

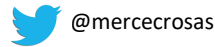
New Reproducibility Workflows with Dataverse:

A path for social science journals to increase
transparency and rigor in research

Mercè Crosas, Ph.D.

Chief Data Science and Technology Officer, IQSS

Harvard University's Research Data Officer, HUIT



“Wishlists and Workflows: Integrating Research Transparency into Editorial and Publishing Processes”, Data-PASS Pre-APSA workshop, Washington, D.C. , August 28

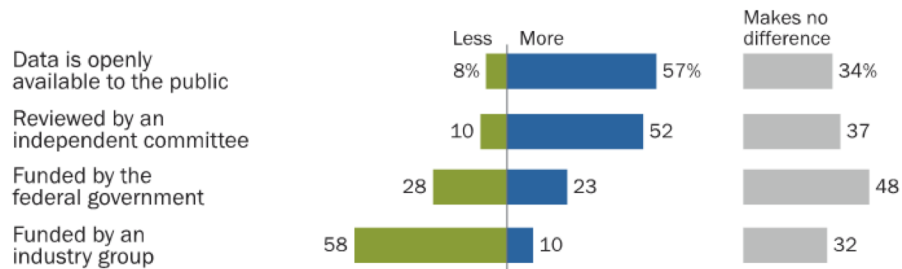
"Americans say open access to data and independent review inspire more trust in research findings"

Trust and Mistrust of American Views on Scientific Experts.

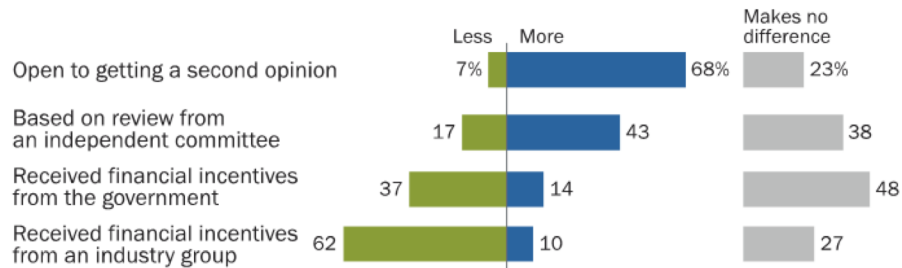
Pew Research Center, August 2, 2019

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

A path for social science journals to increase transparency and rigor in research

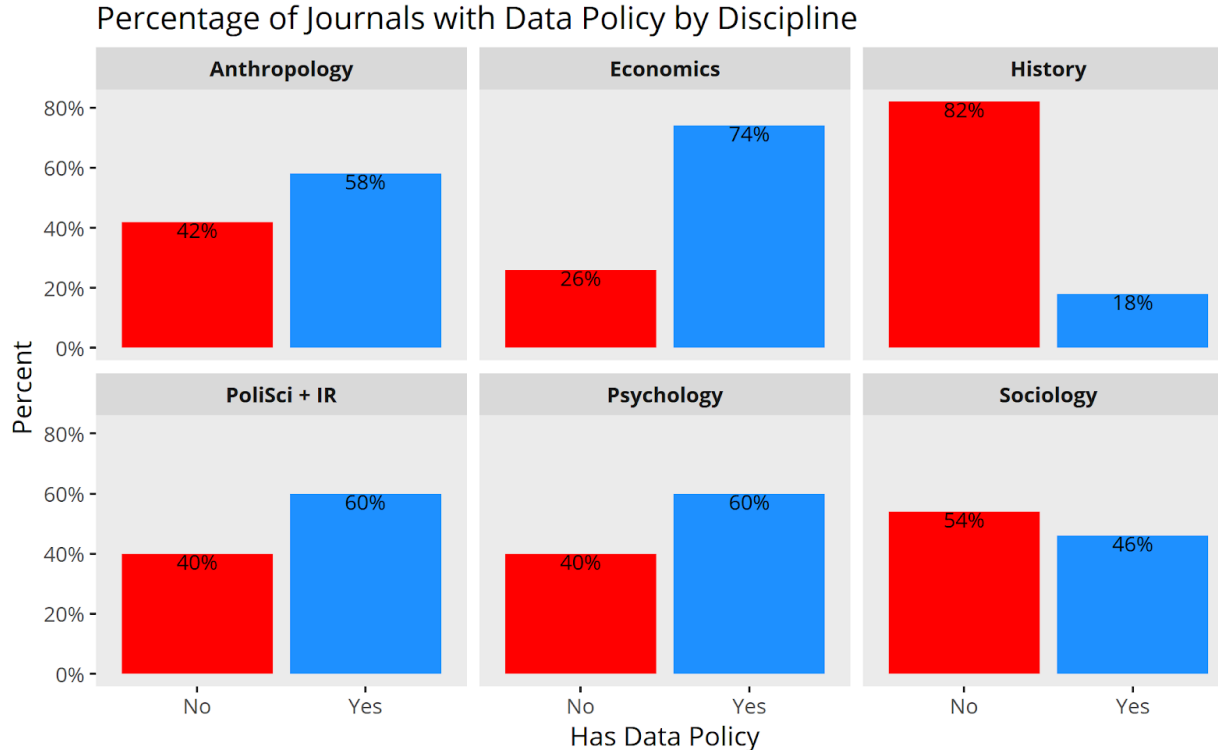
1. **The current** landscape of journal data sharing policies
2. Is data sharing sufficient?
3. New support for computational reproducibility
4. Is computational reproducibility sufficient?

What fraction of social science journals have data sharing policies? Does it vary by discipline?

“we review the data policies of the 50 most influential international peer-reviewed journals according to the Clarivate Analytics (formerly Thomson Reuters) Journal Impact Factor in the disciplines of political science and international relations, economics, sociology, history, psychology, and anthropology.”

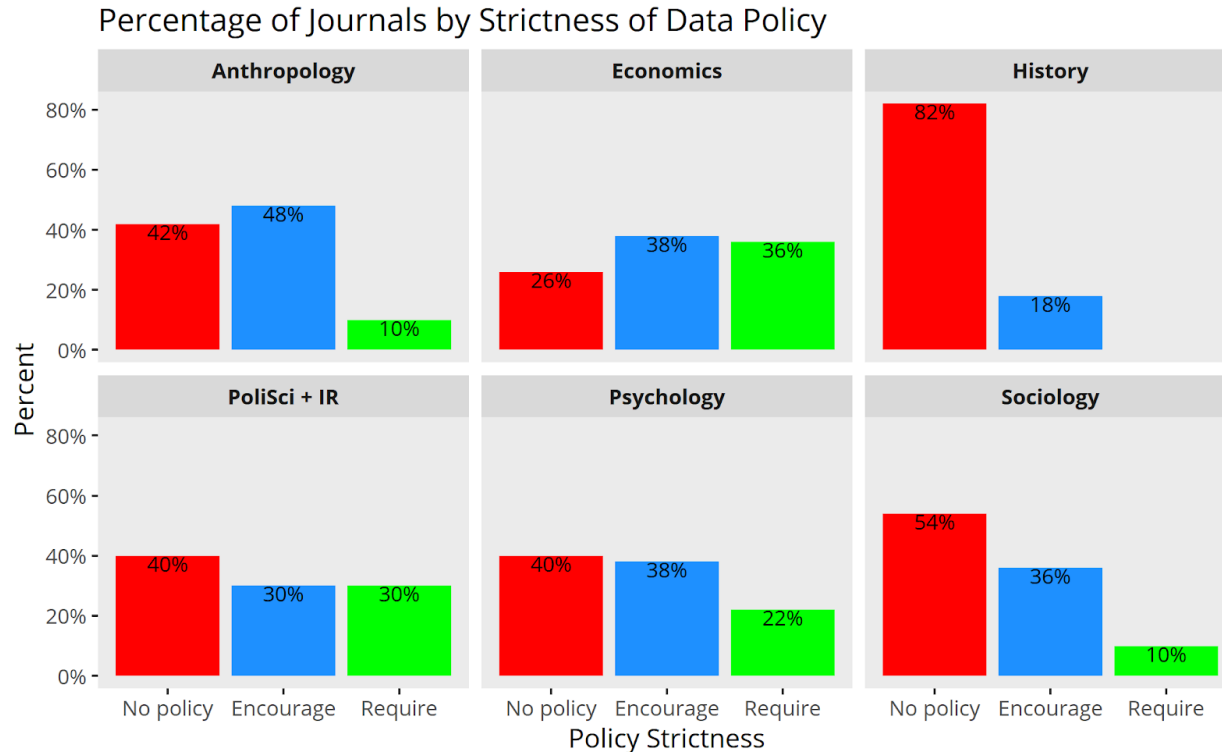
Crosas, Gautier, Karcher, Kirilova, Otalora, Schwartz. Data Policies of Highly-Ranked Social Science Journals, *preprint*, <https://osf.io/preprints/socarxiv/9h7ay>

Half of all journals in our study have a data policy.
For History, only 18 % have a data policy.

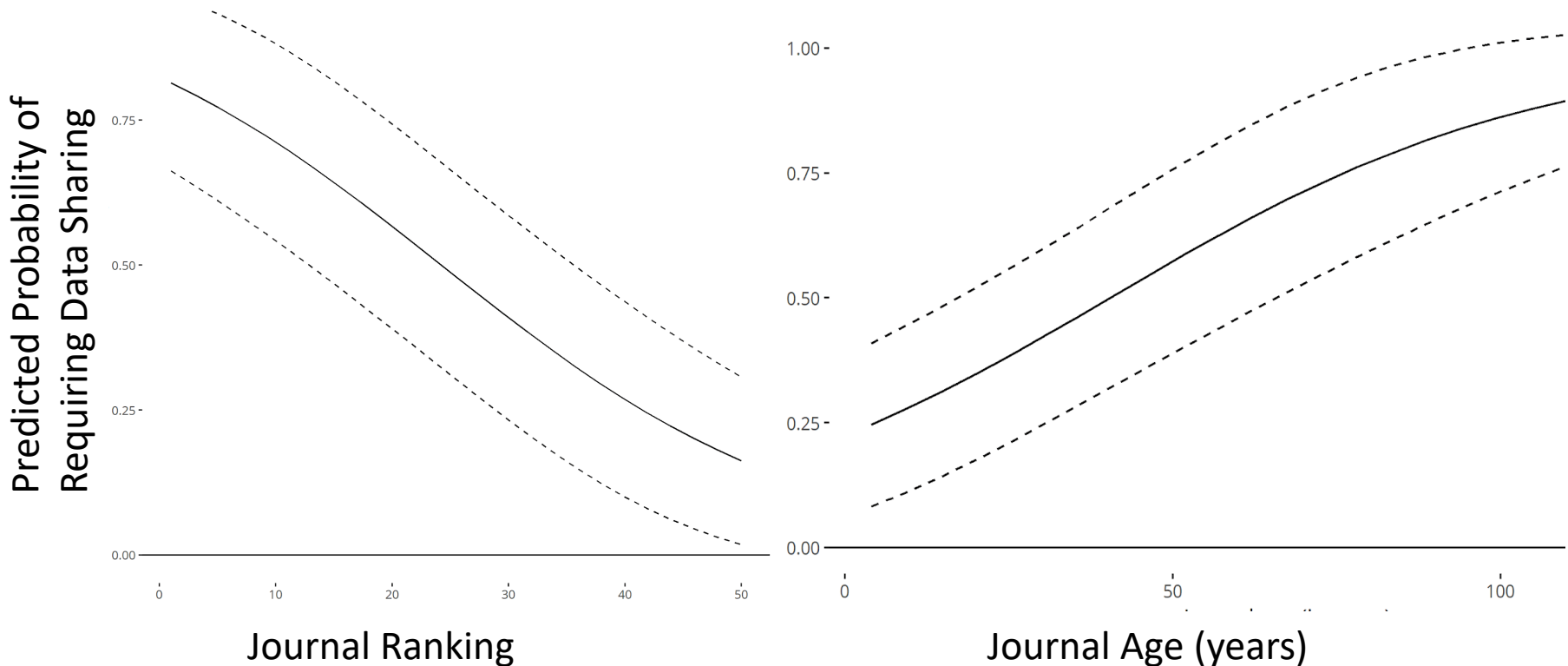


155 of the total 291 unique journals have some sort of data policy

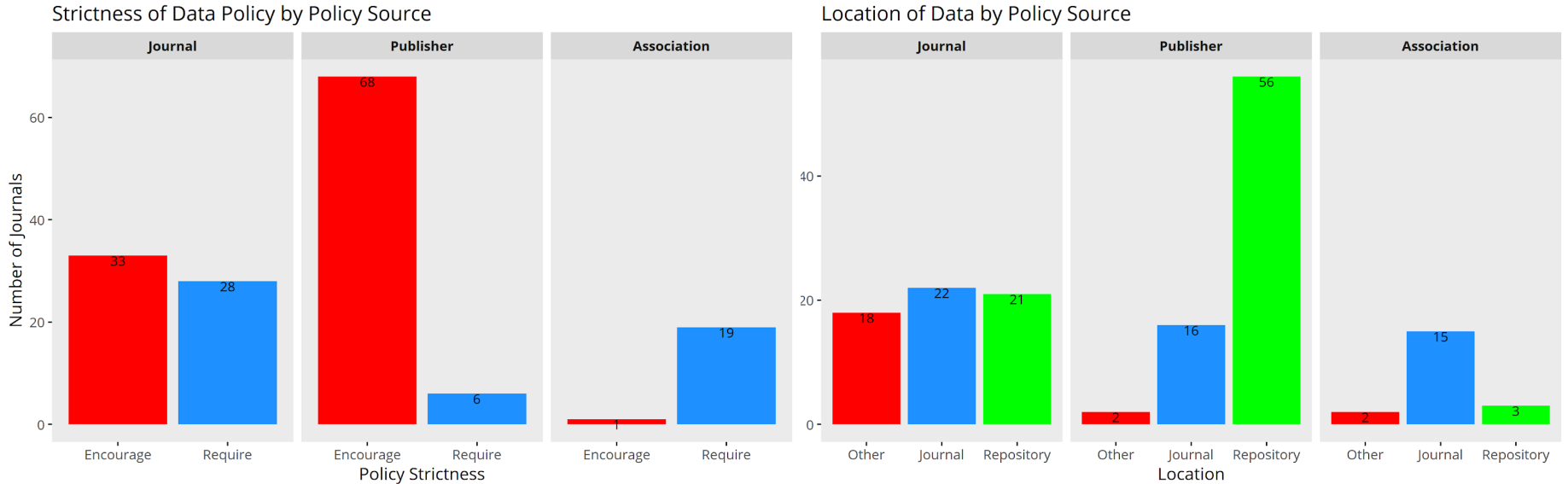
Requiring data sharing is more prominent in Economics and Political Science.



Requiring data sharing is more likely with higher *Rank* and *Age* of the journal.



Policy source impacts data sharing practice.



Policy language from Publishers tends to encourage data sharing in a repository.

Policy language from Associations tends to require data sharing in supplementary materials.

Policy language from journals themselves varies in requirements and recommendations.

[My] Recommendations for Journal Data Policies

- Having any **data policy** is better than no policy at all
- If possible, **require**, not just encourage
- Recommend **data repositories** (community-specific, general purpose)
- Ensure formal **citation** from article to data and from data to article
- Use **clear language** with clear guidance for authors

Dataverse: a Solution for Journal Data Sharing

- A **data citation** with a persistent identifier (DOI)
- Standard **metadata**, plus custom metadata for journals
- **Tiered access** to data as needed:
 - Fully Open, CC0
 - Register to access; Guestbook
 - Restricted with DUA
- **Anonymous** dataset review
- Multiple **versions** of a dataset
- **Branding and customization** for a journal dataverse
- **FAIR principles** support (Findable, Accessible, Interoperable, Reusable data)

Deposit and share your data. Get academic credit.

Harvard Dataverse is a digital repository where you can 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.

90,252 datasets, 161,940 downloads, 507,000 files, 8 million downloads

Add a dataset

Add a dataverse +

• 84 journal dataverses: 5,000 datasets with 50,000 files, 1 million downloads

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

Search over 90,200 datasets...

Find

Browse by subject

Agricultural Sciences 1,264

Computer and Information Science 931

Medicine, Health and Life Sciences 3,011

Arts and Humanities 604

Earth and Environmental Sciences 1,800

Physics 225

Astronomy and Astrophysics 516

Engineering 416

Social Sciences 38,674

Business and Management 428

Law 276

Chemistry 184

Mathematical Sciences 209

+ 45 other Dataverse repositories across 6 continents, including ODUM Dataverse and QDR

A path for social science journals to increase transparency and rigor in research

1. The current landscape of journal data sharing policies
2. Is data sharing sufficient?
3. New support for computational reproducibility
4. Is computational reproducibility sufficient?

8,000 of the 90,000

datasets in Harvard

Dataverse contain the

files to reproduce the

publish results

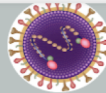
documentation

data













code

HARVARD
Dataverse

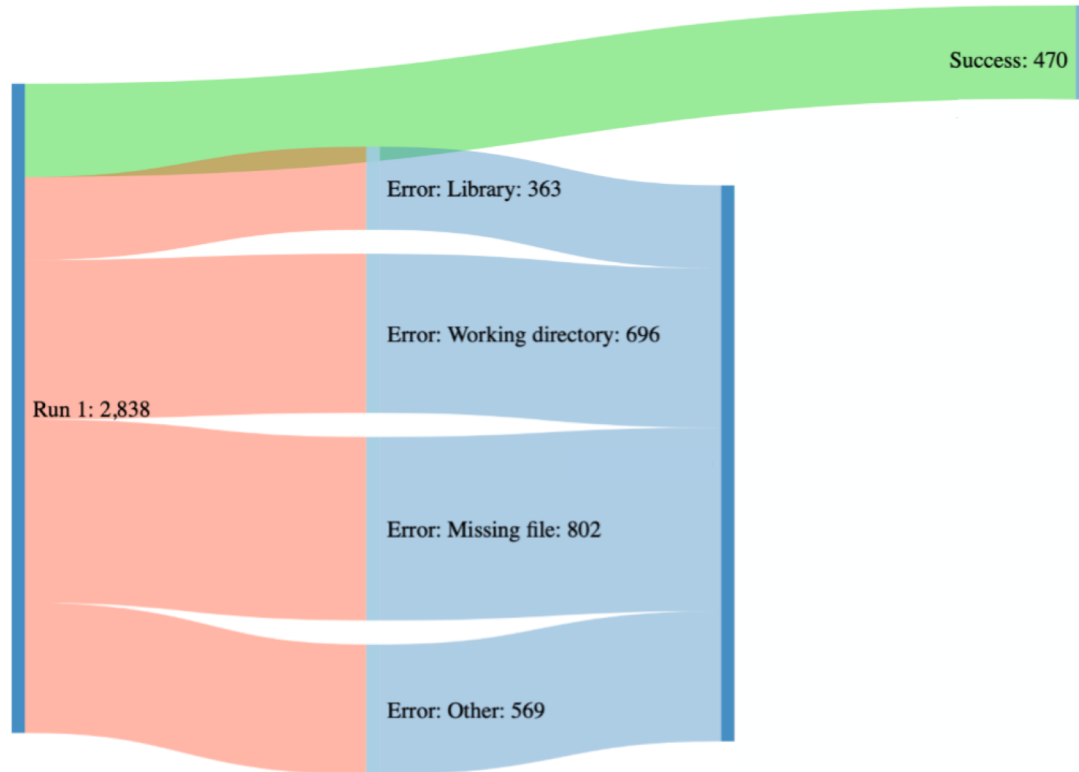
Search - About User Guide Support Sign Up Log In

 Virus Epidemiology and Control (VEC) Dataverse (Kemri Wellcome Trust Research Programme, Kilifi, Kenya) Population dynamics of viral pathogens informing intervention strategies

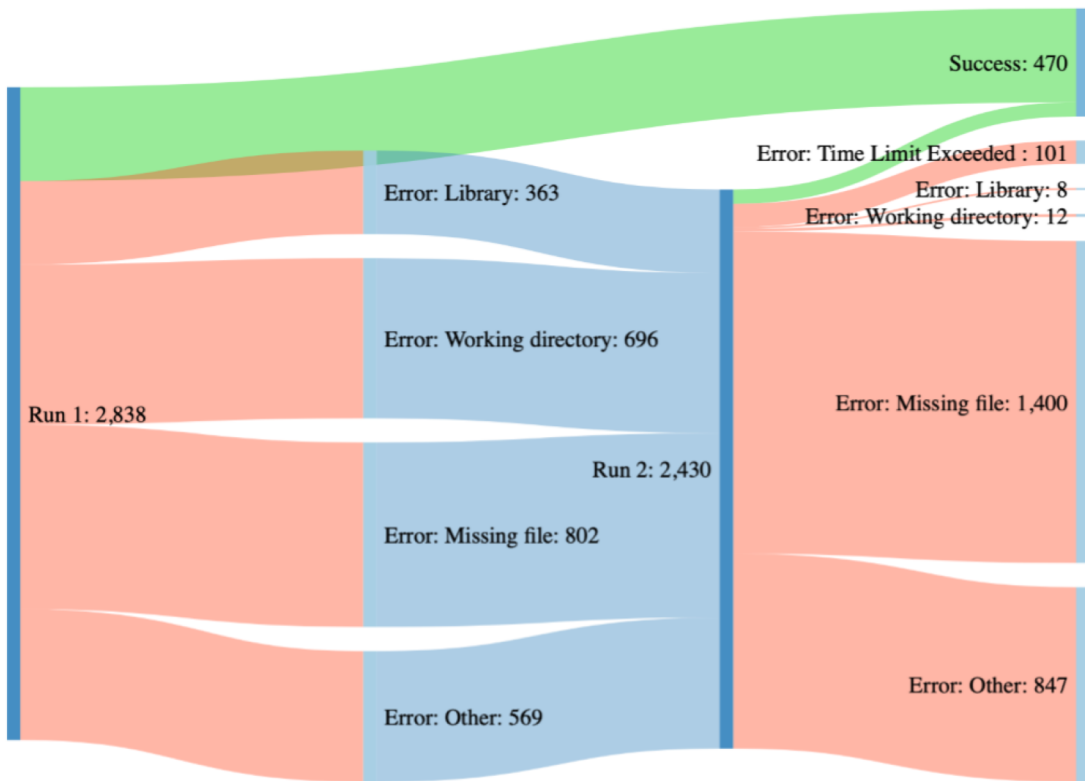
Harvard Dataverse > KWTRP Research Data Repository > Virus Epidemiology and Control (VEC) Dataverse >
Replication Data for: Whole genome sequencing and phylogenetic analysis of Human metapneumovirus strains from Kenya and Zambia

	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	 Download
	Identity_graph_HMPVA_SHgene.csv Comma Separated Values - 2.4 KB - Aug 5, 2019 - 0 Downloads MD5: 5eec8e812e0c9cdd1a81e7d31a7cf551 Data	 Download
	Identity_graph_HMPVB_Ggene.csv Comma Separated Values - 3.6 KB - Aug 5, 2019 - 0 Downloads MD5: 991131141a43d62276cd3083fd78a7d9 Data	 Download
	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: 64531365d4f6caaeaf95549d170fdccd Replication code in R Code	 Download

85.6% of archived R-based studies are not easily re-executable



85.6% of archived R-based studies are not easily re-executable



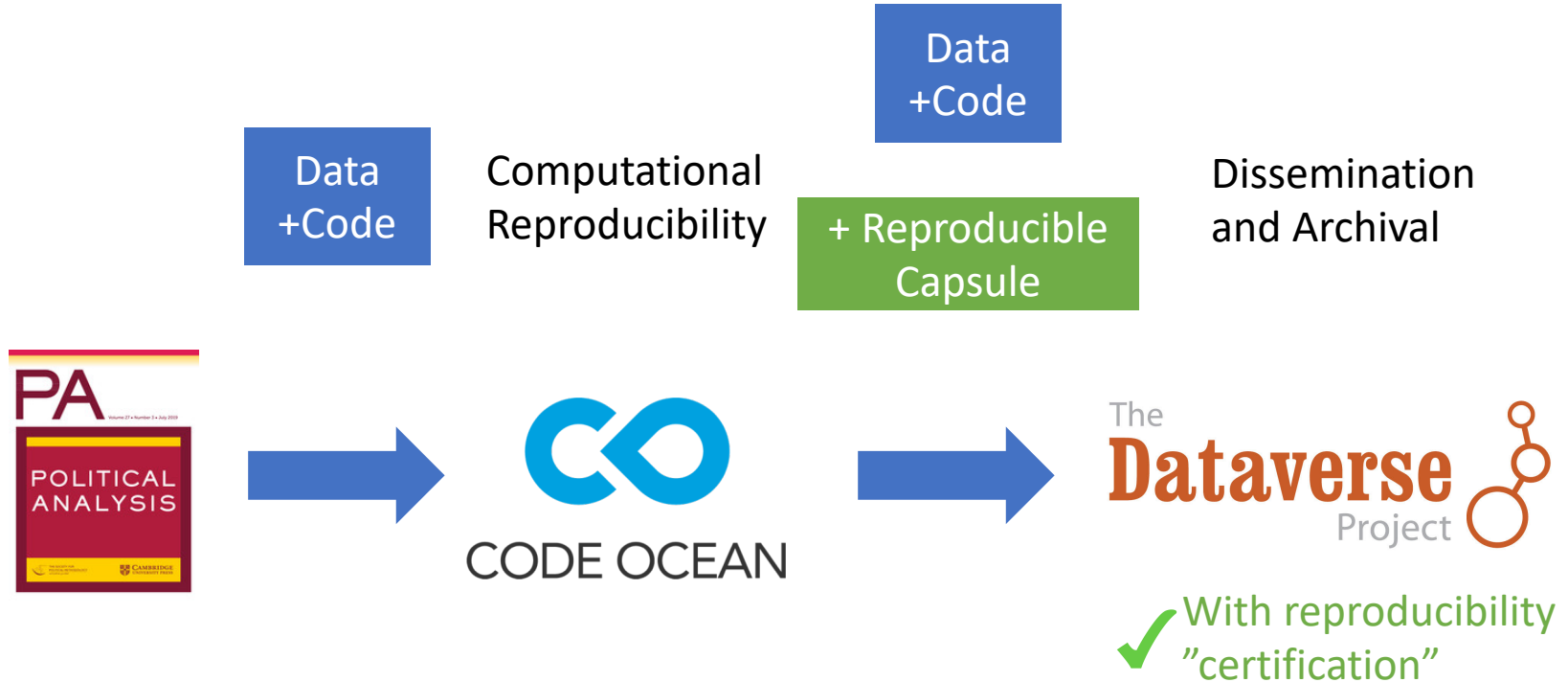
A path for social science journals to increase transparency and rigor in research

1. The current landscape of journal data sharing policies
2. Is data sharing sufficient?
3. **New** support for computational reproducibility
4. Is computational reproducibility sufficient?

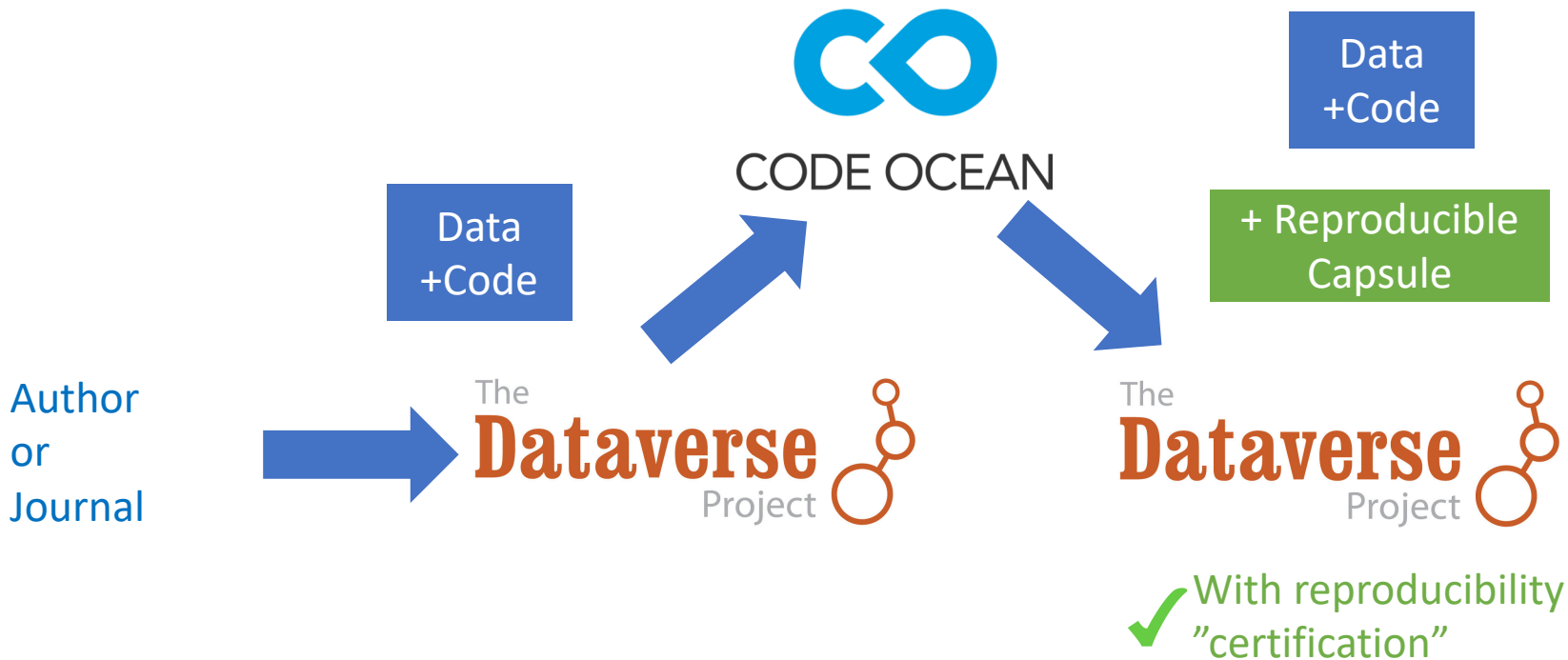
Current Dataverse projects to improve computational reproducibility

- Include [reproducibility as part of peer review](#) workflow [[ODUM as a third-party for reproducibility verification](#)]
- Integrate Dataverse with reproducibility and computational web-based tools (e.g., Code Ocean) to [facilitate code execution](#) [[under development](#)]
- Deposit a [capsule](#) (container with data and code) that has been verified for reproducibility [[under development](#)]
- When possible, [automate code execution](#) upon publishing the data and code [[research project](#)]

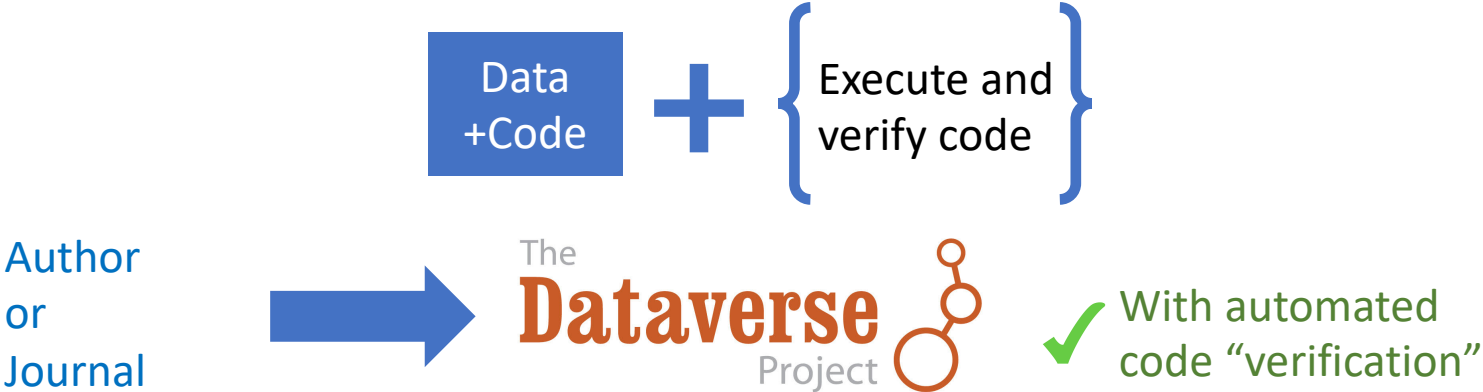
Workflow 1: From journal to Code Ocean, to Dataverse [under development]



Workflow 2: From journal to Dataverse, to Code Ocean, and back to Dataverse [under development]



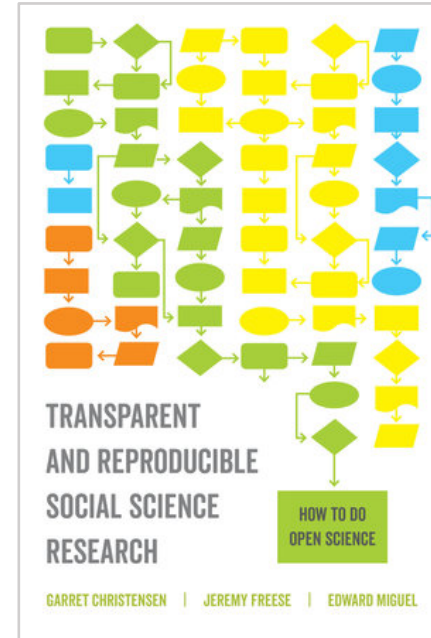
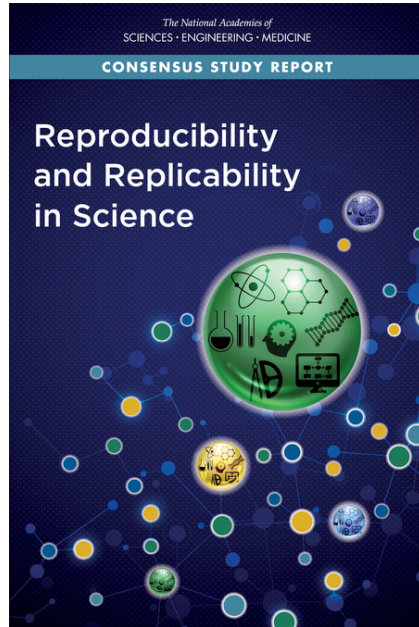
Workflow 3: From journal to Dataverse, verifying code automatically [research project]



A path for social science journals to increase transparency and rigor in research

1. The current landscape of journal data sharing policies
2. Is data sharing sufficient?
3. New support for computational reproducibility
4. Is computational reproducibility sufficient?

A broader context is essential.



NASEM Consensus Study Report on Reproducibility and Replicability in Science, 2019;
Christinsen, Freese, Miguel. Transparent and Reproducible Social Science Research, 2019

“Concerns about reproducibility and replicability have been expressed in both scientific and popular media. As these concerns came to light, **Congress requested that the National Academies of Sciences, Engineering, and Medicine conduct a study** to assess the extent of issues related to reproducibility and replicability and to offer recommendations for improving rigor and transparency in scientific research.”

[NASEM Consensus Study Report Highlights, Reproducibility and Replicability in Science](#)

Beyond Reproducibility, there is Replicability

- **Reproducibility:** equal to **computational reproducibility**—obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis.
- **Replicability:** obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

NASEM Report Highlights

- **No crisis**, but we must do better
- Promote use of open source tools
- ✓ Facilitate transparent sharing and availability of digital artifacts, such as data and code
- ✓ Journals should consider ways to ensure computational reproducibility during peer review

Additional Considerations for Transparency and Rigor

- Include a **clear, specific, and complete description** of how results are reached:
 - all methods, instruments, materials, procedures;
 - decisions for the exclusion or inclusion of data;
 - the analytic decisions and when these decisions were;
 - a discussion of the expected constraints on generality
 - reporting of precision or statistical power; and
 - discussion of the uncertainty of the measurements, results, and inferences;
- Be mindful of **publication bias** and **specification searching**
- Consider **meta-analysis**

<http://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm>

Christinsen, Freese, Miguel, 2019, Transparent and Reproducible Social Science Research

A path for social science journals to increase transparency and rigor in research

1. The current landscape of journal data sharing policies
2. Is data sharing sufficient?
3. New support for computational reproducibility
4. Is computational reproducibility sufficient?

Thank you

@mercecrosas