharing



dataverse.org

- Open-source
- 63 installations
- In 6 continents
- 7,350 dataverses
- 135K datasets
- Metadata shared across the Network of Dataverses

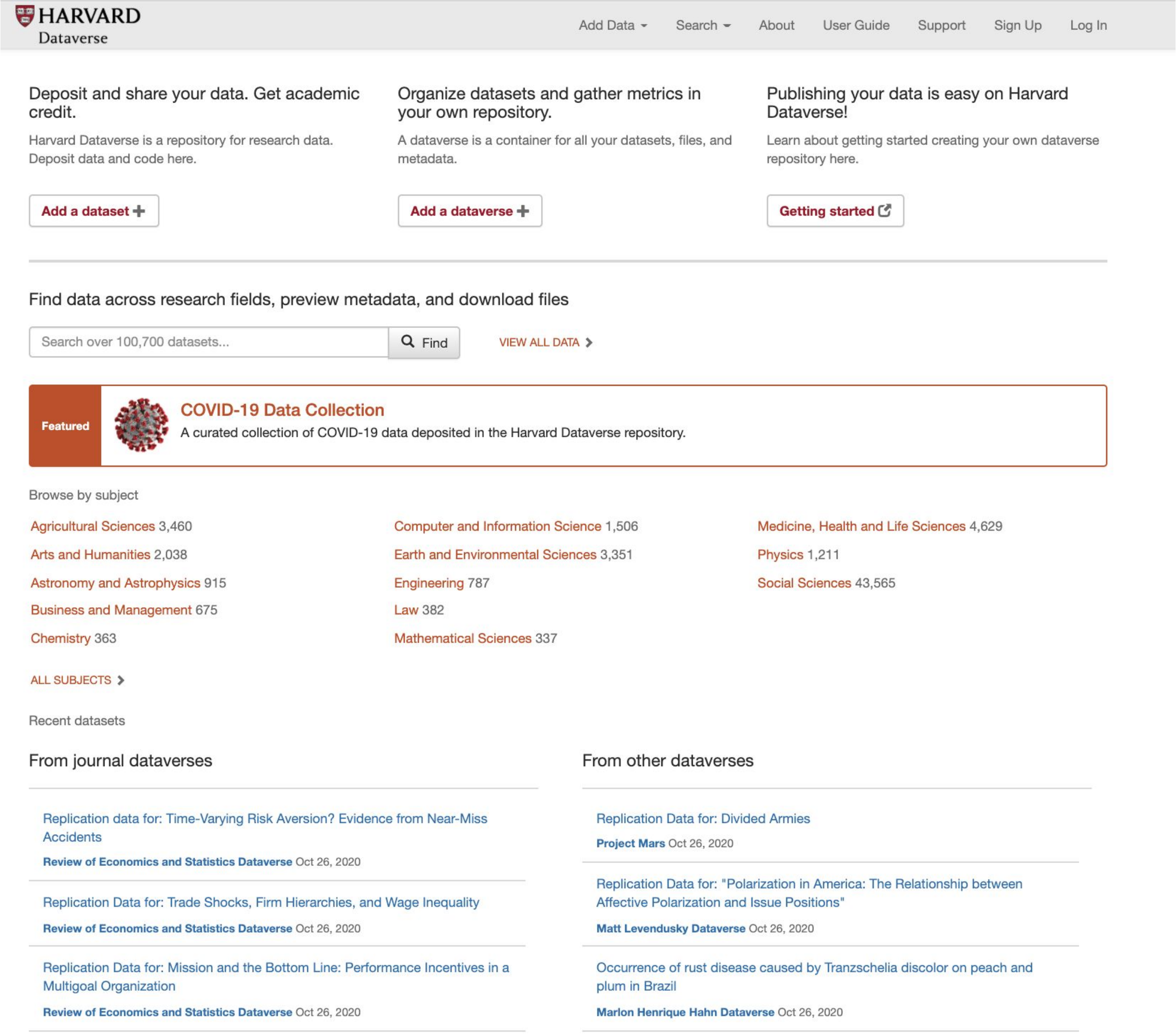
A Growing And Engaged Community

"The **community** is the one thing that makes this work!"

Sherry Lake,
Scholarly Repository Librarian, University of Virginia

- Core development at Harvard
- 117 GitHub contributors
- Regular, popular Community meetings, calls, working groups
- Fast adoption (e.g., in Europe) w/ multiple language support
- Governance and assistance from the Global Dataverse Community Consortium

Dataverse Repositories Examples



Deposit and share your data. Get academic credit.

Harvard Dataverse is a repository for research data. Deposit data and code here.

Organize datasets and gather metrics in your own repository.

A dataverse is a container for all your datasets, files, and metadata.

Publishing your data is easy on Harvard Dataverse!

Learn about getting started creating your own dataverse repository here.

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

COVID-19 Data Collection

A curated collection of COVID-19 data deposited in the Harvard Dataverse repository.

Browse by subject

Agricultural Sciences 3,460	Computer and Information Science 1,506	Medicine, Health and Life Sciences 4,629
Arts and Humanities 2,038	Earth and Environmental Sciences 3,351	Physics 1,211
Astronomy and Astrophysics 915	Engineering 787	Social Sciences 43,565
Business and Management 675	Law 382	
Chemistry 363	Mathematical Sciences 337	

Recent datasets

From journal dataverses

Replication data for: Time-Varying Risk Aversion? Evidence from Near-Miss Accidents
Review of Economics and Statistics Dataverse Oct 26, 2020

Replication Data for: Trade Shocks, Firm Hierarchies, and Wage Inequality
Review of Economics and Statistics Dataverse Oct 26, 2020

Replication Data for: Mission and the Bottom Line: Performance Incentives in a Multigoal Organization
Review of Economics and Statistics Dataverse Oct 26, 2020

From other dataverses

Replication Data for: Divided Armies
Project Mars Oct 26, 2020

Replication Data for: "Polarization in America: The Relationship between Affective Polarization and Issue Positions"
Matt Levendusky Dataverse Oct 26, 2020

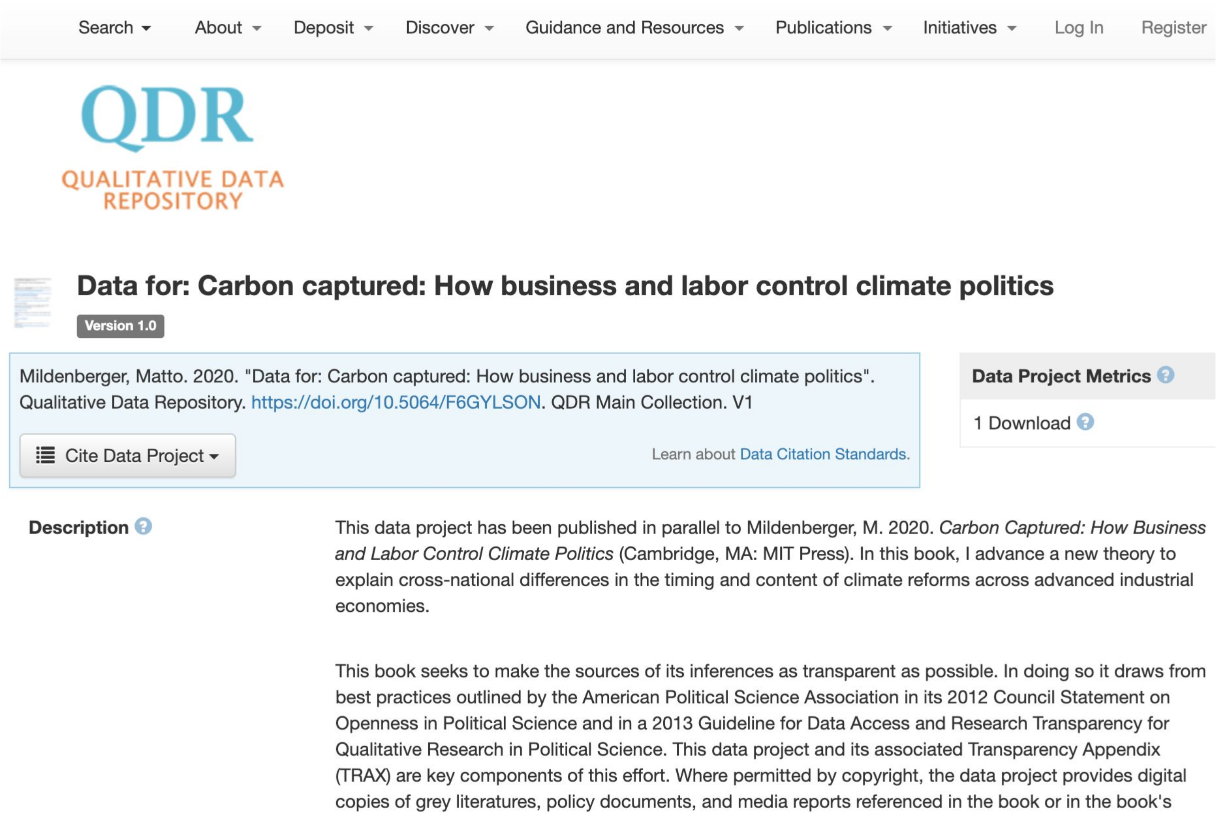
Occurrence of rust disease caused by Tranzschelia discolor on peach and plum in Brazil
Marlon Henrique Hahn Dataverse Oct 26, 2020

HARVARD DATAVERSE:

- Open to all researchers and research fields
- 100K searchable datasets
- 800K files
- 20M file downloads
- 92 journal dataverses
- New fee-based data curation services

dataverse.harvard.edu

QUALITATIVE DATA REPOSITORY:
Dedicated to qualitative and multi-method research



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

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.

Data Project Metrics

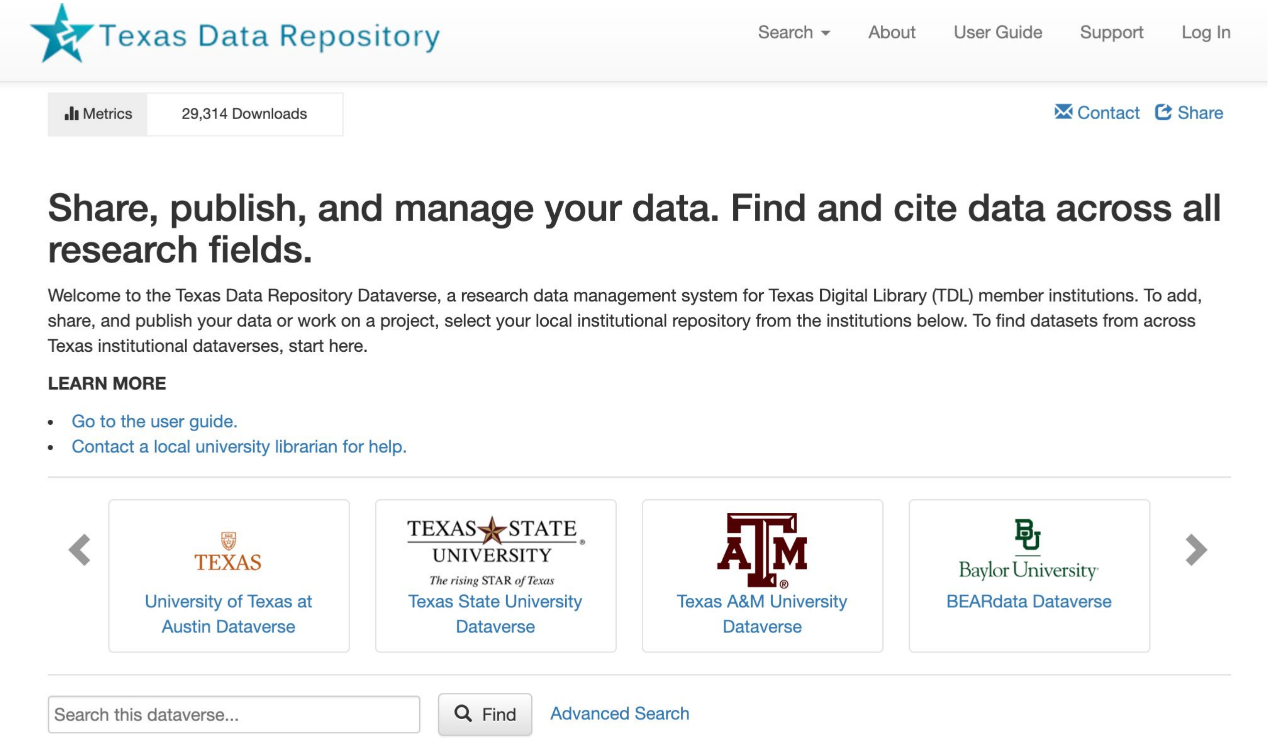
Description

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

CORE TRUST SEAL

TEXAS DATA REPOSITORY:
A Consortium of 11 Universities



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.

University of Texas at Austin Dataverse

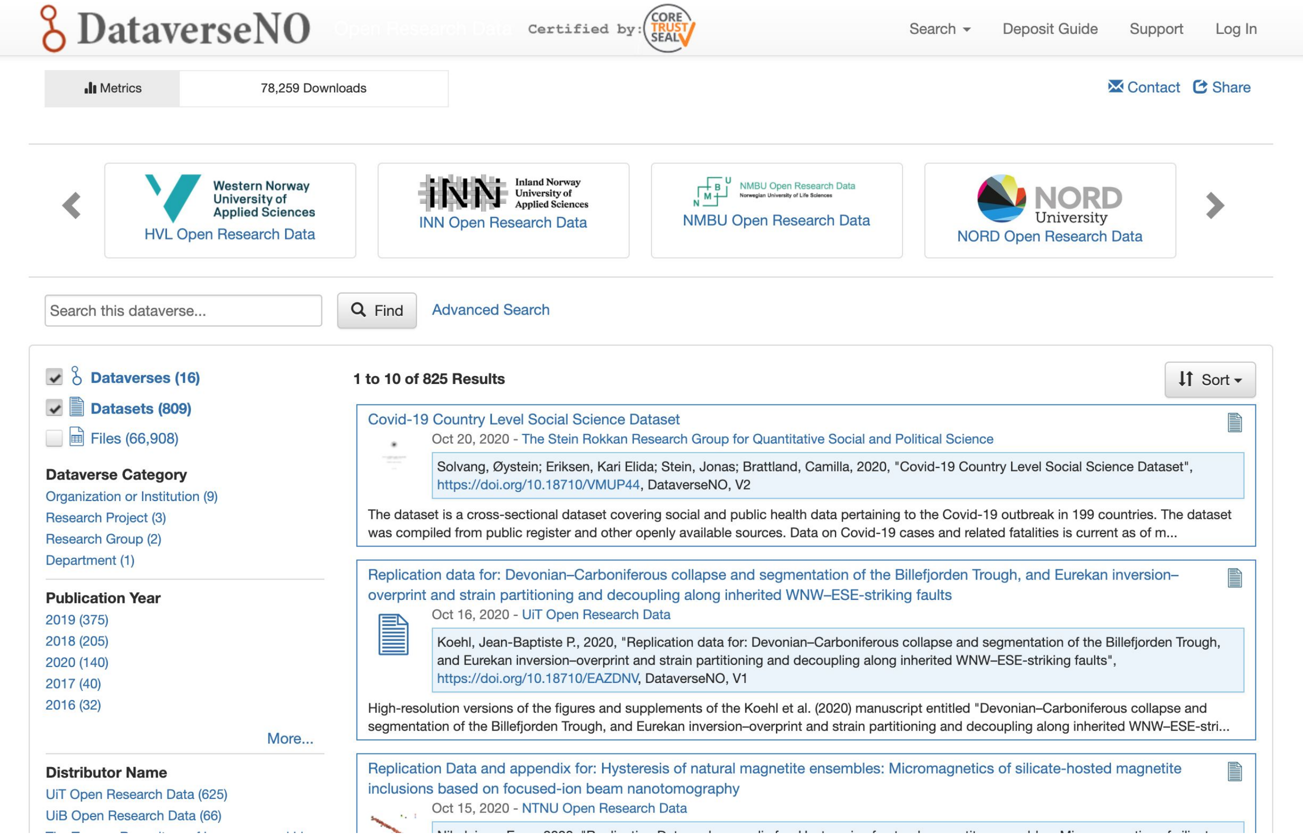
Texas State University Dataverse

Texas A&M University Dataverse

BEARdata Dataverse

CORE TRUST SEAL

DATAVERSENO:
9 Universities in Norway



78,259 Downloads

Western Norway University of Applied Sciences
HVL Open Research Data

INN Inland Norway University of Applied Sciences
INN Open Research Data

NMBU Open Research Data
Norwegian University of Life Sciences
NMBU Open Research Data

NORD University
NORD Open Research Data

Search this dataverse...

☑ Dataverses (16)
☑ Datasets (809)
□ Files (66,908)

Dataverse Category
Organization or Institution (9)
Research Project (3)
Research Group (2)
Department (1)

Publication Year
2019 (375)
2018 (205)
2020 (140)
2017 (40)
2016 (32)
More...

Distributor Name
UiT Open Research Data (625)
UiB Open Research Data (66)

1 to 10 of 825 Results

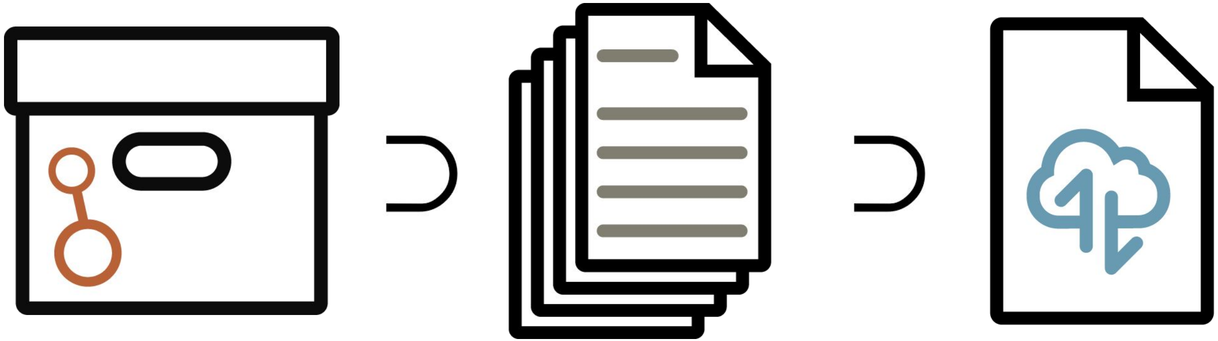
Covid-19 Country Level Social Science Dataset
Oct 20, 2020 - The Stein Rokkan Research Group for Quantitative Social and Political Science
Solvang, Øystein; Eriksen, Kari Elida; Stein, Jonas; Brattland, Camilla, 2020, "Covid-19 Country Level Social Science Dataset", https://doi.org/10.18710/VMUP44, DataverseNO, V2
The dataset is a cross-sectional dataset covering social and public health data pertaining to the Covid-19 outbreak in 199 countries. The dataset was compiled from public register and other openly available sources. Data on Covid-19 cases and related fatalities is current as of m...

Replication data for: Devonian–Carboniferous collapse and segmentation of the Billefjorden Trough, and Eurekan inversion–overprint and strain partitioning and decoupling along inherited WNW–ESE-striking faults
Oct 16, 2020 - UiT Open Research Data
Koehl, Jean-Baptiste P., 2020, "Replication data for: Devonian–Carboniferous collapse and segmentation of the Billefjorden Trough, and Eurekan inversion–overprint and strain partitioning and decoupling along inherited WNW–ESE-striking faults", https://doi.org/10.18710/EAZDNV, DataverseNO, V1
High-resolution versions of the figures and supplements of the Koehl et al. (2020) manuscript entitled "Devonian–Carboniferous collapse and segmentation of the Billefjorden Trough, and Eurekan inversion–overprint and strain partitioning and decoupling along inherited WNW–ESE-stri...

Replication Data and appendix for: Hysteresis of natural magnetite ensembles: Micromagnetics of silicate-hosted magnetite inclusions based on focused-ion beam nanotomography
Oct 15, 2020 - NTNU Open Research Data

Key Dataverse Features

Organization of a Dataverse Repository



Dataverse

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

dataset

Citation
Metadata
Versioning
Terms/permissions
Collection of Files

File

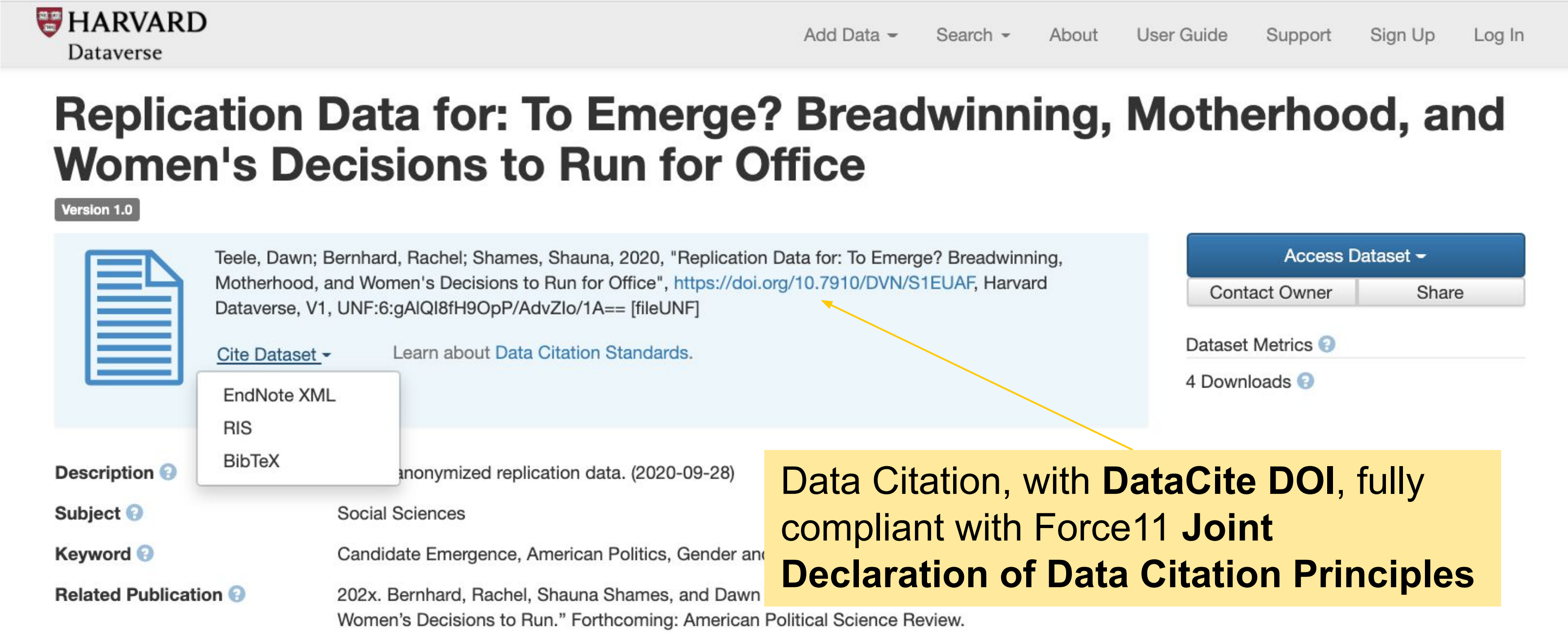
Citation
Preview/Explore
Metadata
Versioning
Permissions

Features for FAIR and Responsible Data Sharing

- Data Citation with DOI for datasets and files
- Credit to data authors
- Link from data to related article
- Standards-based and custom metadata
    - DDI, Schema.org, DataCite, Dublin Core, OAI-ORE, OpenAire, JSON-LD
- 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 and hierarchical folder structure: Web UI, API, Standalone Client
- Integration with external tools through extensive API
- Make Data Count (coming soon)

FINDABILITY:

Full, standard data citation automatically generated



Replication Data for: To Emerge? Breadwinning, Motherhood, and Women's Decisions to Run for Office

Teele, Dawn; Bernhard, Rachel; Shames, Shauna, 2020, "Replication Data for: To Emerge? Breadwinning, Motherhood, and Women's Decisions to Run for Office", https://doi.org/10.7910/DVN/S1EUAF, Harvard Dataverse, V1, UNF:6:gAlQI8fH9OpP/AdvZlo/1A== [fileUNF]

Learn about Data Citation Standards.

Dataset Metrics

4 Downloads

Description anonymized replication data. (2020-09-28)

Subject Social Sciences

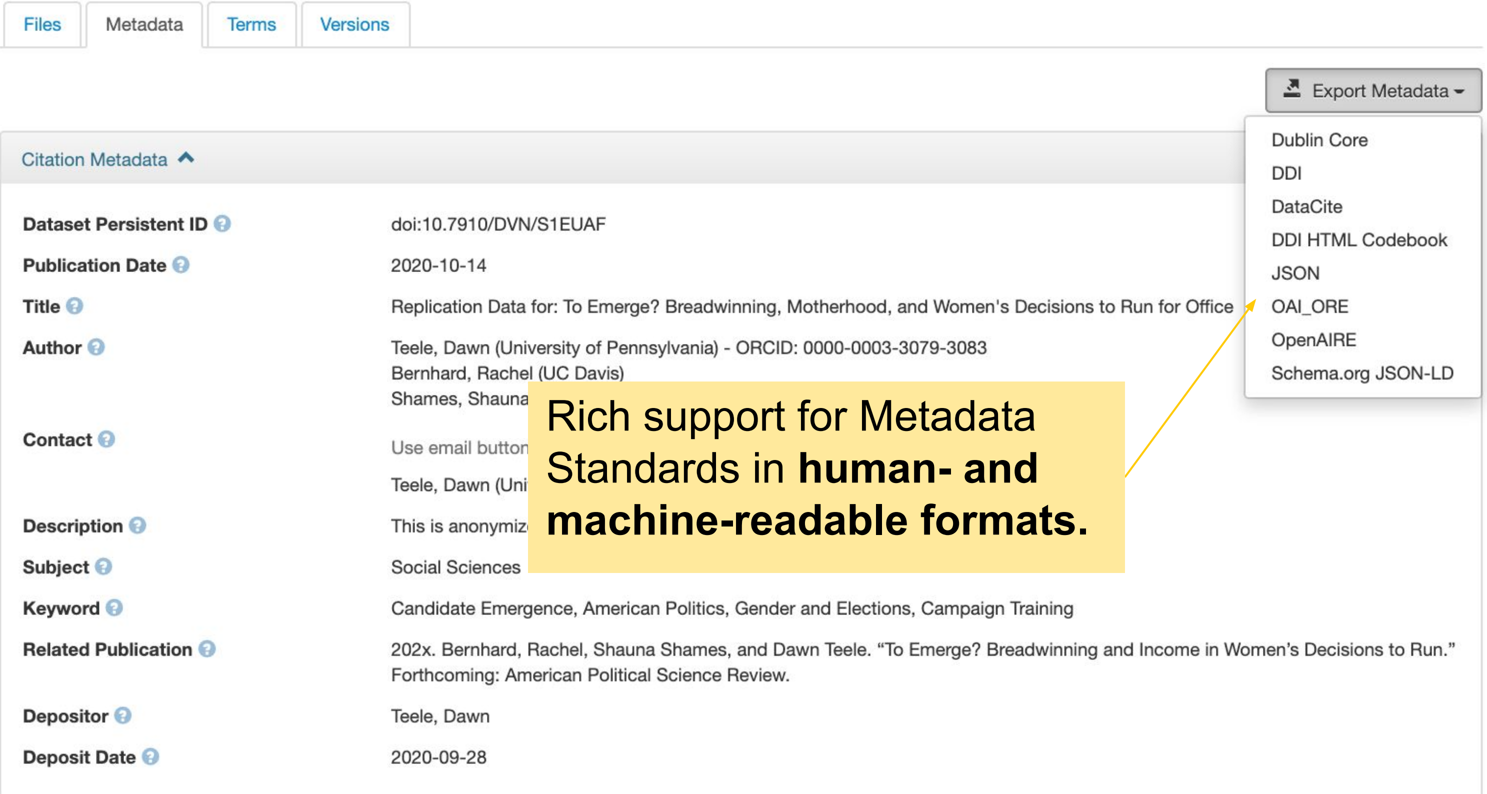
Keyword Candidate Emergence, American Politics, Gender anc

Related Publication 202x. Bernhard, Rachel, Shauna Shames, and Dawn
Women's Decisions to Run." Forthcoming: American Political Science Review.

Data Citation, with **DataCite DOI**, fully compliant with Force11 **Joint Declaration of Data Citation Principles**

FINDABILITY AND REUSABILITY:

Support for multiple metadata standards

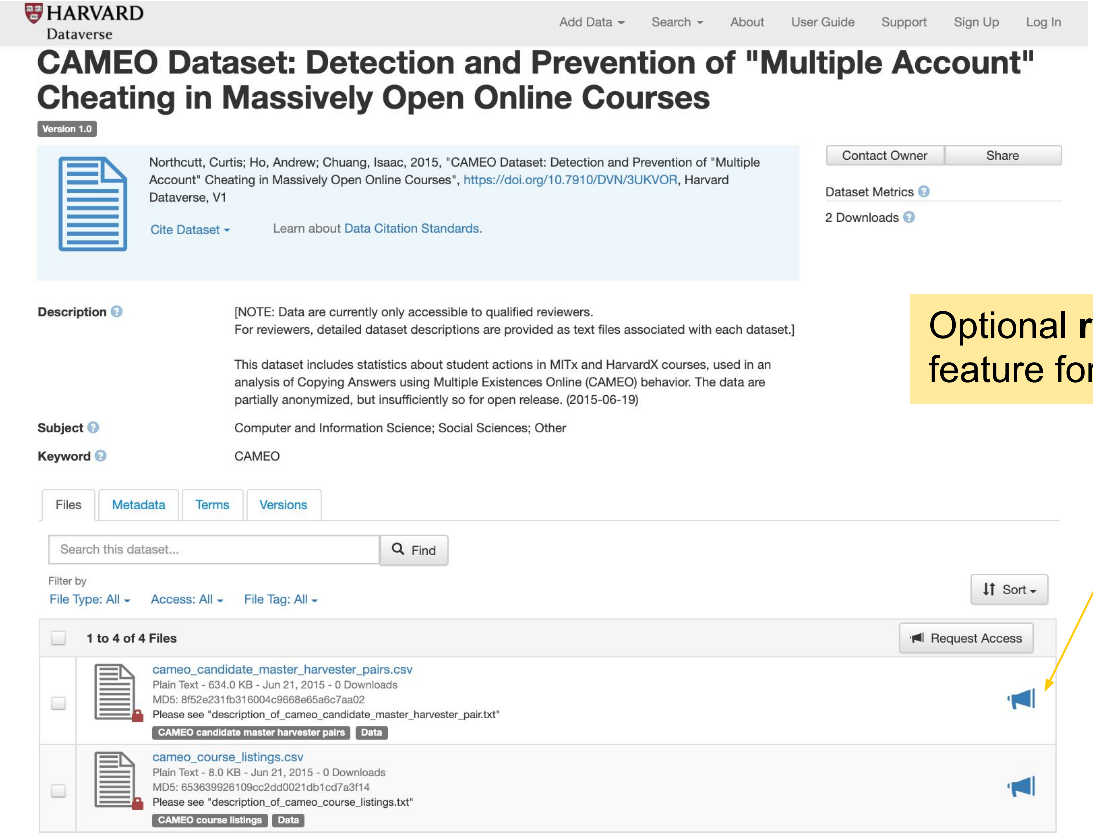


Dataset Persistent ID	doi:10.7910/DVN/S1EUAF
Publication Date	2020-10-14
Title	Replication Data for: To Emerge? Breadwinning, Motherhood, and Women's Decisions to Run for Office
Author	Teele, Dawn (University of Pennsylvania) - ORCID: 0000-0003-3079-3083 Bernhard, Rachel (UC Davis) Shames, Shauna
Contact	Use email button Teele, Dawn (Uni
Description	This is anonymiz
Subject	Social Sciences
Keyword	Candidate Emergence, American Politics, Gender and Elections, Campaign Training
Related Publication	202x. Bernhard, Rachel, Shauna Shames, and Dawn Teele. “To Emerge? Breadwinning and Income in Women’s Decisions to Run.” Forthcoming: American Political Science Review.
Depositor	Teele, Dawn
Deposit Date	2020-09-28

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

ACCESSIBILITY:

Access terms available for restricted data



CAMEO Dataset: Detection and Prevention of "Multiple Account" Cheating in Massively Open Online Courses

Version 1.0

Northcutt, Curtis; Ho, Andrew; Chuang, Isaac, 2015, "CAMEO Dataset: Detection and Prevention of "Multiple Account" Cheating in Massively Open Online Courses", https://doi.org/10.7910/DVN/3UKVOR, Harvard Dataverse, V1

Learn about Data Citation Standards.

Dataset Metrics

2 Downloads

Description [NOTE: Data are currently only accessible to qualified reviewers.
For reviewers, detailed dataset descriptions are provided as text files associated with each dataset.]

This dataset includes statistics about student actions in MITx and HarvardX courses, used in an analysis of Copying Answers using Multiple Existences Online (CAMEO) behavior. The data are partially anonymized, but insufficiently so for open release. (2015-06-19)

Subject Computer and Information Science; Social Sciences; Other

Keyword CAMEO

□ **1 to 4 of 4 Files**

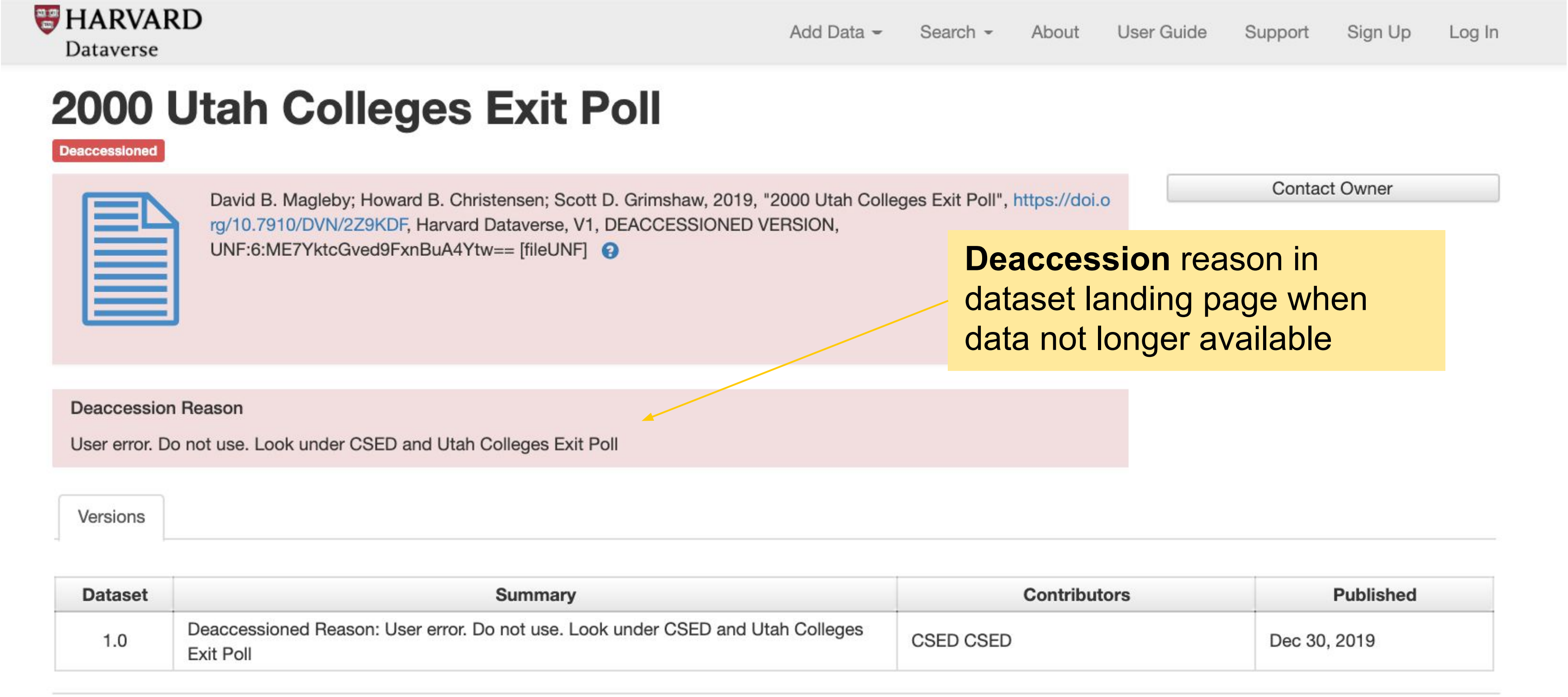
□ cameo_candidate_master_harvester_pairs.csv
Plain Text - 634.0 KB - Jun 21, 2015 - 0 Downloads
MD5: 8f52e231fb316004c9668e65a6c7aa02
Please see "description_of_cameo_candidate_master_harvester_pair.txt"
CAMEO candidate master harvester pairs | Data

□ cameo_course_listings.csv
Plain Text - 8.0 KB - Jun 21, 2015 - 0 Downloads
MD5: 653639926109cc2dd0021db1cd7a3f14
Please see "description_of_cameo_course_listings.txt"
CAMEO course listings | Data

Optional **request access** feature for restricted data

ACCESSIBILITY:

Metadata always available



2000 Utah Colleges Exit Poll

Deaccessioned

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 in dataset landing page when data not longer available

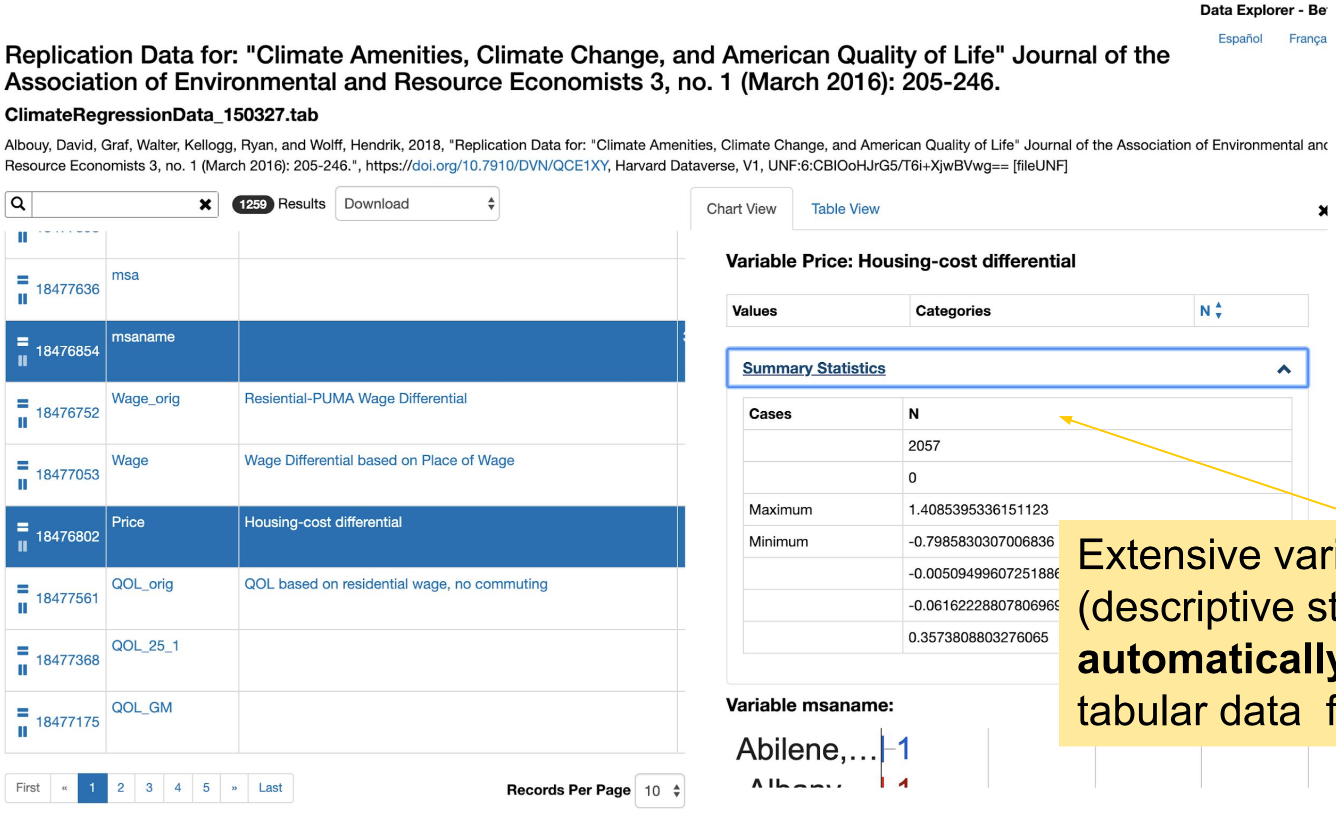
Deaccession Reason

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

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 statistics metadata derived from each variable in data file



Data Explorer - Be

Español França

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

Albouy, 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 anc 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]

18477636	msa	
18476854	msaname	
18476752	Wage_orig	Resiential-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	

Variable Price: Housing-cost differential

Values	Categories	N

Summary Statistics

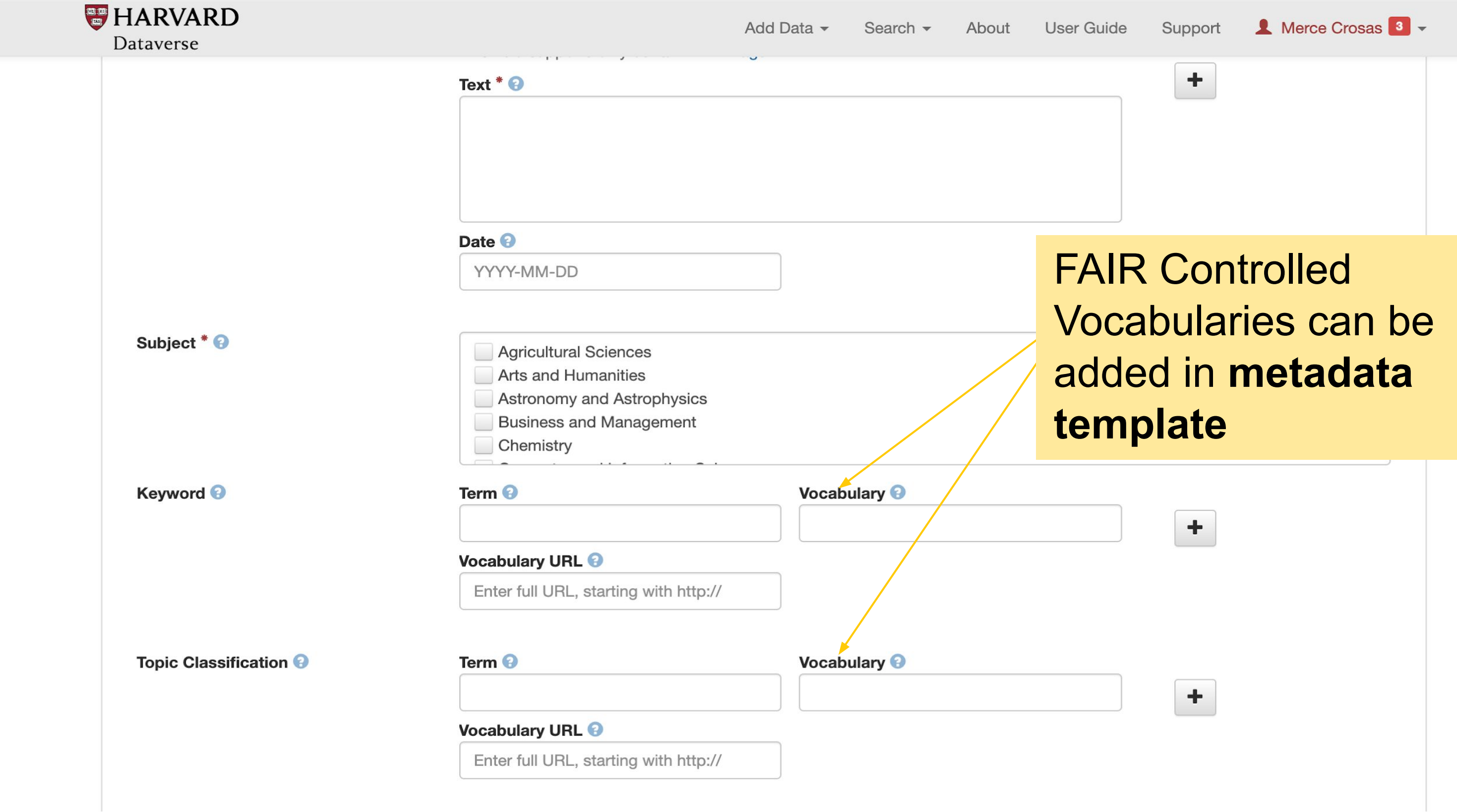
Cases	N
	2057
	0
Maximum	1.4085395336151123
Minimum	-0.7985830307006836
	-0.00509499607251886
	-0.0616222880780696
	0.3573808803276065

Variable msaname:

Extensive variable metadata (descriptive statistics) **automatically** derived from tabular data file in DDI format

INTEROPERABLE:

Enable use of standard vocabularies



Text *

Date

YYYY-MM-DD

Subject *

□ Agricultural Sciences
□ Arts and Humanities
□ Astronomy and Astrophysics
□ Business and Management
□ Chemistry

Keyword

Term

Vocabulary

Vocabulary URL

Enter full URL, starting with http://

Topic Classification

Term

Vocabulary

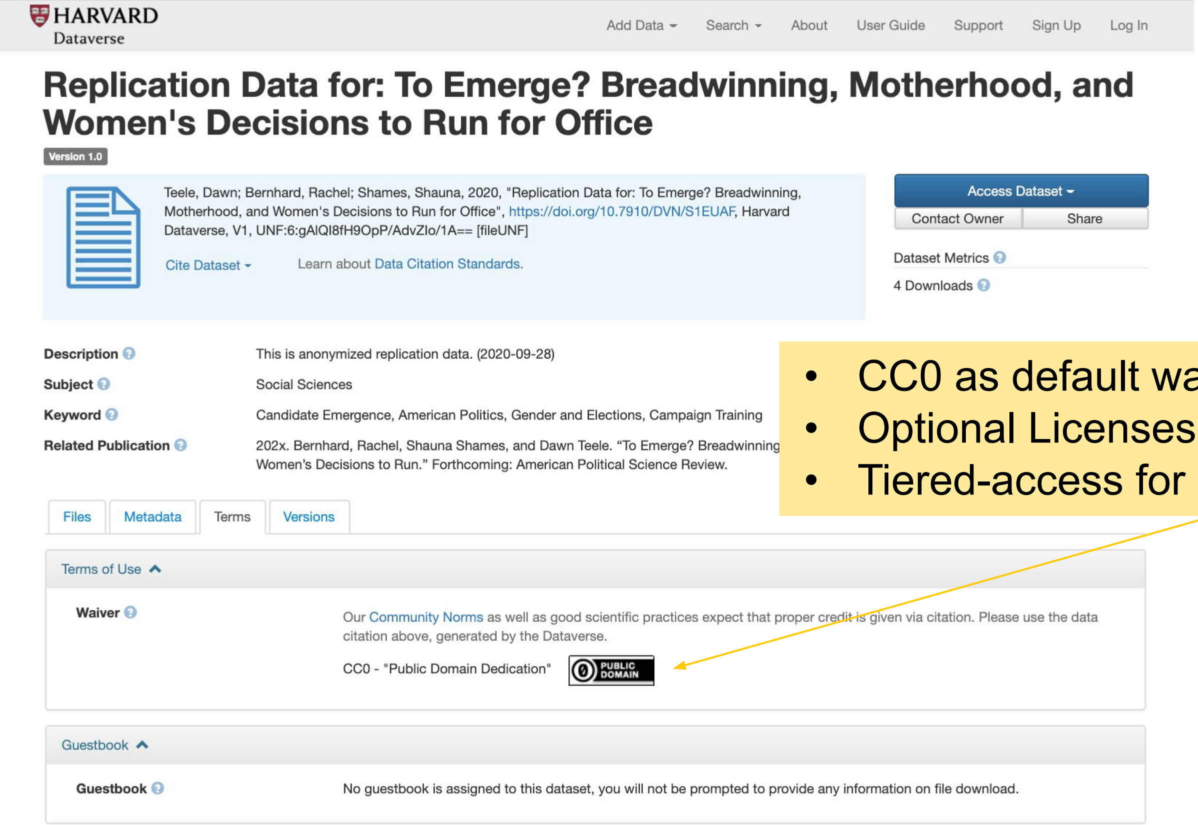
Vocabulary URL

Enter full URL, starting with http://

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

REUSABLE:

Licenses, Terms, and Tiered-Access to Data



Replication Data for: To Emerge? Breadwinning, Motherhood, and Women's Decisions to Run for Office

Teele, Dawn; Bernhard, Rachel; Shames, Shauna, 2020, "Replication Data for: To Emerge? Breadwinning, Motherhood, and Women's Decisions to Run for Office", https://doi.org/10.7910/DVN/S1EUAF, Harvard Dataverse, V1, UNF:6:gAlQI8fH9OpP/AdvZlo/1A== [fileUNF]

Learn about Data Citation Standards.

Dataset Metrics

4 Downloads

Description	This is anonymized replication data. (2020-09-28)
Subject	Social Sciences
Keyword	Candidate Emergence, American Politics, Gender and Elections, Campaign Training
Related Publication	202x. Bernhard, Rachel, Shauna Shames, and Dawn Teele. “To Emerge? Breadwinning Women’s Decisions to Run.” Forthcoming: American Political Science Review.

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"

Guestbook

Guestbook: No guestbook is assigned to this dataset, you will not be prompted to provide any information on file download.

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

What’s Next

Sensitive Data

Support for Sensitive Data in Dataverse

Non-Sensitive DataTags (TODAY)

- Data uploaded to Dataverse via one of the current options
- Stored locally

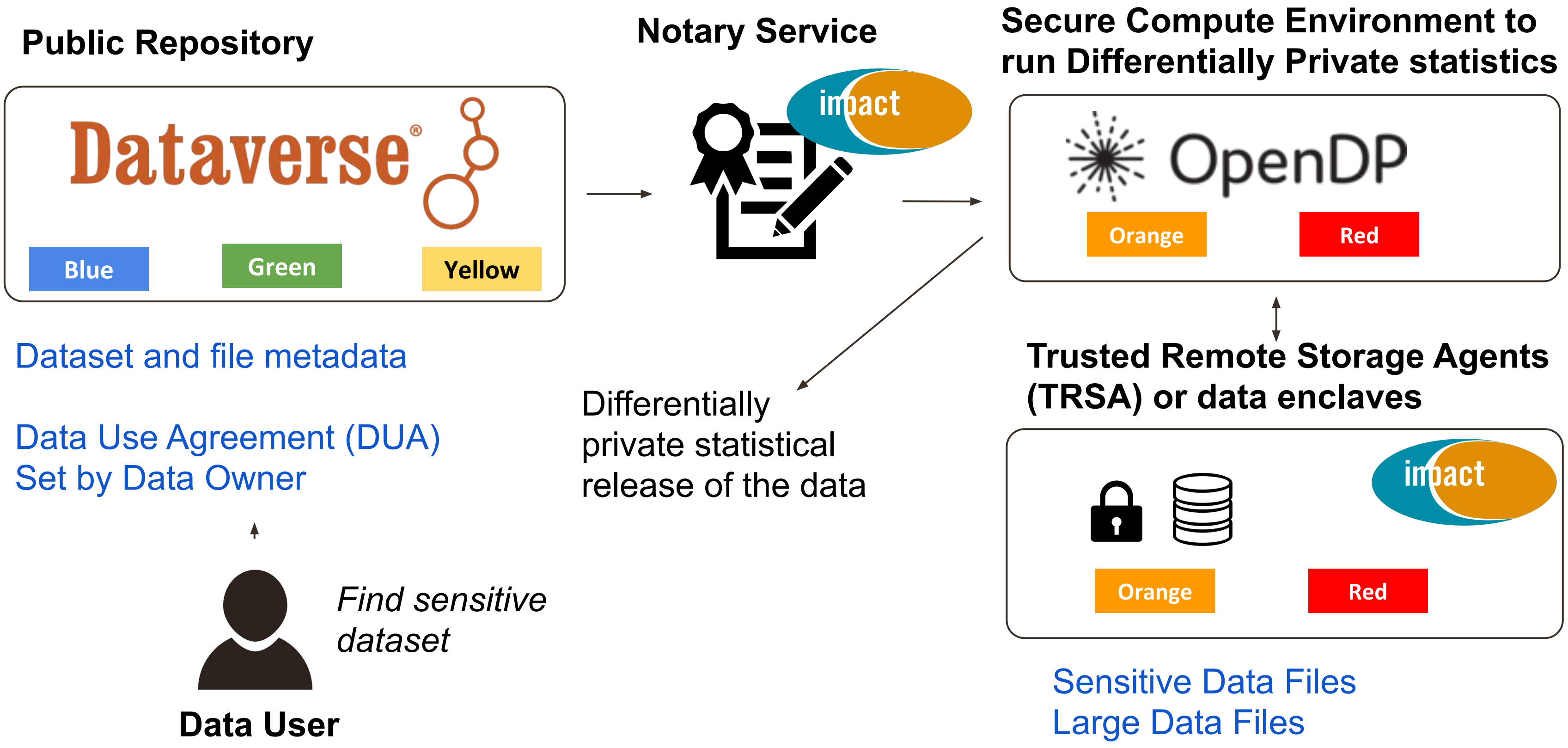
Blue	Publicly open, no barriers
Green	Publicly open, but need to register to access
Yellow	Restricted, need to be granted permissions, but non-sensitive

Sensitive DataTags (FUTURE RELEASE)

- Stored in a **Trusted Remote Storage Agent,** accessed through notary service
- Metadata published in Dataverse

Orange	Requires Data Use Agreement (DUA); requires data enclave ***(moderate sensitivity)***
Red	Requires DUA; stricter security requirements and audits ***(high sensitivity)***
Crimson	Only metadata and no link to data; data stored outside network ***(maximum sensitivity)***

Dataverse + Impact + OpenDP



Public Repository

Dataset and file metadata

Data Use Agreement (DUA)
Set by Data Owner

Find sensitive dataset

Data User

Notary Service

Differentially private statistical release of the data

Secure Compute Environment to run Differentially Private statistics

Trusted Remote Storage Agents (TRSA) or data enclaves

Sensitive Data Files
Large Data Files

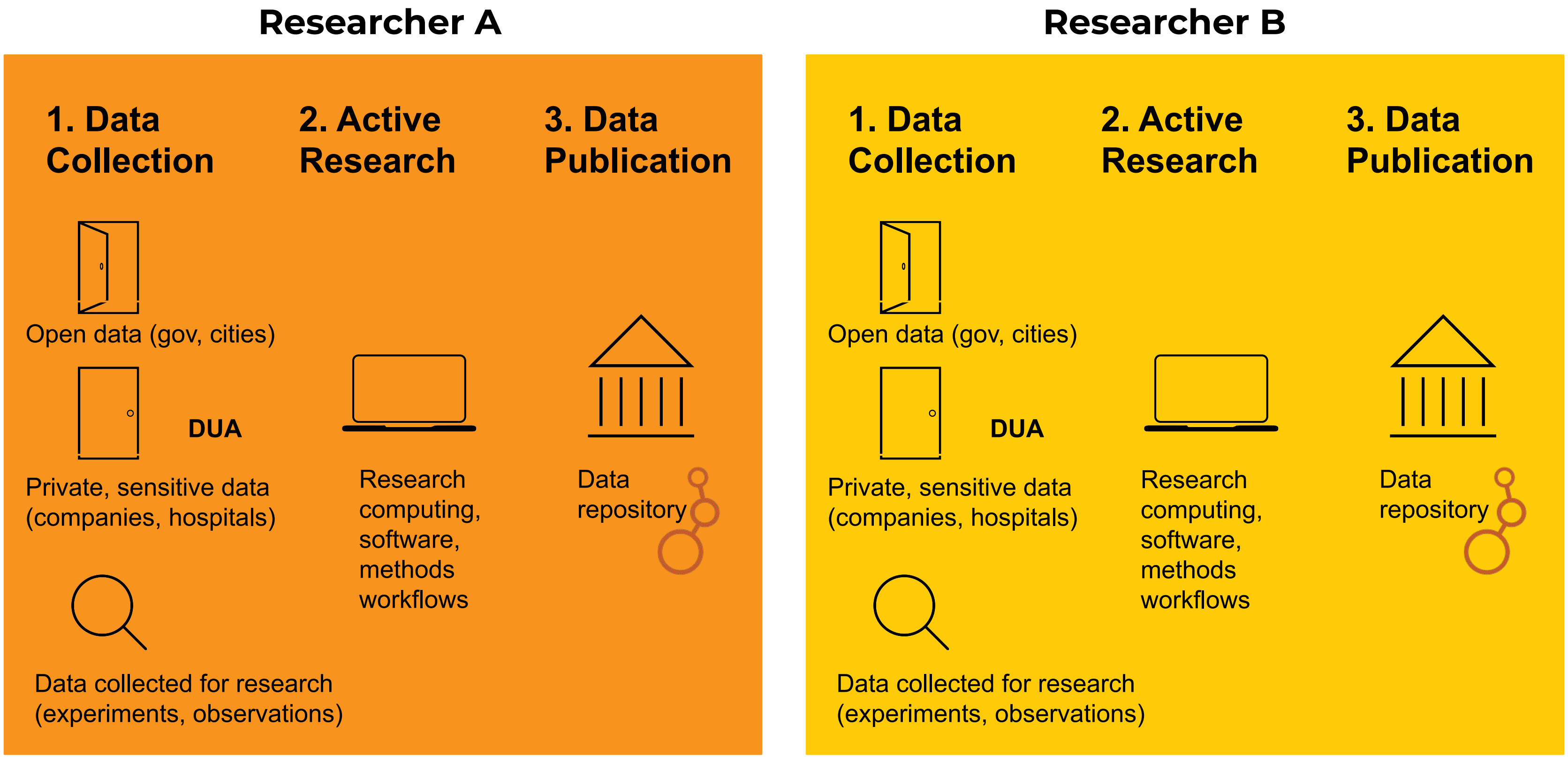
Towards a Data Commons

“a **data commons** brings together (or co-locates) **data with cloud computing** infrastructure and **commonly used** software services, tools & applications for **managing, analyzing and sharing data** to create an **interoperable** resource for a research community.”

[Robert Grossman, on the NIH Data Commons Consortium initiative]

The Problem

- During Active Research, researcher A doesn’t know about data from Researcher B; **not easy to collaborate**
- Research lifecycle steps 1, 2, & 3 are **duplicative** and **not connected**; not easy to manage data throughout lifecycle and publish research output



Researcher A

1. Data Collection

Open data (gov, cities)

DUA

Private, sensitive data (companies, hospitals)

Data collected for research (experiments, observations)

2. Active Research

Research computing, software, methods workflows

3. Data Publication

Data repository

Researcher B

1. Data Collection

Open data (gov, cities)

DUA

Private, sensitive data (companies, hospitals)

Data collected for research (experiments, observations)

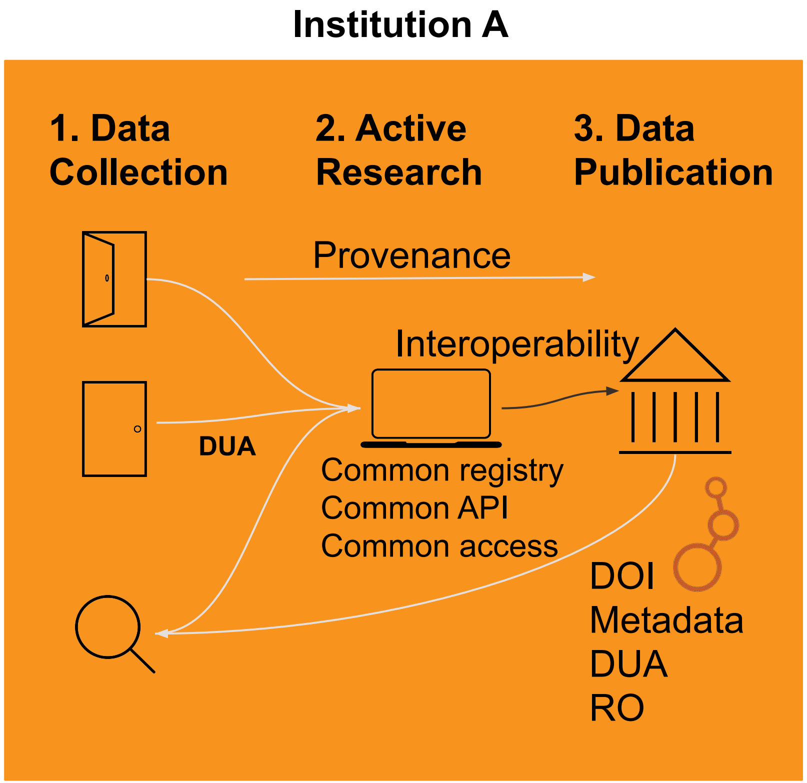
2. Active Research

Research computing, software, methods workflows

3. Data Publication

Data repository

A Solution for an Institution

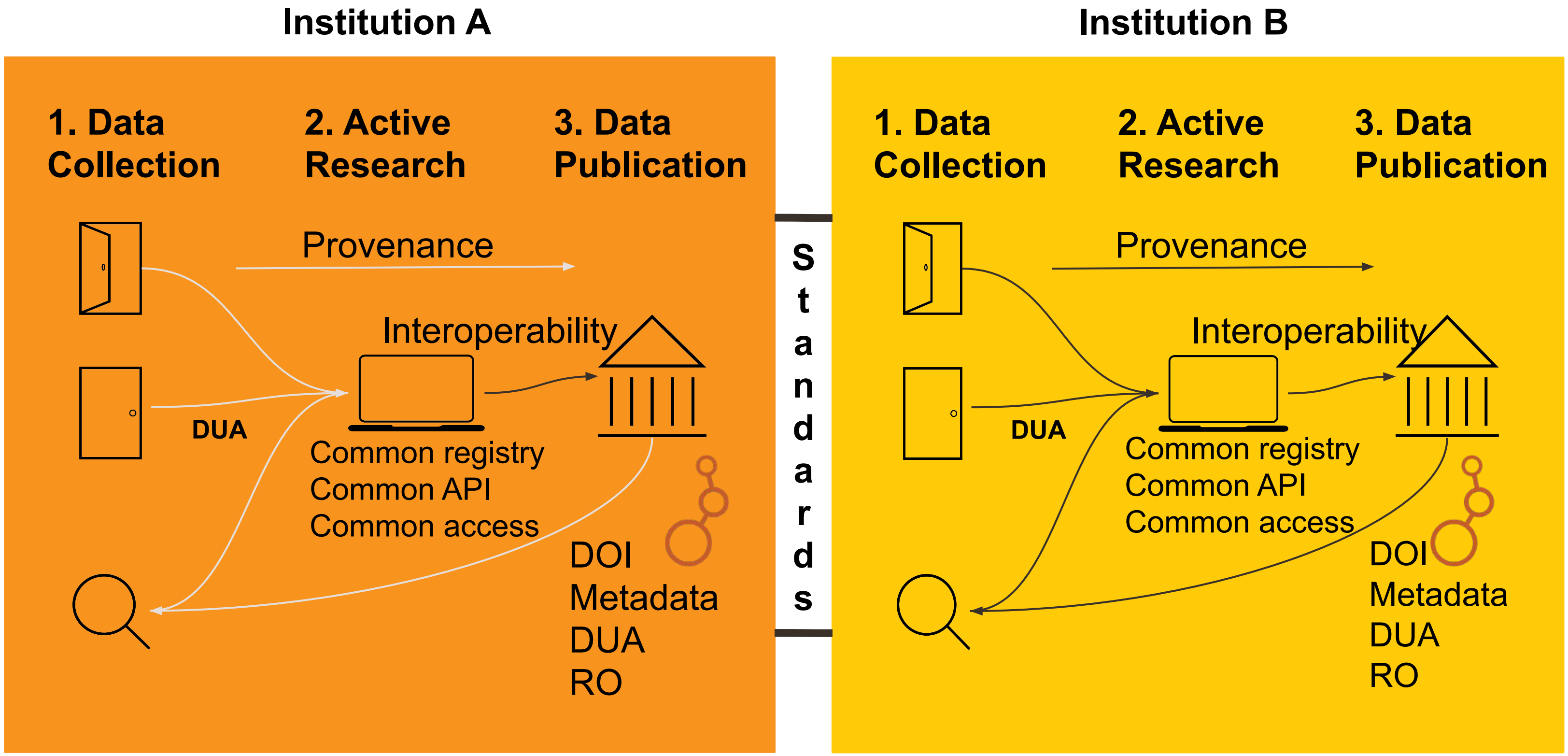


A Commons vision with Dataverse:

- **Common registry** with **metadata**, so Researcher A can find and access data from Researcher B during Active Research
- **Common API** to integrate with research tools and computing to access the data seamlessly
- **DUA and controlled access** tracked and throughout lifecycle and shared through Dataverse
- **Provenance** tracked throughout the lifecycle to produce reproducible research outputs – package data, code, workflows in a Research Object(RO)

A Global Solution

Agree on standards so Data Commons A can talk to Data Commons B



THANKS

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