



Dataverse

A repository platform for sharing data and code

*Workshop of the OECD's Committee for Scientific and Technological Policy
Revision of Recommendation concerning access to research data from public funding
OECD Conference Center, Paris, October 15 2019*

Mercè Crosas, Ph.D.

Chief Data Science and Technology Officer, IQSS

Harvard University's Research Data Officer, HUIT

@mercecrosas

Dataverse – Achievements

Facilitating data sharing with technology, standards, and incentives

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.

Two Examples: Consortium Dataverse sites



- **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/>



- **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
- 800 datasets created in last 2 years
- <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

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

HARVARD Dataverse

Search - About User Guide Support Sign Up Log In

Journal of the Association of Environmental and Resource Economists Dataverse (University of Chicago Press) JAERE homepage

Harvard Dataverse > Journal of the Association of Environmental and Resource Economists Dataverse > Replication Data for: Demand for off-grid solar electricity – Experimental evidence from Rwanda

Contact Share

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

Version 1.0

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:e4XldMssTZduaSKrLW43wg== [fileUNF]

Cite Dataset - 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, Willingness to Pay, Electrification, Energy access

Files Metadata Terms Versions

Search this dataset... Find

Filter by File Type: All Access: All Sort

1 to 3 of 3 Files Download

<input type="checkbox"/>	 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)	Download
<input type="checkbox"/>	 Grimm et al (2019)_WTP Data(publish).tab Tabular Data - 567.5 KB - Oct 9, 2019 - 1 Download 502 Variables, 323 Observations - UNF:6:e4XldMssTZduaSKrLW43wg== Data (STATA)	Explore Download

Dataverse – Next

Improving research reproducibility and data reuse

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

documentation

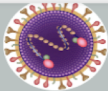
data

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

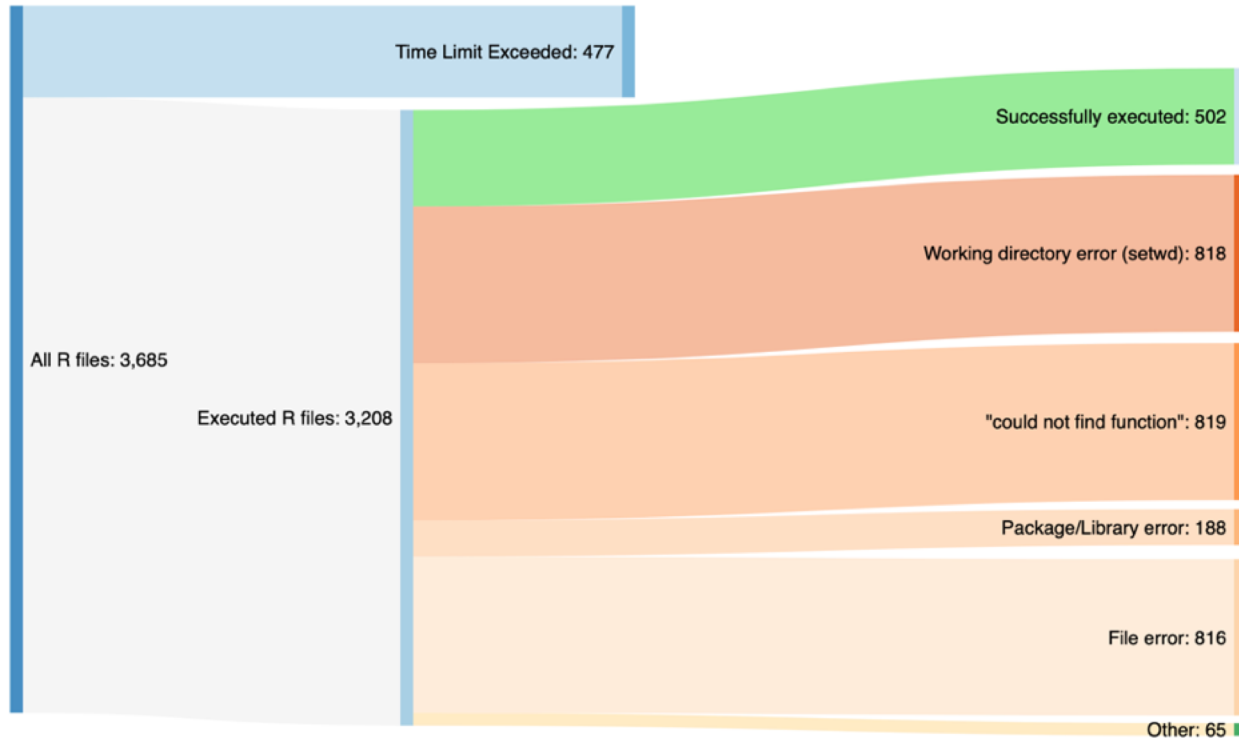
Harvard Dataverse > KWTRP Research Data Repository > Virus Epidemiology and Control (VEC) Dataverse > Replication Data for: Whole genome sequencing and phylogenetic analysis of Human metapneumovirus strains from Kenya and Zambia

<input type="checkbox"/>	 EKamau_HMPV_WGS_Readme.txt Plain Text - 4.5 KB - Aug 5, 2019 - 0 Downloads MD5: 94e1f85ded6a0a8b4e99f460ba7de65f Dataset readme file Documentation	<input type="button" value="Download"/>
<input type="checkbox"/>	 Identity_graph_HMPVA_Ggene.csv Comma Separated Values - 3.2 KB - Aug 5, 2019 - 0 Downloads MD5: 85b9d82a093f56f425a618da56dbba64 Data	
<input type="checkbox"/>	 Identity_graph_HMPVA_SHgene.csv Comma Separated Values - 2.4 KB - Aug 5, 2019 - 0 Downloads MD5: 5eec8e812e0c9cdd1a81e7d31a7cf551 Data	
<input type="checkbox"/>	 Identity_graph_HMPVB_Ggene.csv Comma Separated Values - 3.6 KB - Aug 5, 2019 - 0 Downloads MD5: 991131141a43d62276cd3083fd78a7d9 Data	
<input type="checkbox"/>	 Identity_graph_HMPVB_SHgene.csv Comma Separated Values - 2.6 KB - Aug 5, 2019 - 0 Downloads MD5: c8a3d807c5e88443678bcf3b68291802 Data	<input type="button" value="Download"/>
<input type="checkbox"/>	 script_2Jul2019.R R Syntax - 3.0 KB - Aug 5, 2019 - 0 Downloads MD5: 64531365d4f6caaeaf95549d170fdccd Replication code in R Code @mercecosas	<input type="button" value="Download"/>



A diagram illustrating the components of a dataset. A central cylinder labeled 'Dataset' has an orange arrow pointing to it from a box labeled 'Data files'. Another orange arrow points from the 'Dataset' to a box labeled 'Documentation'. A third orange arrow points from the 'Dataset' to a box labeled 'Code'. A blue curved arrow points from a box labeled 'Descriptive Metadata' back to the 'Dataset'.

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

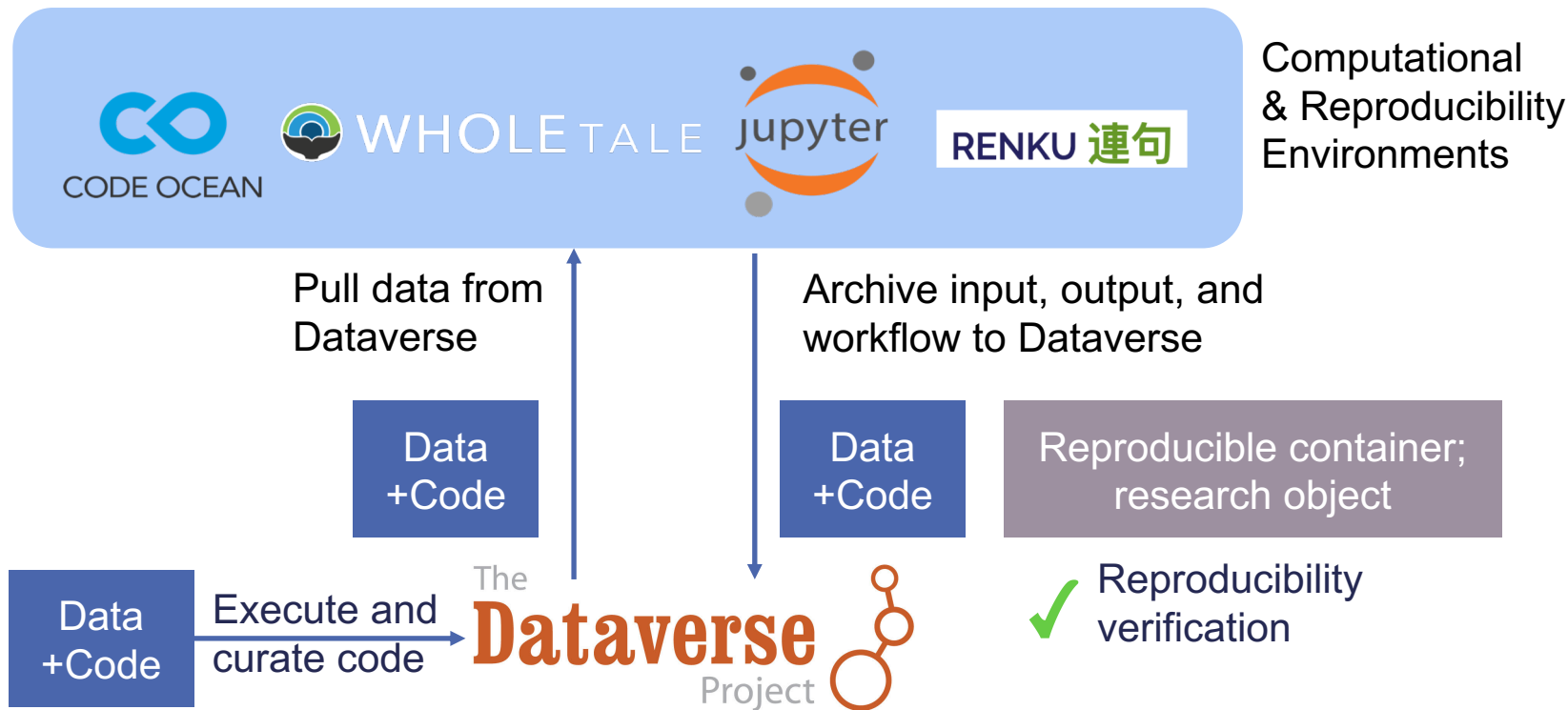
New Features: computational reproducibility

- 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



Thanks



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

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