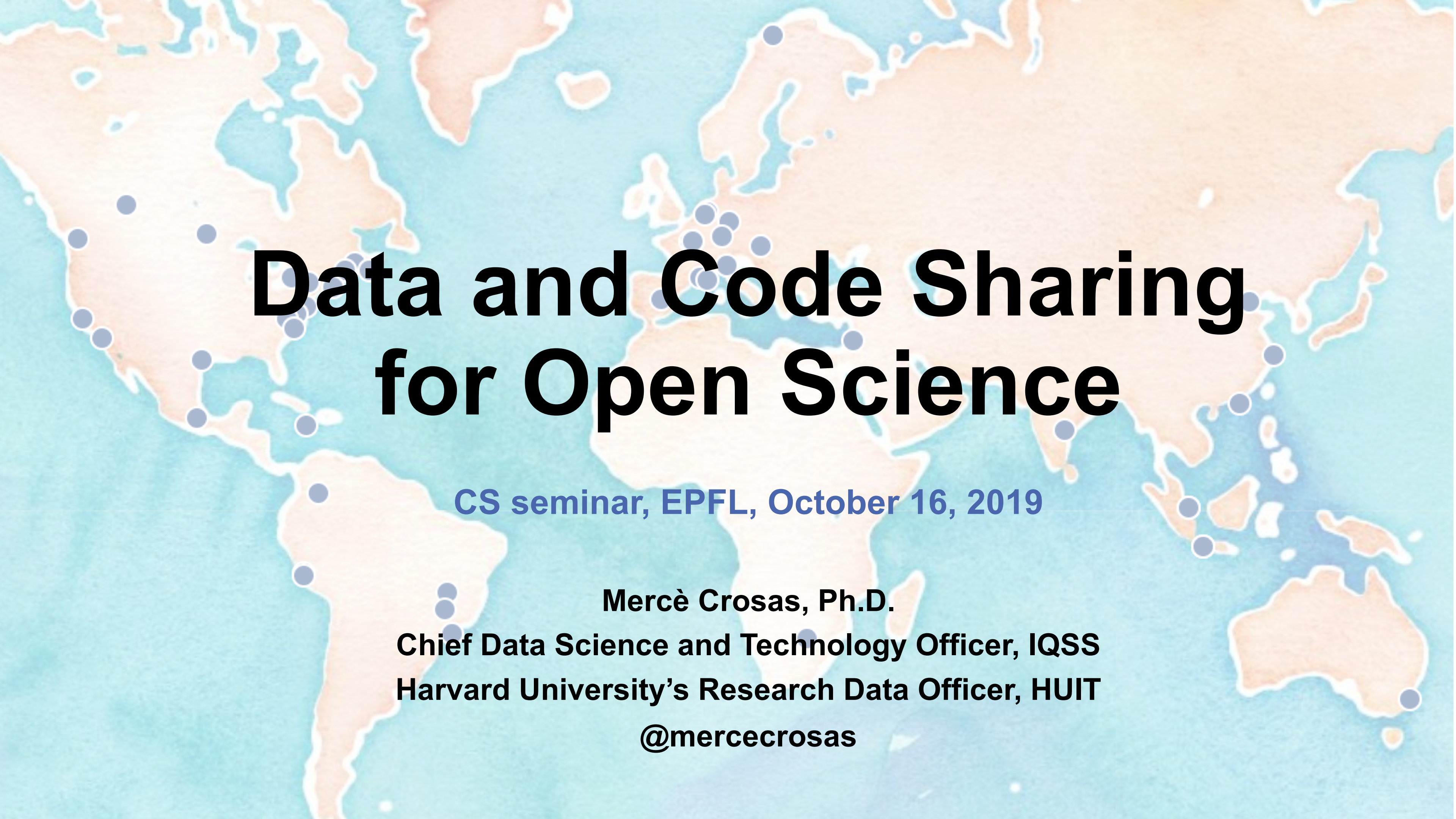


Data and Code Sharing for Open Science

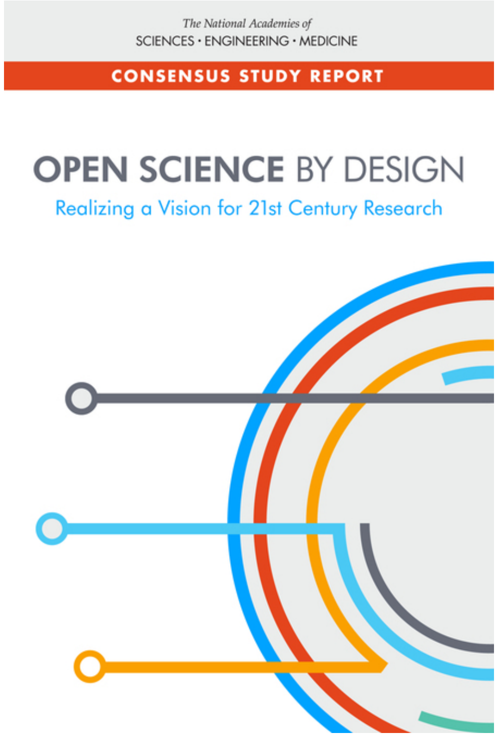
CS seminar, EPFL, October 16, 2019

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

This Talk

1. Open Science awareness
2. Open Science implementation considerations
3. Our Contribution to data sharing and reproducibility: Dataverse
4. Our Contribution to sensitive data sharing and analysis: DataTags and OpenDP

National Academies of Sciences New Report



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

OPEN SCIENCE BY DESIGN

Realizing a Vision for 21st Century Research

Researcher at the center, both contributes to open science and takes advantage of the open science practices:

- **Knowledge generation:** collect data, conduct research using tools compatible with **open sharing;** use **automated workflow tools** to ensure accessibility of research outputs.
- **Validation:** prepare data and tools for **reproducibility and reuse** and participate in replication studies.
- **Dissemination:** use **appropriate licenses** for sharing research outputs; report all results and supporting information.
- **Preservation:** deposit research outputs in **FAIR archives** and ensure long-term access to research results

OpenAire New White Paper

OpenAIRE

Achieving Open Science in the European Open Science Cloud

Setting out OpenAIRE's vision and contribution to EOSC

Position Paper - September 2019

Researcher-centric, with openness as default in scientific communication:

- Publishing all kinds of **scientific products**
- Publishing **semantic links**
- Publishing research products as packages of **workflows**
- **Innovation** in publishing and dissemination practices
- Quality control for securing the **quality, reproducibility, FAIRness** of research results
- Assessment and **reward** by intelligently combining diverse/open/auditable metrics

New Survey by Pew Research

"Americans say open access to data and independent review inspire more trust in research findings"

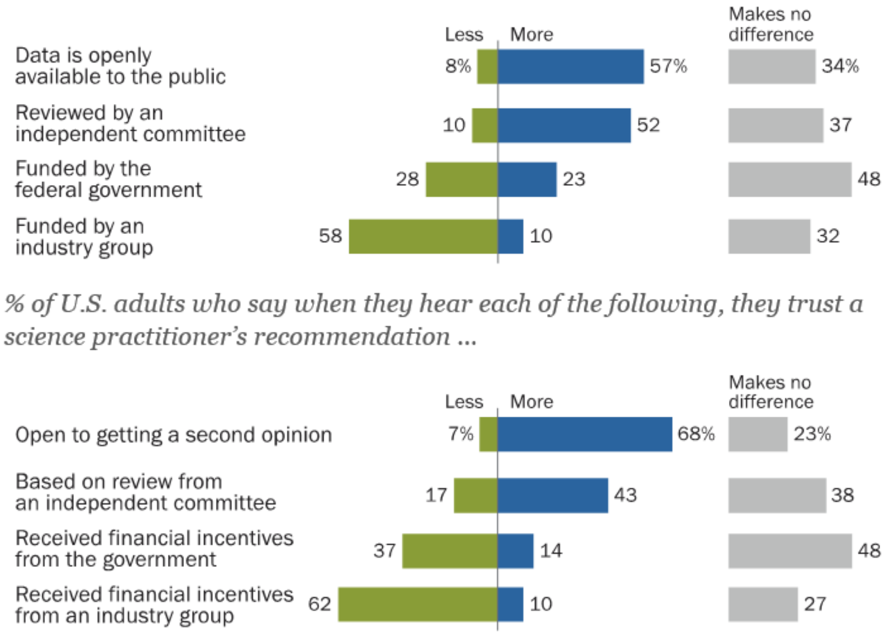
Trust and Mistrust of American Views on Scientific Experts.

Pew Research Center, August 2, 2019

Majority of Americans say they are more apt to trust research when the data is openly available

% of U.S. adults who say when they hear each of the following, they trust scientific research findings …

% of U.S. adults who say when they hear each of the following, they trust a science practitioner's recommendation …



Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 7-21, 2019.
"Trust and Mistrust in Americans' Views of Scientific Experts"

PEW RESEARCH CENTER

What Open Science Is and Isn’t

- **Open science can be more beneficial to humanity**
  - Wider reach of data and knowledge, greater impact to human welfare
- **Open Science does not guarantee good science**
  - Facilitates reproducibility validation, but data/research might still be low quality
- **Open Science does not always mean fully public data**
  - Research data should be discoverable, but protected if sensitive

Open Science Implementation

An open science technology implementation must consider and include:

- **Incentives** to share data, code, and other research outputs
- **Standards** for metadata discovery
- **Public metadata** (at a minimum for citation) even when data are restricted
- Sufficient information to **reuse** the data, code, workflows (all research outputs)
- Support for data and code **terms and use agreements**
- Integration of archival **repositories** with **computational platforms** (clouds, tools)
- Solutions for collaborations that access **sensitive, private data**

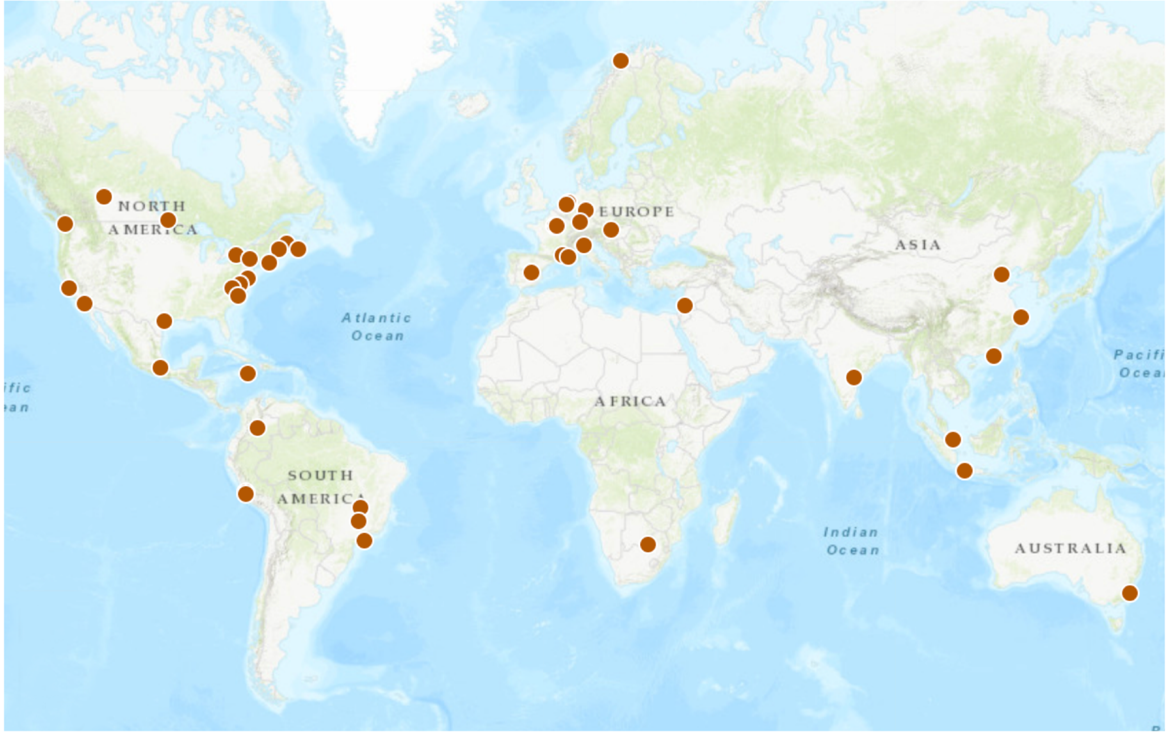
Dataverse – Achievements

Facilitating data and code sharing for Open Science

15 years of Dataverse

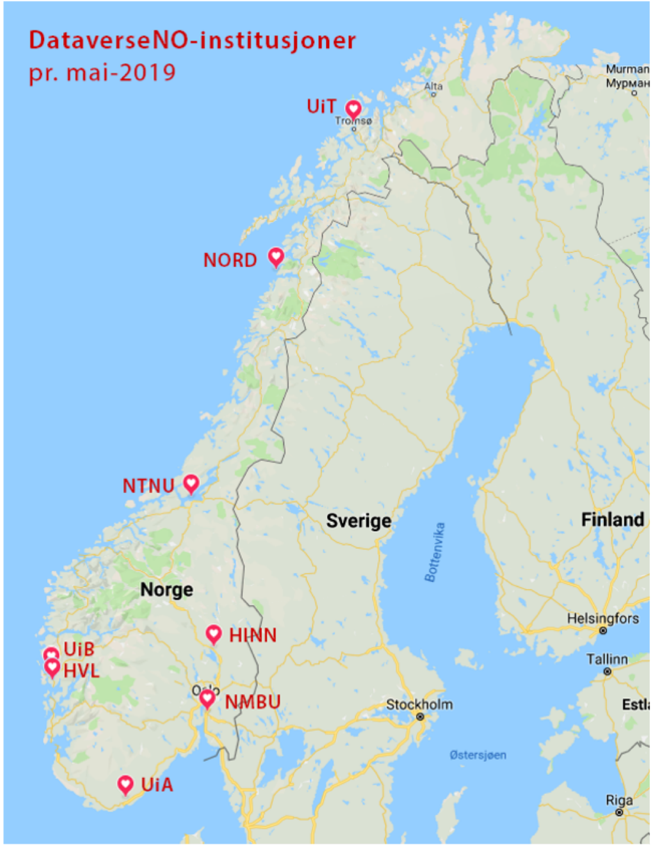
- Launched in **2006** by Harvard's Institute for Quantitative Social Science
- Now **48** Dataverse installations sites throughout 6 continents
- Each site can support **multiple** Universities or research organizations
- **Harvard Dataverse** is the largest Dataverse repository site with:
  - **3,400** dataverses (collections of datasets)
  - **91,000** datasets
  - **560,000** files
  - **8.9** million downloads
- An active, growing **open-source community** with 100 contributors, 41 releases

Sites harvest metadata from each other



- 48 Dataverse sites can be federated via the **Open Archives Initiative Protocol** for Metadata Harvesting.
- Data **searchable** in one Dataverse site while hosted in another.

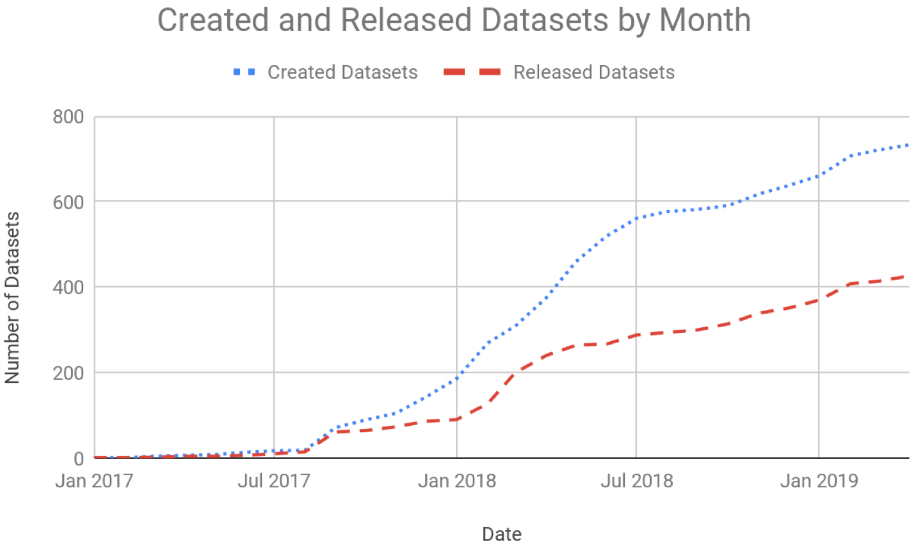
An National Example: DataverseNO



DataverseNO
Dataverse Network Norway

- **8 universities in Norway** as members; the other 3 to join soon
- **Policies and guidelines common** to all DataverseNO members
- Global and local support
- Applied for Core Trust Seal certificate
- https://site.uit.no/dataverseno/

A Consortium Example: Texas Dataverse

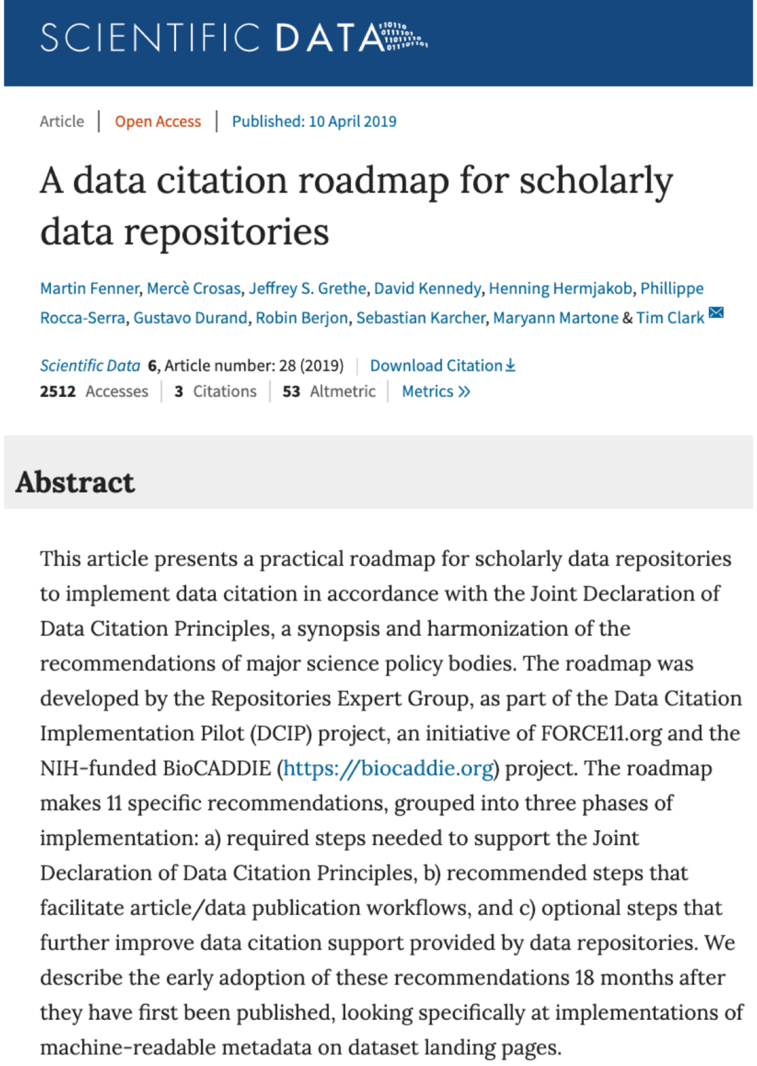


Texas Data Repository

- **11 universities in Texas** as members
- Led by Texas Digital Library **Consortium**
- Texas Data Repository steering committee focuses on **outreach**
- Working on Core Trust Seal
- Differing policies/practices per member
- https://www.tdl.org/texas-data-repository/

Combines technology, standards, incentives

- A **data citation** with a globally unique persistent identifier and credit for data author
- Standard **metadata**, with variable-level metadata, plus rich custom metadata
  - Schema.org JSON-LD, DataCite, Dublin Core, DDI, OAI-ORE, OpenAire, PROV
- **Tiered data access**, depending on data restrictions:
  - Fully Open, CC0; Guestbook; Restricted w/ Data Use Agreement
- **Data publishing workflows**, with anonymous review, and reviewer/curator roles
- Multiple **versions** of a dataset
- **Branding and customization** for each dataverse (collection of datasets)
- Overall, follows **FAIR** guiding principles



SCIENTIFIC DATA

Article | Open Access | Published: 10 April 2019

A data citation roadmap for scholarly data repositories

Martin Fenner, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone & Tim Clark ✉

Scientific Data **6**, Article number: 28 (2019) | Download Citation ⤓

2512 Accesses | **3** Citations | **53** Altmetric | Metrics »

Abstract

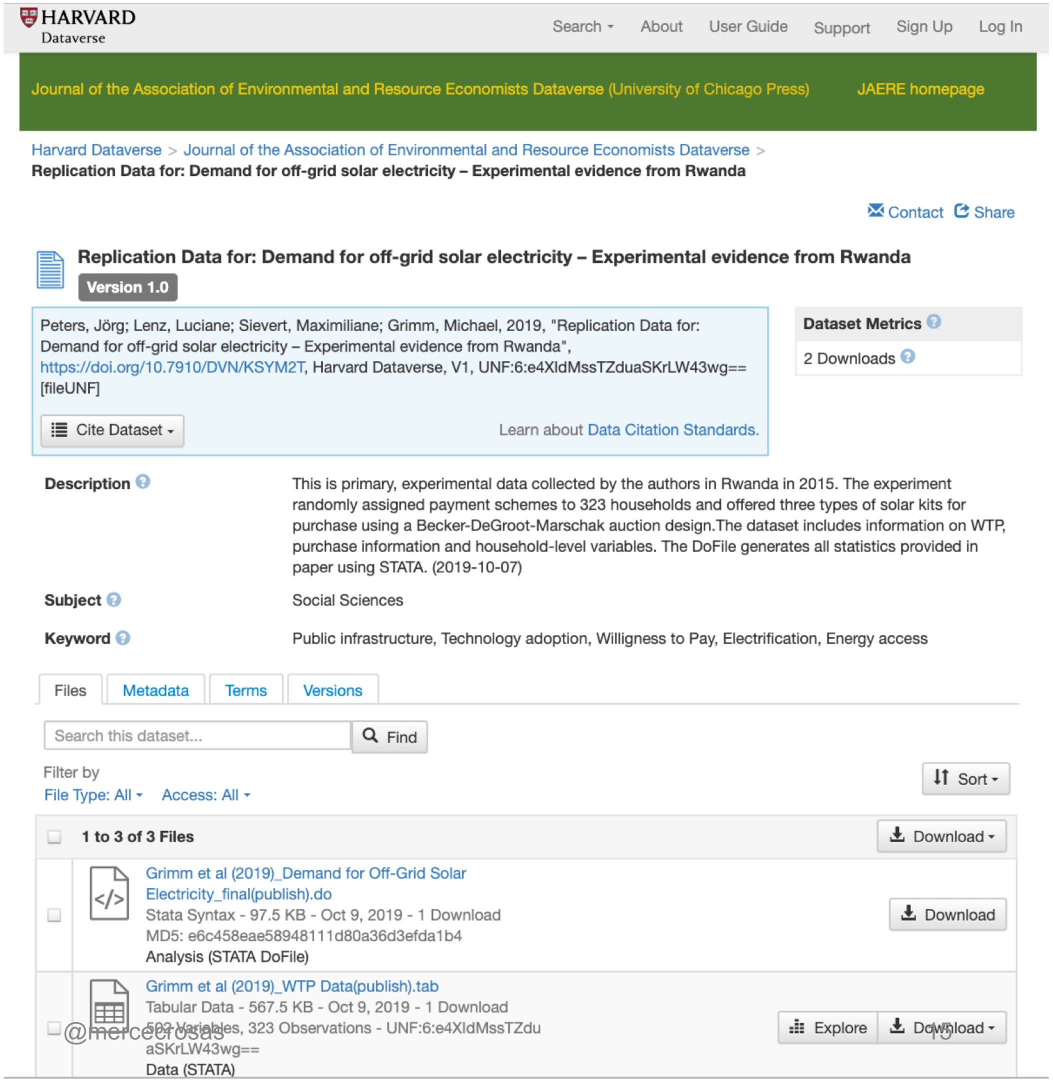
This article presents a practical roadmap for scholarly data repositories to implement data citation in accordance with the Joint Declaration of Data Citation Principles, a synopsis and harmonization of the recommendations of major science policy bodies. The roadmap was developed by the Repositories Expert Group, as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH-funded BioCADDIE (https://biocaddie.org) project. The roadmap makes 11 specific recommendations, grouped into three phases of implementation: a) required steps needed to support the Joint Declaration of Data Citation Principles, b) recommended steps that facilitate article/data publication workflows, and c) optional steps that further improve data citation support provided by data repositories. We describe the early adoption of these recommendations 18 months after they have first been published, looking specifically at implementations of machine-readable metadata on dataset landing pages.

Data Citation Implementation

- Global unique persistent identifiers (GUID)
- Allow to cite specific version or subset
- GUID resolves to dataset landing page
- Landing page w/ instructions to access data
- Machine-readable citation metadata
- Schema.org JSON-LD in landing page
- Standard Bibliographic formats

Dataset Landing Page

- Dataset and file citation
- Data files review and downloads
  - Automatic transforms tabular files to multiple formats
  - Extracts variable metadata
- Dataset Metadata
- Metrics: Make Data Count
- Terms and data use agreements
- Versions and Provenance
- APIs to access metadata & data
- Machine-actionable



Replication Data for: Demand for off-grid solar electricity – Experimental evidence from Rwanda

Replication Data for: Demand for off-grid solar electricity – Experimental evidence from Rwanda

Peters, Jörg; Lenz, Luciane; Sievert, Maximiliane; Grimm, Michael, 2019, "Replication Data for: Demand for off-grid solar electricity – Experimental evidence from Rwanda", https://doi.org/10.7910/DVN/KSYM2T, Harvard Dataverse, V1, UNF:6:e4XIdMssTZduaSKrLW43wg== [fileUNF]

Learn about Data Citation Standards.

Dataset Metrics

2 Downloads

Description This is primary, experimental data collected by the authors in Rwanda in 2015. The experiment randomly assigned payment schemes to 323 households and offered three types of solar kits for purchase using a Becker-DeGroot-Marschak auction design.The dataset includes information on WTP, purchase information and household-level variables. The DoFile generates all statistics provided in paper using STATA. (2019-10-07)

Subject Social Sciences

Keyword Public infrastructure, Technology adoption, Willigness to Pay, Electrification, Energy access

1 to 3 of 3 Files

Grimm et al (2019)_Demand for Off-Grid Solar Electricity_final(publish).do
Stata Syntax - 97.5 KB - Oct 9, 2019 - 1 Download
MD5: e6c458eae58948111d80a36d3efda1b4
Analysis (STATA DoFile)

Grimm et al (2019)_WTP Data(publish).tab
Tabular Data - 567.5 KB - Oct 9, 2019 - 1 Download
502 Variables, 323 Observations - UNF:6:e4XIdMssTZdu aSKrLW43wg==
Data (STATA)

Dataverse – Next

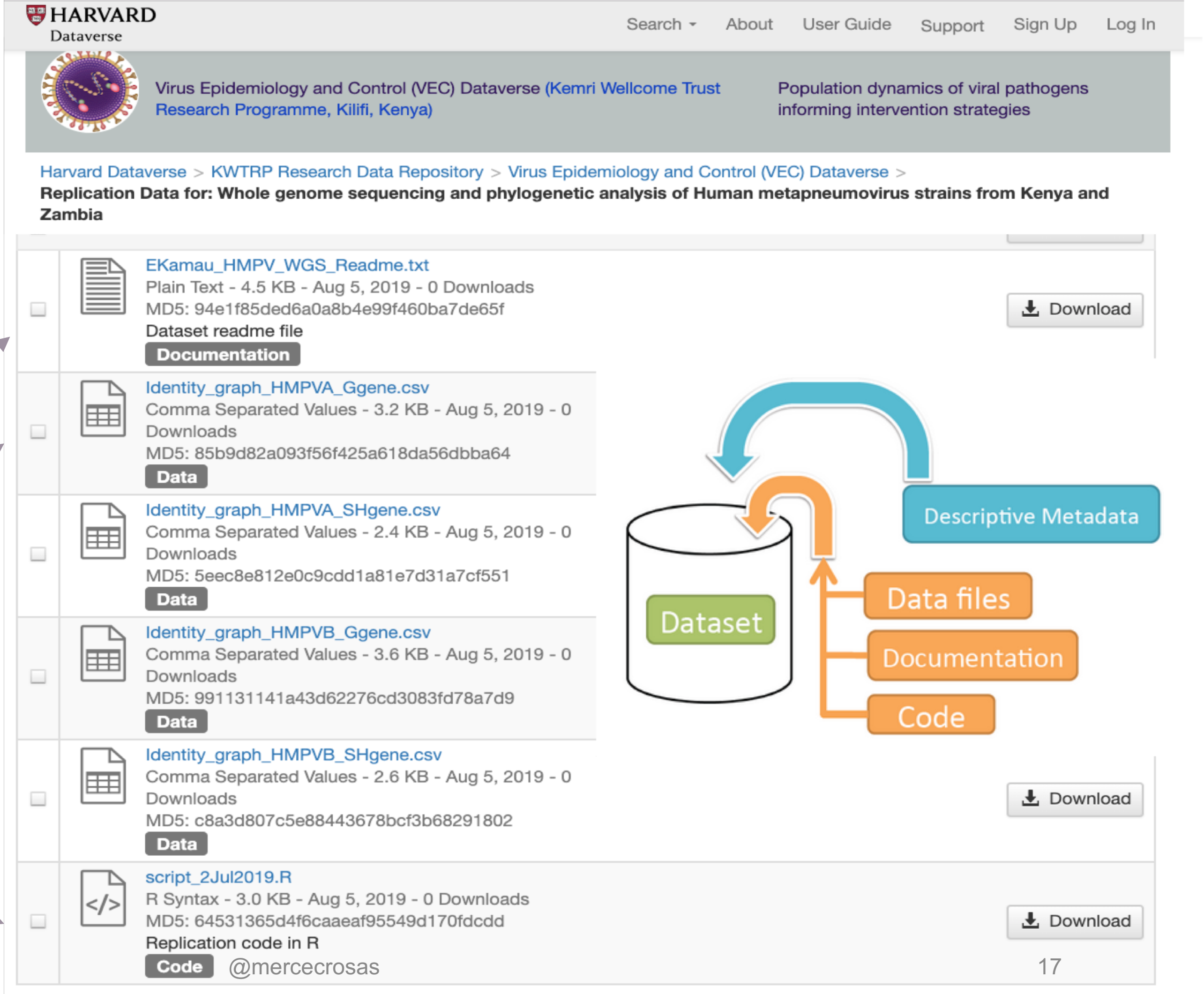
Improving research reproducibility and data reuse

8,000 of the 90,000 datasets in Harvard Dataverse contain the files to reproduce the published results

documentation

data

code



HARVARD Dataverse

Search ▾ About User Guide Support Sign Up Log In

Virus Epidemiology and Control (VEC) Dataverse (Kemri Wellcome Trust Research Programme, Kilifi, Kenya)

Population dynamics of viral pathogens informing intervention strategies

Harvard Dataverse > KWTRP Research Data Repository > Virus Epidemiology and Control (VEC) Dataverse >

Replication Data for: Whole genome sequencing and phylogenetic analysis of Human metapneumovirus strains from Kenya and Zambia

□ EKamau_HMPV_WGS_Readme.txt
Plain Text - 4.5 KB - Aug 5, 2019 - 0 Downloads
MD5: 94e1f85ded6a0a8b4e99f460ba7de65f
Dataset readme file
Documentation
Download

□ Identity_graph_HMPVA_Ggene.csv
Comma Separated Values - 3.2 KB - Aug 5, 2019 - 0 Downloads
MD5: 85b9d82a093f56f425a618da56dbba64
Data

□ Identity_graph_HMPVA_SHgene.csv
Comma Separated Values - 2.4 KB - Aug 5, 2019 - 0 Downloads
MD5: 5eec8e812e0c9cdd1a81e7d31a7cf551
Data

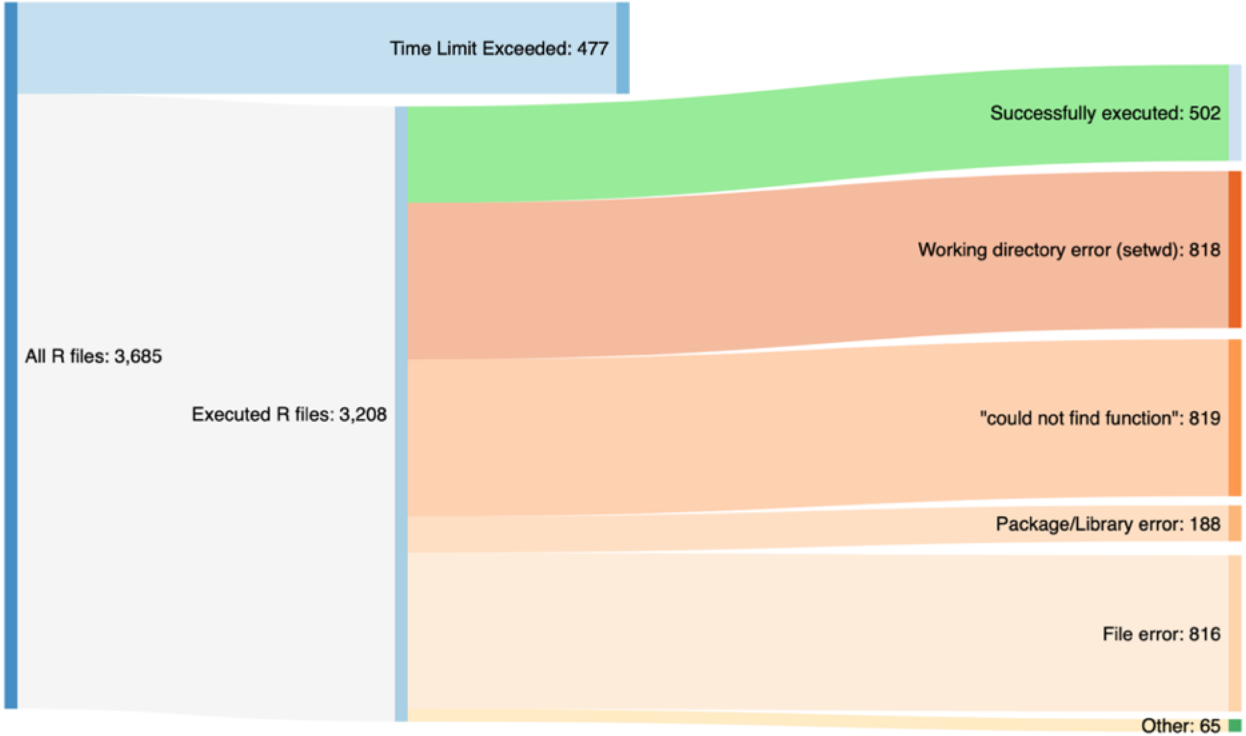
□ Identity_graph_HMPVB_Ggene.csv
Comma Separated Values - 3.6 KB - Aug 5, 2019 - 0 Downloads
MD5: 991131141a43d62276cd3083fd78a7d9
Data

□ Identity_graph_HMPVB_SHgene.csv
Comma Separated Values - 2.6 KB - Aug 5, 2019 - 0 Downloads
MD5: c8a3d807c5e88443678bcf3b68291802
Data
Download

□ script_2Jul2019.R
R Syntax - 3.0 KB - Aug 5, 2019 - 0 Downloads
MD5: 64531365d4f6caaeaf95549d170fdcdd
Replication code in R
Code
Download

@mercecrosas

Re-execution of R Code in published datasets



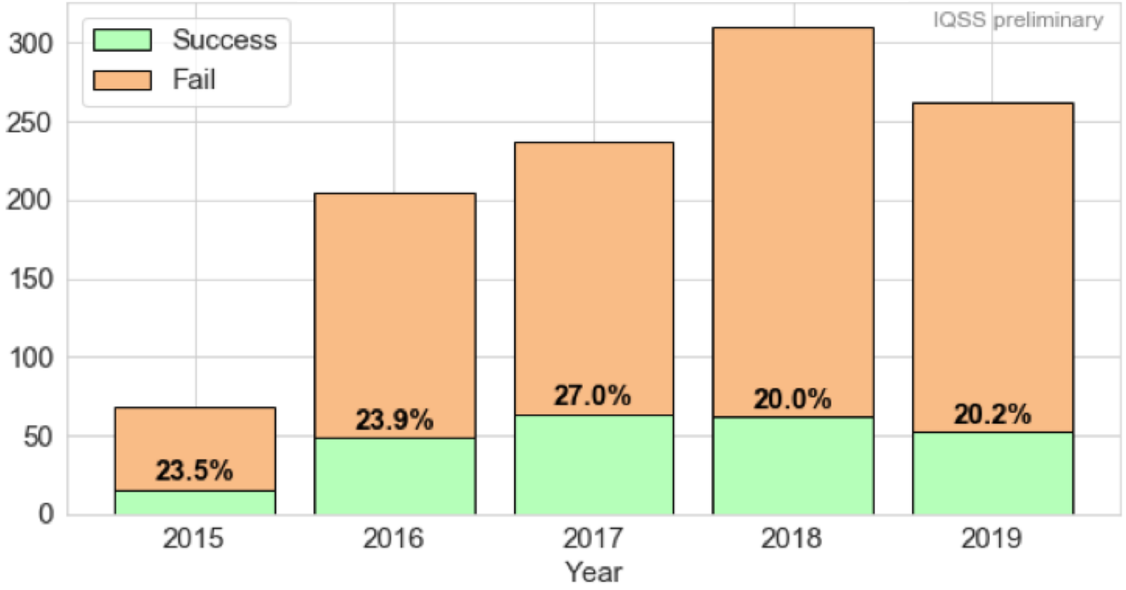
84.4% R code files fail to execute

77.5% of datasets with R files contain non-executable code

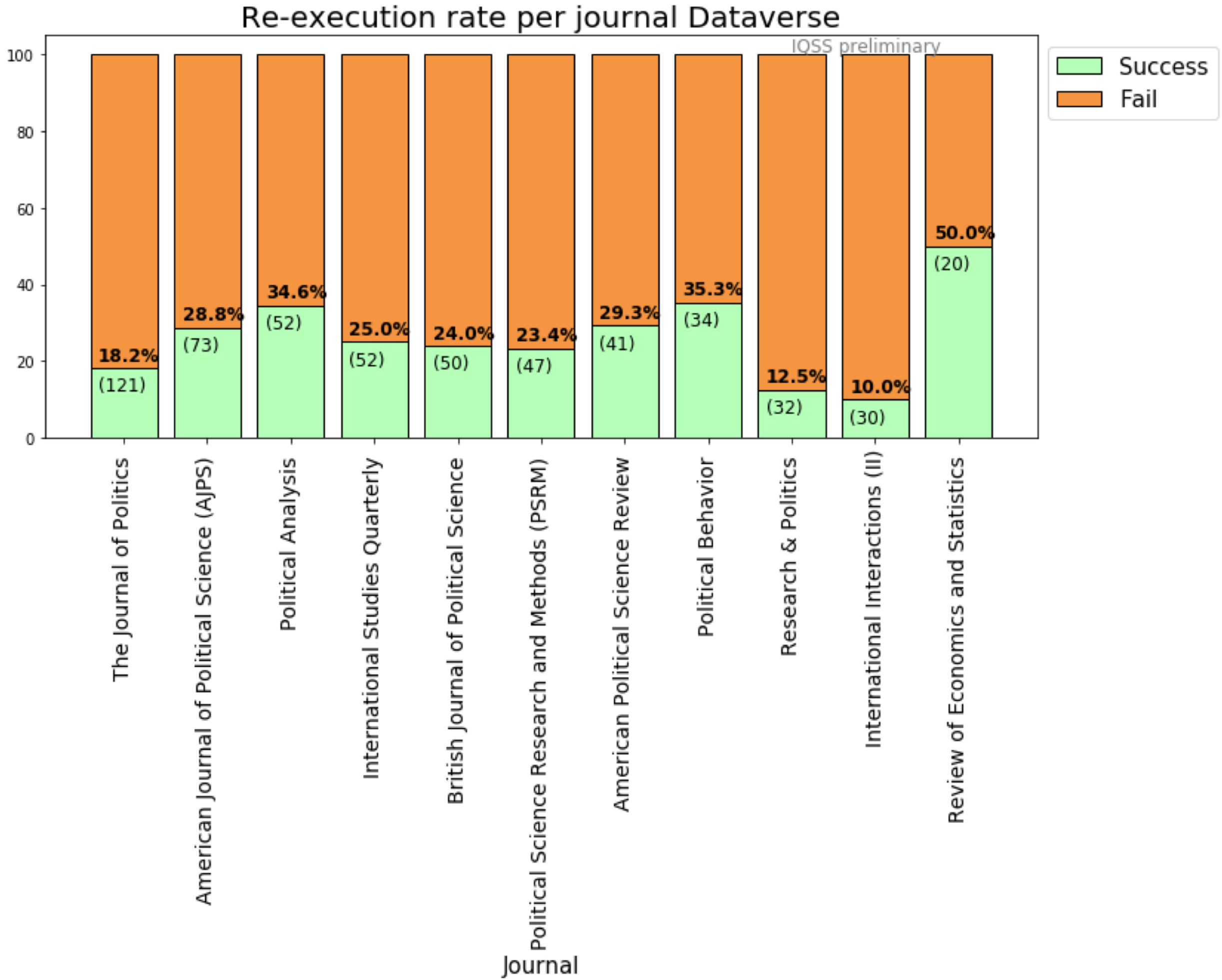
Source: Current study on code execution for datasets published in Dataverse by Ana Trisovic (Harvard’s IQSS) funded by Sloan Foundation

Re-execution of R code per year

Re-execution of R code in datasets published in Harvard Dataverse



Re-execution of R code in Journals Replication Data published in Harvard Dataverse



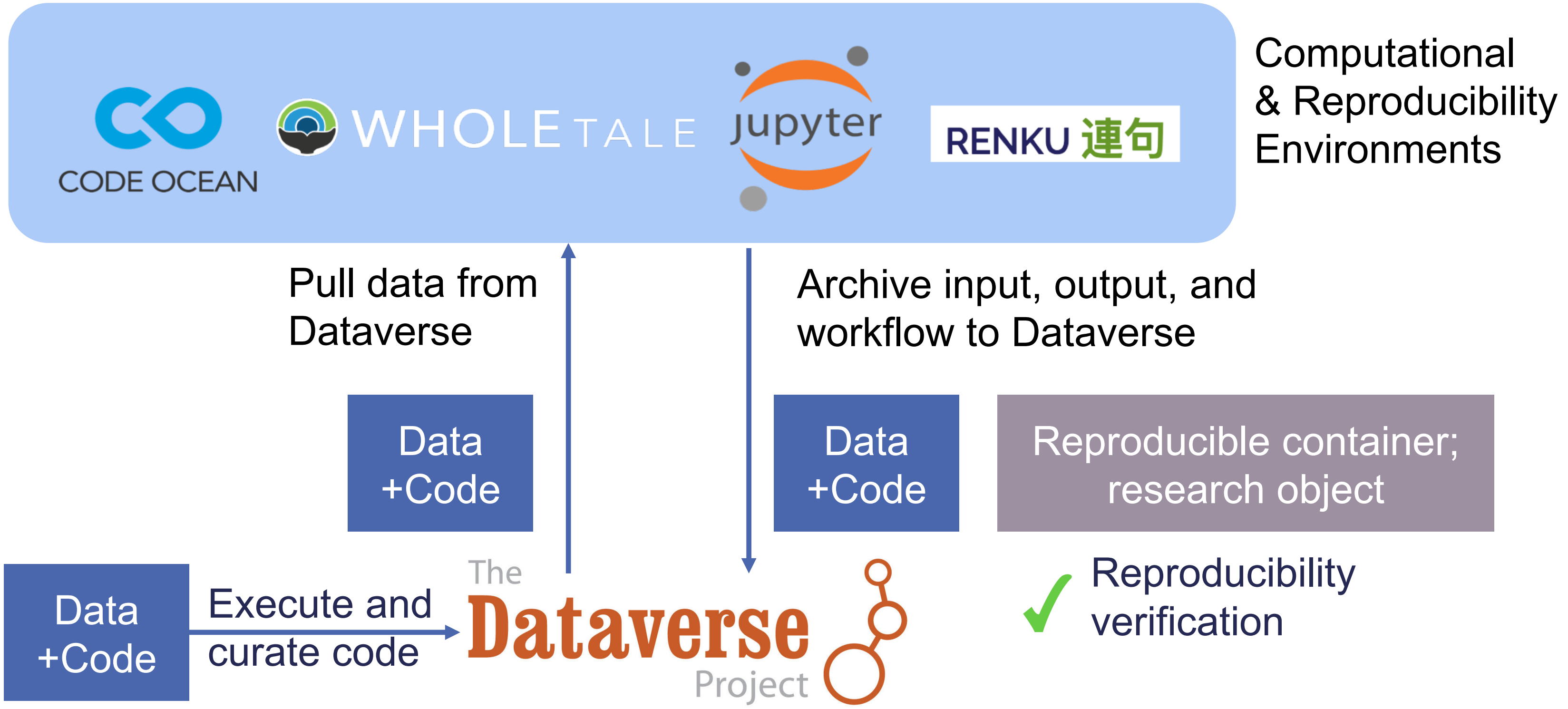
New Features: computational reproducibility

- Include **reproducibility as part of review** workflow
- Integrate Dataverse with computational tools to **facilitate code execution**
- Deposit a **capsule** (container with data and code) once verified for reproducibility
- When possible, **automate code execution** upon publishing the data and code

Also, in the process of evaluating integration with:

- **Data Curation Tools**
- **Research Objects**
- **Citation and metadata for software**

Integration with computational tools



Computational & Reproducibility Environments

Pull data from Dataverse

Archive input, output, and workflow to Dataverse

Data +Code

Data +Code

Reproducible container; research object

Data +Code

Execute and curate code

✓ Reproducibility verification

DataTags

Mercè Crosas, Latanya Sweeney, Michael Bar-Sinai

Dataverse Team

A **datatag** is a set of security features and access requirements for file handling.

A **datatags repository** is one that stores and shares data files in accordance with a standardized and ordered levels of security and access requirements

A DataTags repository must:

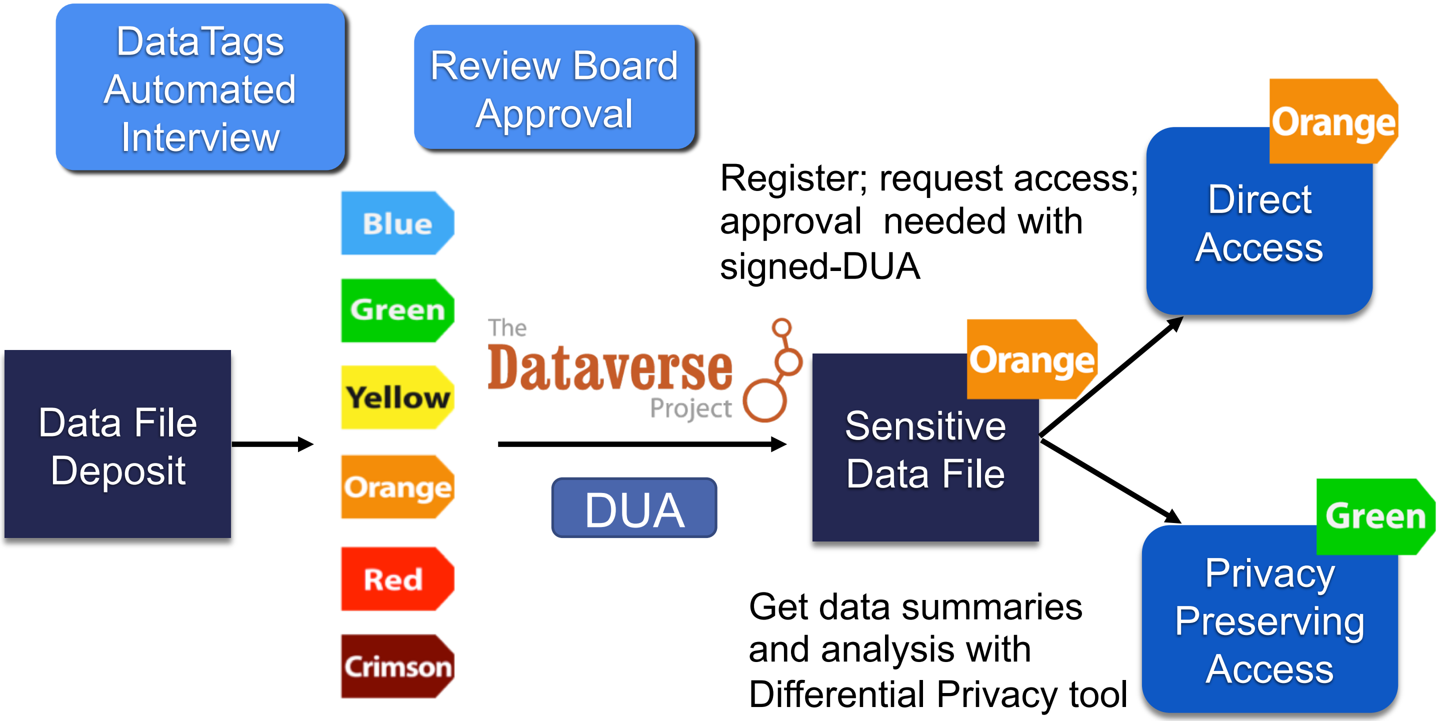
1. Support more than one **datatag**
2. Each file in the repository must have one and only one **datatag**
    a. additional requirements cannot weaken the file security
    b. and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the **repository** must:
    a. satisfy file's access requirements,
    b. produce sufficient credentials as requested,
    c. and agree to any terms of use required to acquire the file.
4. Provide **technological guarantees** for requirements 1, 2 and 3.

Standardized Access and Security Levels

DataTag	Access	Authorization	Data Use Agreement	Encryption
Blue	Public			
Green	Public	+ Register		
Yellow	Restricted	+ Approval Needed	+ Click-thru DUA	+ Encrypted transit
Orange	Restricted	+ Approval Needed	+ Signed DUA	+ Encrypted transit + Encrypted storage
Red	Restricted	+ Approval Needed	+ Signed DUA + Two-factor Auth	+ Encrypted transit + Encrypted storage
Crimson	Restricted	+ Approval Needed	+ Signed DUA + Two-factor Auth	+ Encrypted transit + Multi-encrypted storage

Sweeney, Crosas, Bar-Sinai, 2015. Sharing Sensitive Data with Confidence: The DataTags System, Technology Science

DataTags, Dataverse, Privacy-Preserving Tools



DataTags Automated Interview

Review Board Approval

Data File Deposit

Blue

Green

Yellow

Orange

Red

Crimson

The Dataverse Project

DUA

Orange

Sensitive Data File

Register; request access; approval needed with signed-DUA

Orange

Direct Access

Get data summaries and analysis with Differential Privacy tool

Green

Privacy Preserving Access

OpenDP

Salil Vadhan, James Honaker, Gary King, Mercè Crosas

Harvard Privacy Tools team

OpenDP: A New Project for Sensitive Data

A **community effort** to build a trustworthy and **open-source** suite of **differential privacy** tools that can be **easily adopted** by custodians of sensitive data to make it available for statistical research.

- To be launched in **2020** by Sloan Foundation funding
- Initially led by **Harvard Privacy Tools** project

*A tool (algorithm) is **differentially private** if its output cannot reveal whether any individual's data was included in the original dataset or not*.

OpenDP: Use Cases Motivation

- **Archival data repositories** (e.g., Dataverse) enabling secondary reuse and reproducibility
- **Government agencies** making data available to the public, both for official statistics and open data mandates
- **Companies** sharing data for academic research or internal research.
- **Focus on “centralized model” for Differential Privacy:**
    - A central aggregator has accessed to the raw data. The aggregator transforms the data with a differentially private mechanism.

OpenDP: Principles

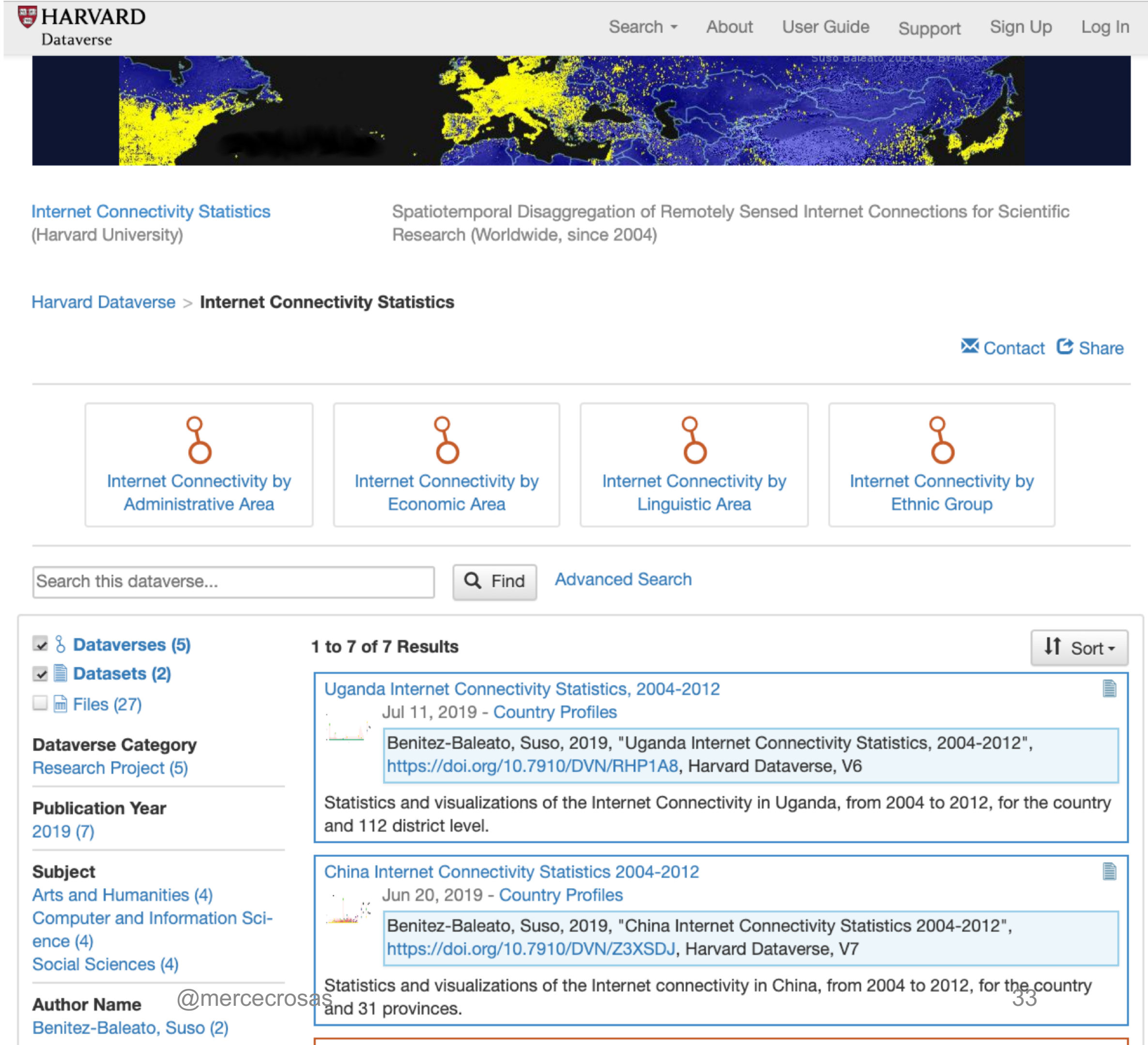
- **Open Source**
  - worldwide open-source community
  - processes and recognition for contribution
- **Security & Privacy**
  - careful vetting of any security-critical or privacy-critical code
  - can ship code to the sensitive data
- **Scalability**
  - handle petabyte-scale data
- **Extensibility**
  - can grow from the continuing research developments in the field

Example: Enabling Reproducibility with Privacy Preserving Data Sharing

Suso Baleato, James Honaker, Mercè Crosas

Internet Connectivity Statistics

- Data from measurements of the used **IPv4** address space across countries
- Close **correlation with internet penetration statistics** from International Telecommunications Union (ITU) and **OECD**



Internet Connectivity Statistics
(Harvard University)

Spatiotemporal Disaggregation of Remotely Sensed Internet Connections for Scientific Research (Worldwide, since 2004)

Harvard Dataverse > **Internet Connectivity Statistics**

Internet Connectivity by Administrative Area

Internet Connectivity by Economic Area

Internet Connectivity by Linguistic Area

Internet Connectivity by Ethnic Group

Search this dataverse...

☑ **Dataverses (5)**

☑ **Datasets (2)**

□ Files (27)

Dataverse Category
Research Project (5)

Publication Year
2019 (7)

Subject
Arts and Humanities (4)
Computer and Information Science (4)
Social Sciences (4)

Author Name
Benitez-Baleato, Suso (2)

1 to 7 of 7 Results

Uganda Internet Connectivity Statistics, 2004-2012
Jul 11, 2019 - Country Profiles

Benitez-Baleato, Suso, 2019, "Uganda Internet Connectivity Statistics, 2004-2012", https://doi.org/10.7910/DVN/RHP1A8, Harvard Dataverse, V6

Statistics and visualizations of the Internet Connectivity in Uganda, from 2004 to 2012, for the country and 112 district level.

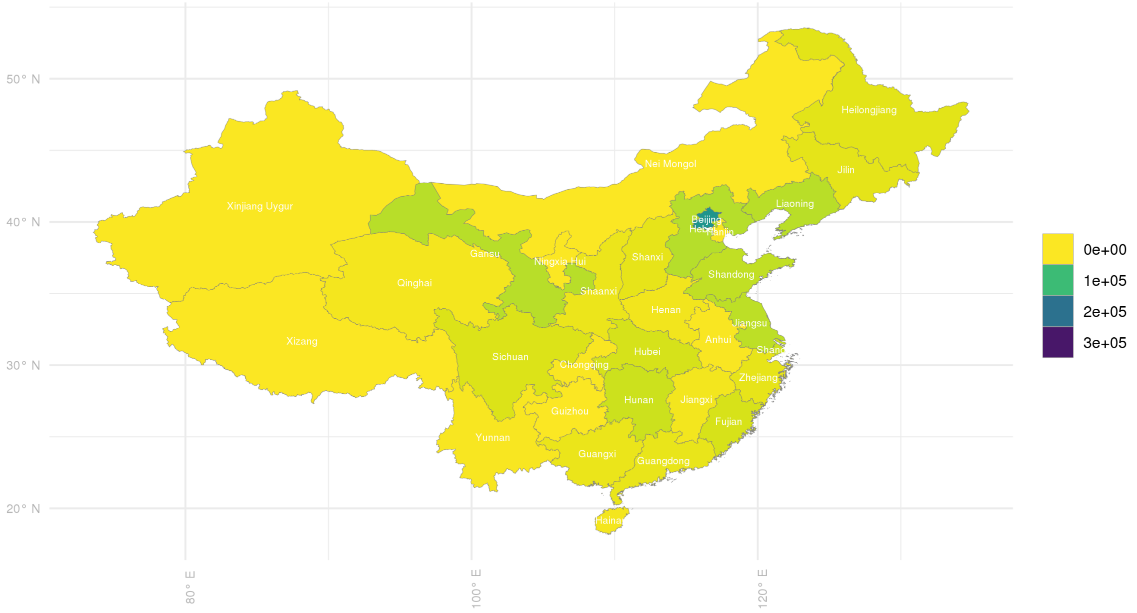
China Internet Connectivity Statistics 2004-2012
Jun 20, 2019 - Country Profiles

Benitez-Baleato, Suso, 2019, "China Internet Connectivity Statistics 2004-2012", https://doi.org/10.7910/DVN/Z3XSDJ, Harvard Dataverse, V7

Statistics and visualizations of the Internet connectivity in China, from 2004 to 2012, for the country and 31 provinces.

Differentially private data released in Dataverse

- Original data might contain **sensitive** information:
  - **DataTag = Orange**
- Statistics published in Dataverse are **differentially private**:
  - **DataTag = Green**
- Map with DP epsilon (noise) = 0.01, error added only 10^{-4}
- Error = 10^{-5} with Epsilon = 0.2, smaller epsilon, more privacy protection



Suso Baleato CC 3.0 BY-NC-SA <susobaleato@iq.harvard.edu>
Source & cite: https://doi.org/10.7910/DVN/Z3XSDJ Run: 20190322-11:01:50 (0.1-kaweco)

Conclusions

- Dataverse integrated with computational environments and tools is a solution for implementing Open Science, for a university, for a consortium, or nationally
- Dataverse with DataTags and OpenDP is a solution for sharing sensitive data
- A technology solution should include standards and incentives to share openly
- A technology solution should automatically assist in improving the quality of research outputs and tracking provenance
- But data and code authors must do their part to make their research outputs high-quality and reusable

Thanks

The Dataverse® Project

IQSS
The Institute for Quantitative Social Science

HARVARD
UNIVERSITY

GDCC The Global Dataverse Community Consortium
Supporting Dataverse repositories around the world.

Harvard University
Privacy Tools Project

dataverse.org | dataversecommunity.global/ | scholar.harvard.edu/mercecrosas