METASCIENCE 2021 CONFERENCE

Evidence-based steps toward a culture for replicability and reproducibility

Ana Trisovic

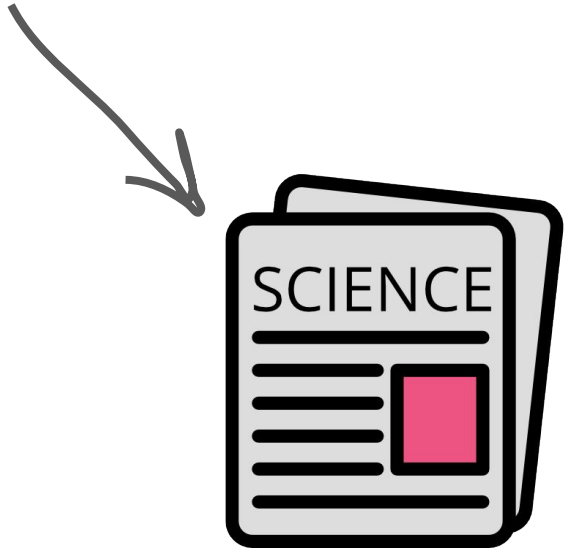
Harvard University

COS CENTER FOR OPEN SCIENCE

RoRI RESEARCH ON RESEARCH INSTITUTE

aimos

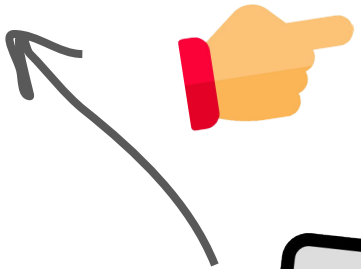
Researchers

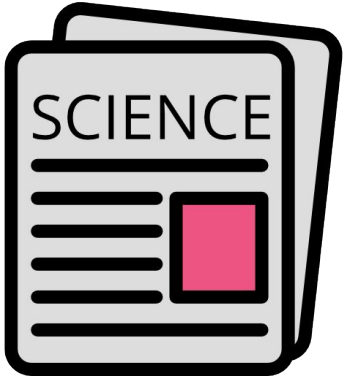


Journal

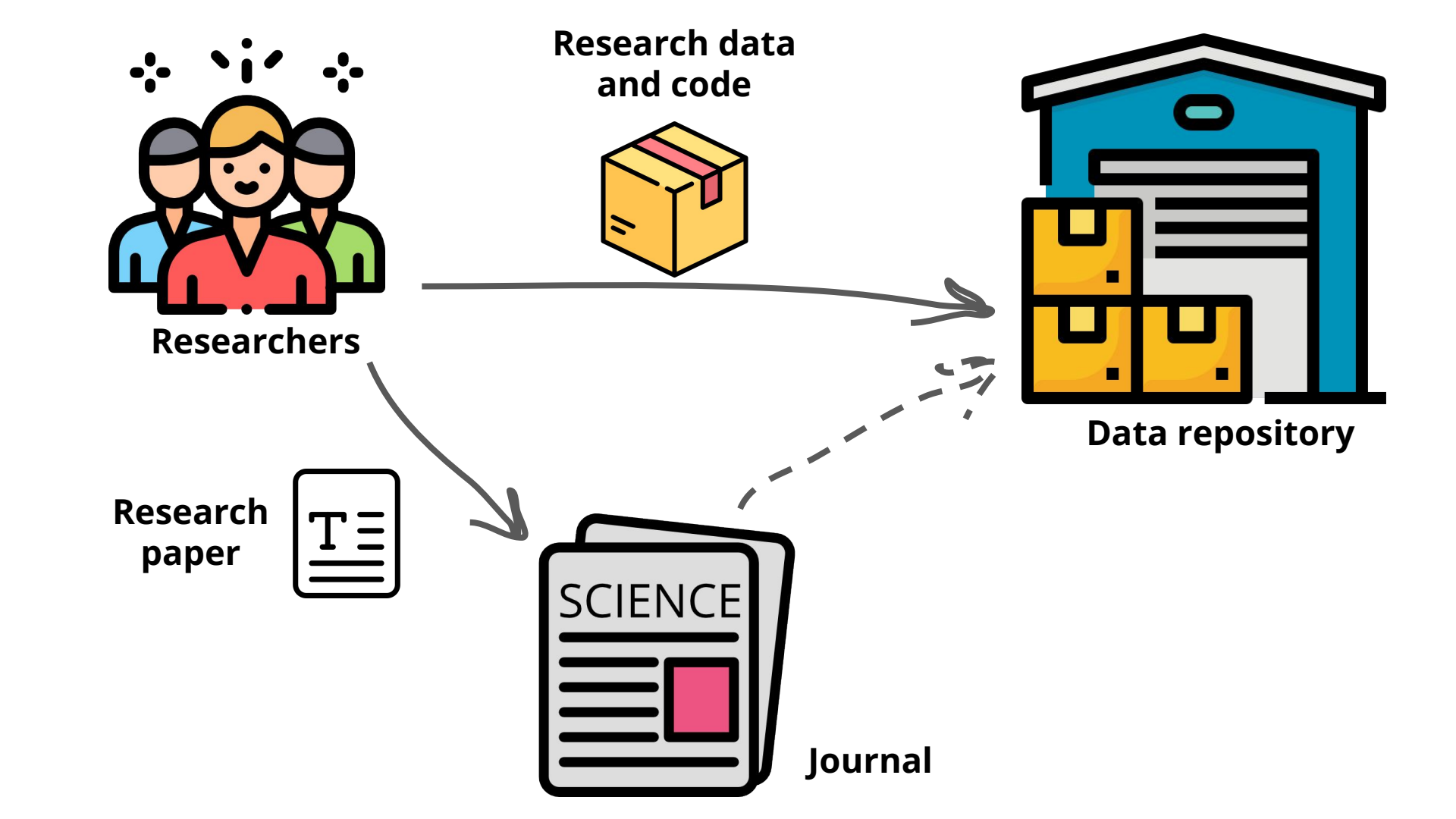
Researchers

Data repository





Journal


Research data
and code
Researchers
Data repository
Research
paper
SCIENCE
Journal

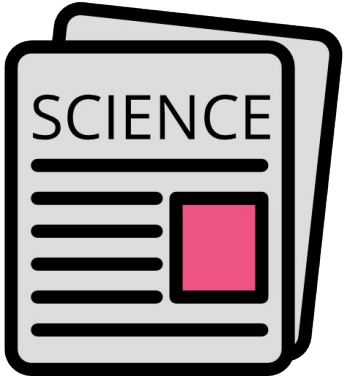
Presentation agenda

1. Dataverse research data repository
2. A large-scale study on code quality and execution
3. Results and discussion
4. What can researchers do
5. What can repositories do
6. What can journals do

Who are we?

Researchers

Data repository



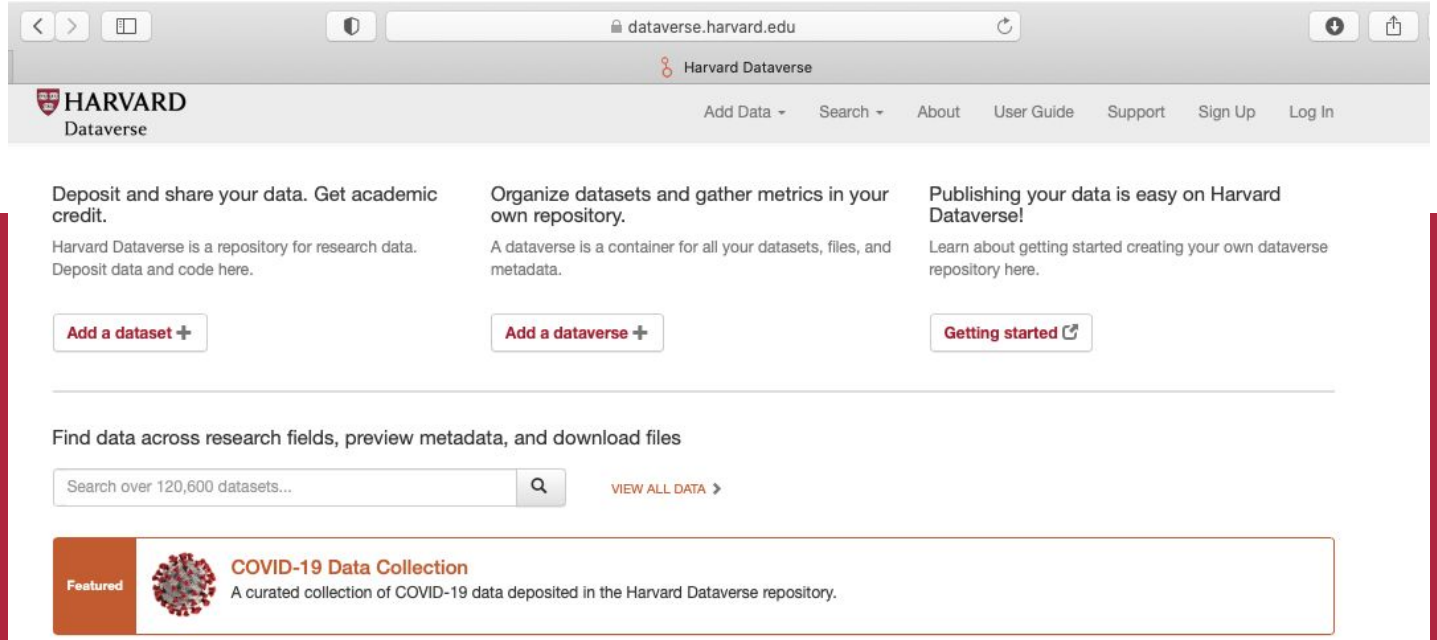
Journal

- A free and open-source software platform to archive, share, and cite research data
  - Focus on data sharing and making data available
- Provides data repository software that can be installed at institutions
  - Supports research communities for entire countries (NO, NL)

70 institutions around the globe run Dataverse installations as their official data repository

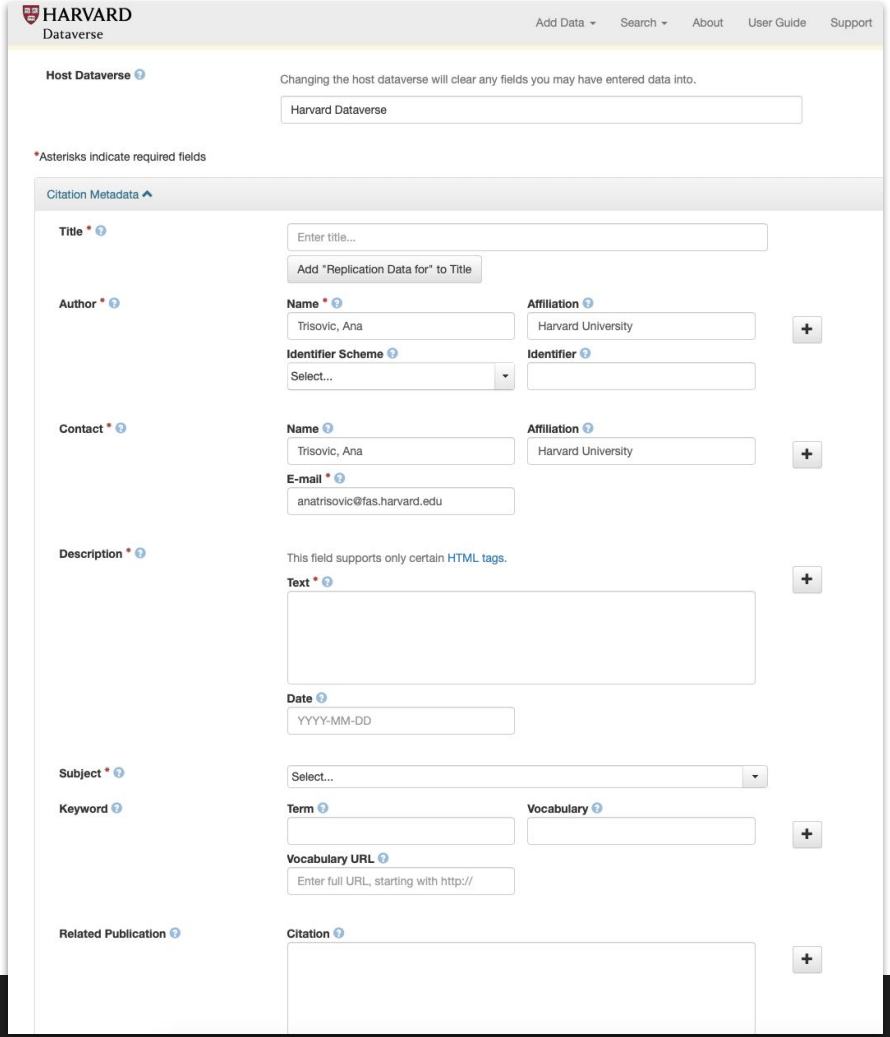
Harvard Dataverse

https://dataverse.harvard.edu



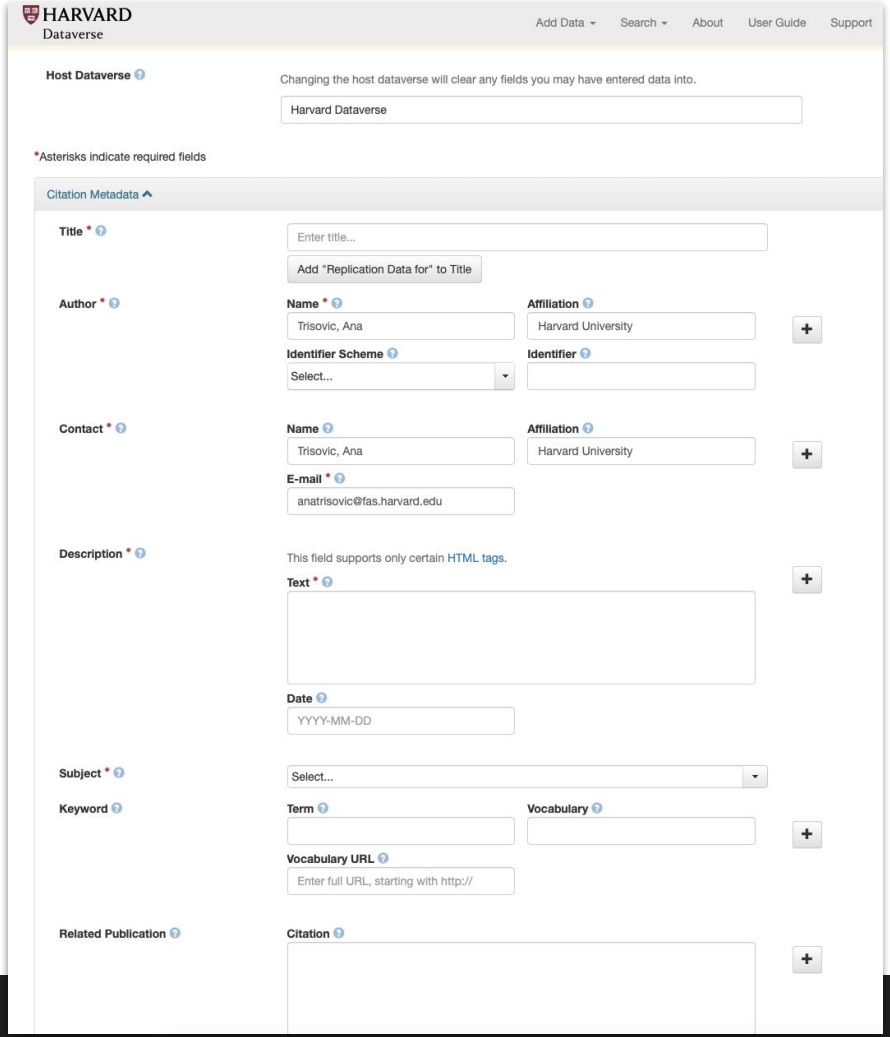
Data sharing

- Stand-alone or institutional account for depositing data (Dataverse, GitHub, Google, ORCID, University credentials)
- Individuals, institutes or journals may have own dataverse collections



Data sharing

- Stand-alone or institutional account for depositing data (Dataverse, GitHub, Google, ORCID, University credentials)
- Individuals, institutes or journals may have own dataverse collections



Replication dataset

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

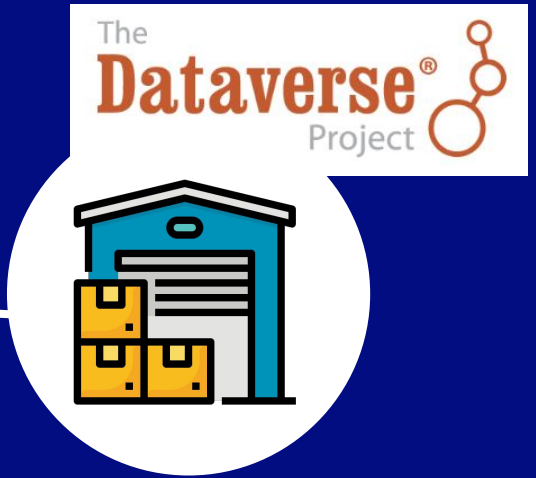
Replication dataset

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

Summary

- Dataverse data repositories have versatile support for data sharing
- Research data and code are shared in a “replication dataset” that often belong to a journal or institutional collection

How reusable are our replication datasets?



A large-scale study on research code quality and re-execution

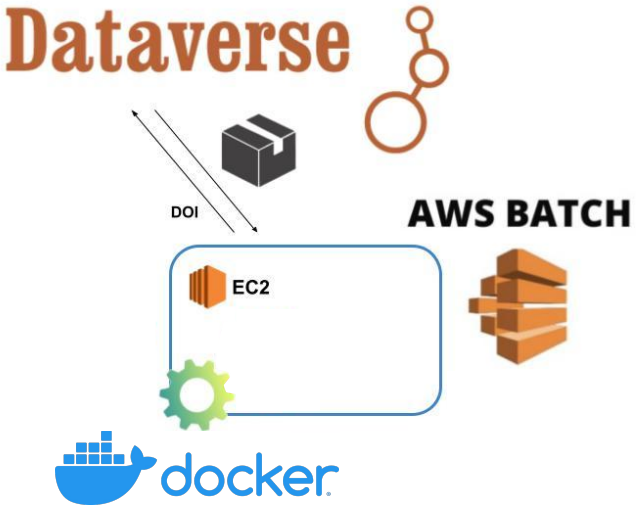
A large-scale study on research code quality and re-execution

reusable replication datasets

A large-scale study on research code quality and re-execution

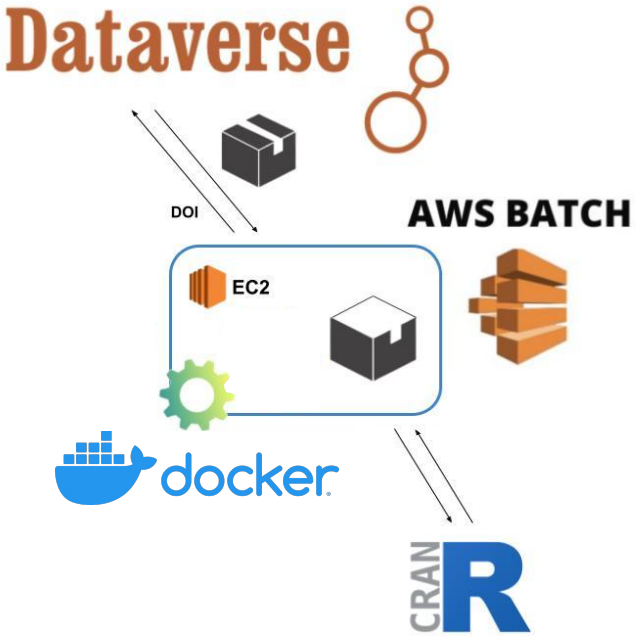
Data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, code, install used libraries etc.
3. We attempt code re-execution for an allocated time of 1h per file and 5h in total
4. The re-execution result and other collected data are passed to the backend database for analysis



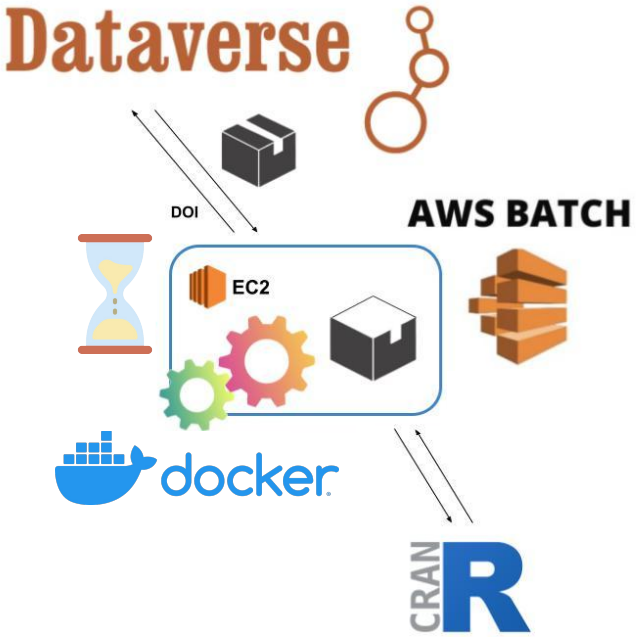
Data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, code, install used libraries etc.
3. We attempt code re-execution for an allocated time of 1h per file and 5h in total
4. The re-execution result and other collected data are passed to the backend database for analysis



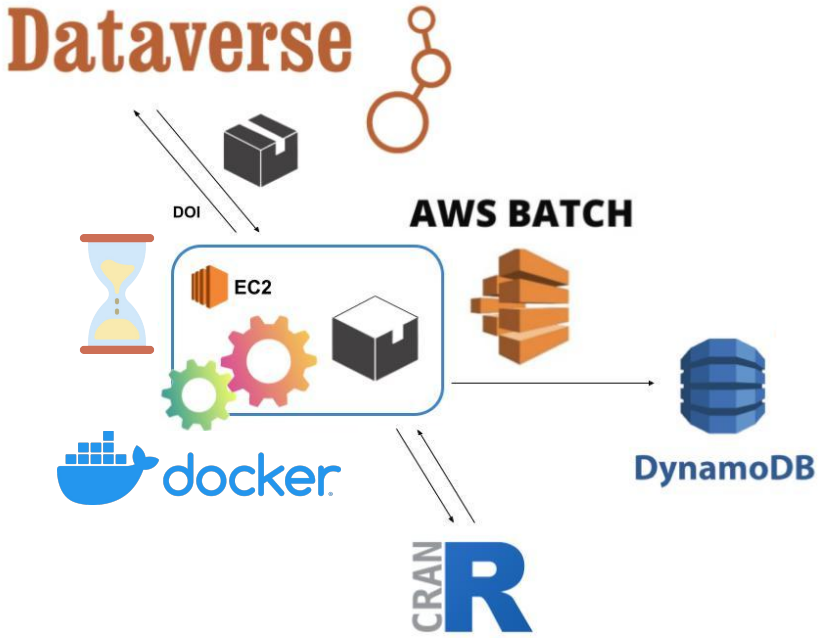
Data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, code, install used libraries etc.
3. We attempt code re-execution for an allocated time of 1h per file and 5h in total
4. The re-execution result and other collected data are passed to the backend database for analysis



Data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, code, install used libraries etc.
3. We attempt code re-execution for an allocated time of 1h per file and 5h in total
4. The re-execution result and other collected data are passed to the backend database for analysis



Results (basic properties)

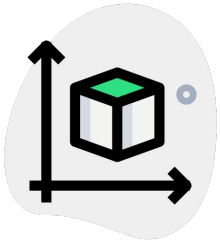
Retrieved 2109 publicly available replication datasets containing 9078 R files

Over 94% of the datasets belonged to social sciences

Results (basic properties)

Retrieved 2109 publicly available replication datasets containing 9078 R files

Over 94% of the datasets belonged to social sciences



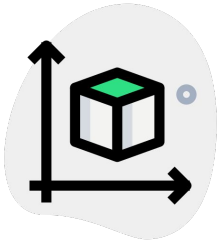
Dataset size (median): 3.2 MB

Number of files (median): 8 (typically less than 15)

Results (basic properties)

Retrieved 2109 publicly available replication datasets containing 9078 R files

Over 94% of the datasets belonged to social sciences



Dataset size (median): 3.2 MB

Number of files (median): 8 (typically less than 15)

- File name length: 10-20 characters
- Documentation present in 57% of the datasets
- Comments comprise 20% of the shared R code

Presence of conventional files

DESCRIPTION
README.md
LICENSE
NAMESPACE
Dockerfile
R Markdown
.Rproj
install.R

Marwick, Ben, Carl Boettiger, and Lincoln Mullen. "Packaging data analytical work reproducibly using R (and friends)." The American Statistician 72.1 (2018).

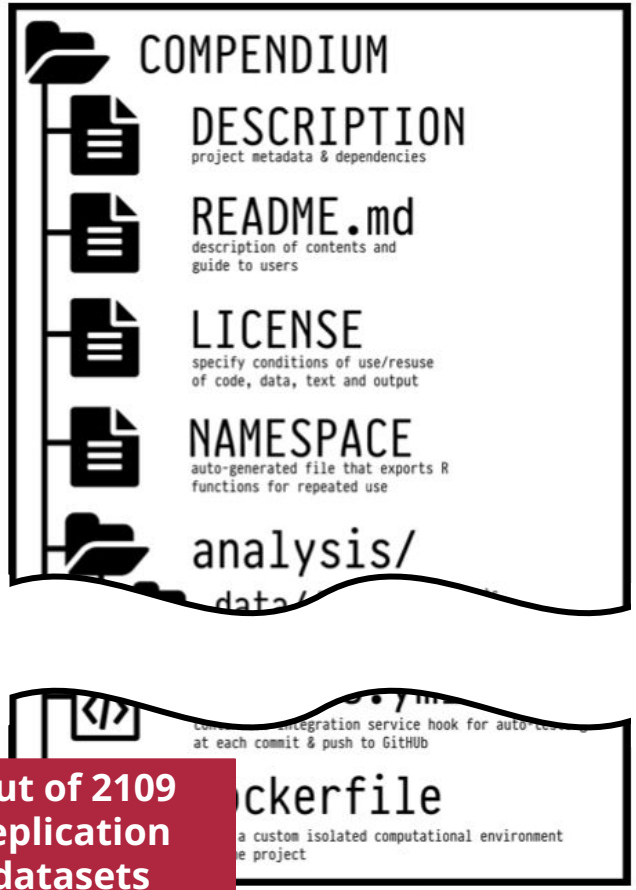
Presence of conventional files

DESCRIPTION	< 1%
README.md	
LICENSE	< 1%
NAMESPACE	< 1%
Dockerfile	< 1%
R Markdown	
.Rproj	< 1%
install.R	< 1%

Marwick, Ben, Carl Boettiger, and Lincoln Mullen. "Packaging data analytical work reproducibly using R (and friends)." The American Statistician 72.1 (2018).

Presence of conventional files

DESCRIPTION	< 1%
README.md	48%
LICENSE	< 1%
NAMESPACE	< 1%
Dockerfile	< 1%
R Markdown	3%
.Rproj	< 1%
install.R	< 1%



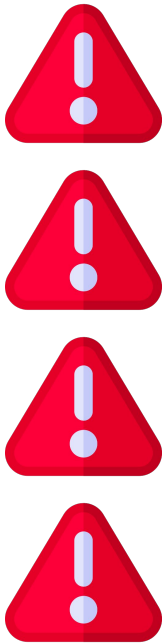
Out of 2109 replication datasets

Marwick, Ben, Carl Boettiger, and Lincoln Mullen. "Packaging data analytical work reproducibly using R (and friends)." The American Statistician 72.1 (2018).

Most used libraries in research code:

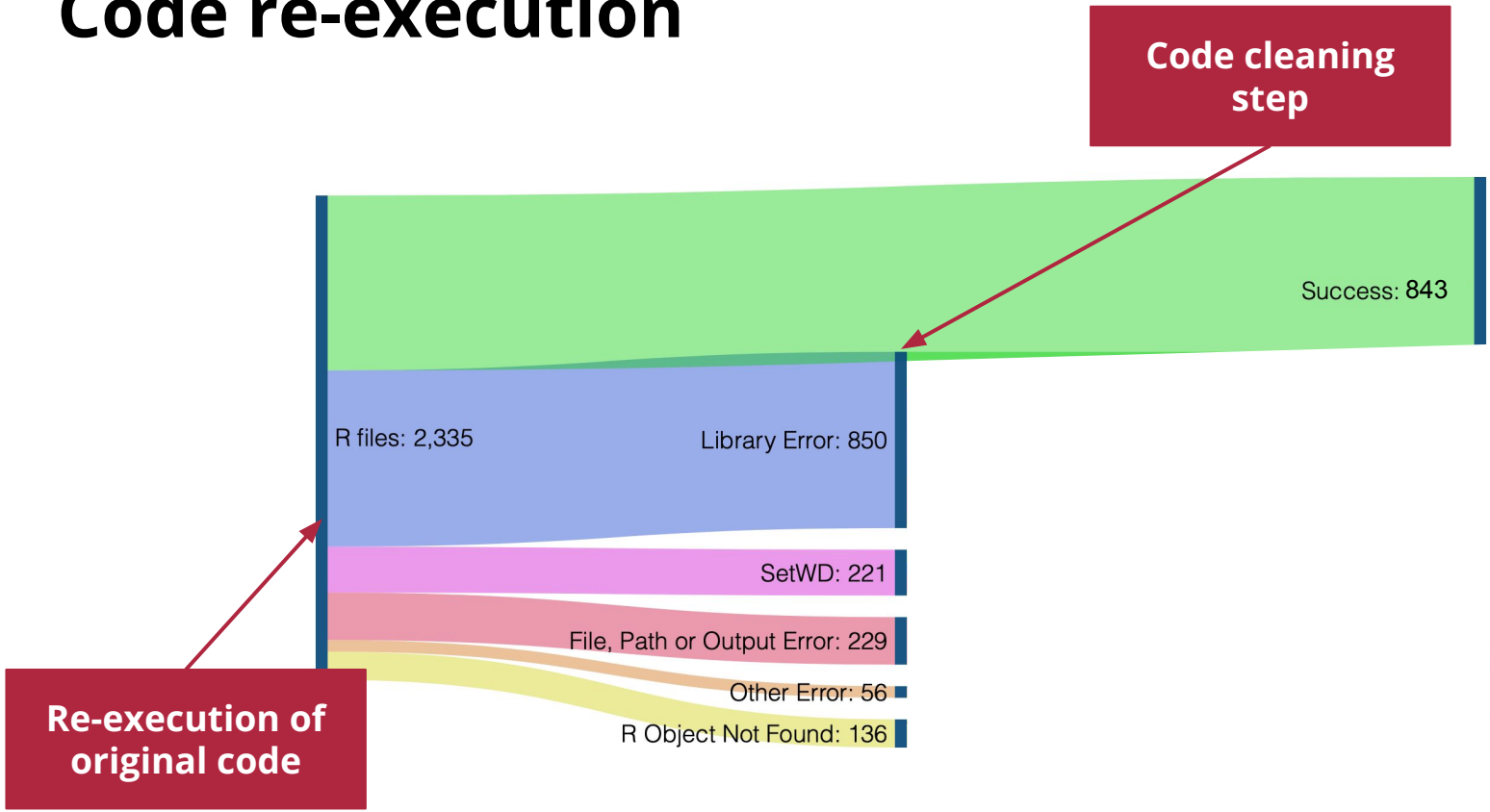
1. **Data visualization and plotting** (`ggplot2, lattice`)
2. **Data wrangling and display in a tabular form** (`xtable`)
3. **Data import and export** (`foreign, dplyr, plyr, reshape2`)
4. **Statistical analysis** (`stargazer, MASS, lmetest, car`)

In the research code, no libraries detected for:



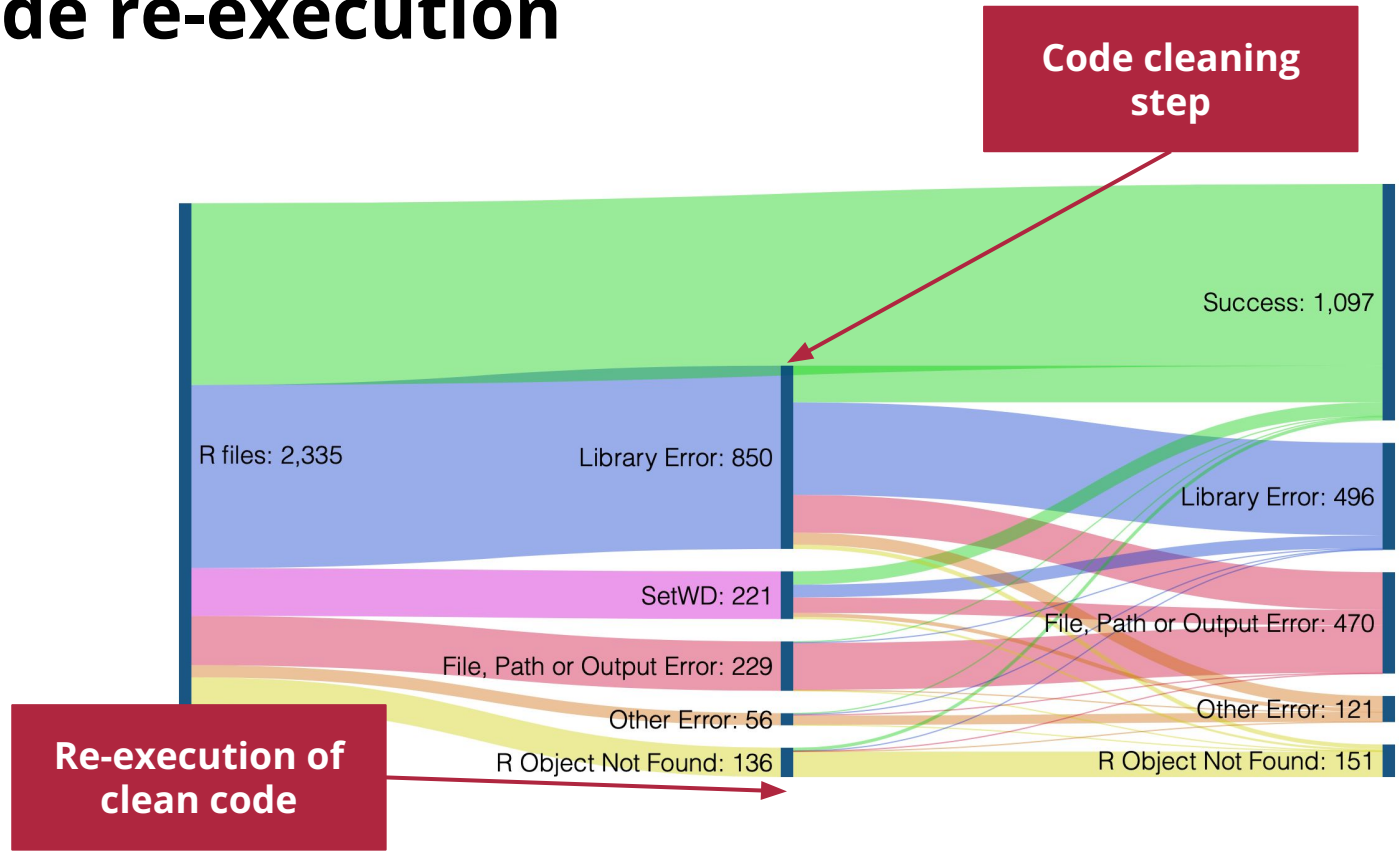
1. **Code testing** (`runit, testthat, tinytest, unitizer`)
2. **Provenance tracking** (`provR, provenance, RDTlite, provTraceR`)
3. **Runtime environment management** (`packrat, pacman`)
4. **Workflow libraries** (`workflowR, workflows, drake`)

Code re-execution



Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv preprint arXiv:2103.12793 (2021).

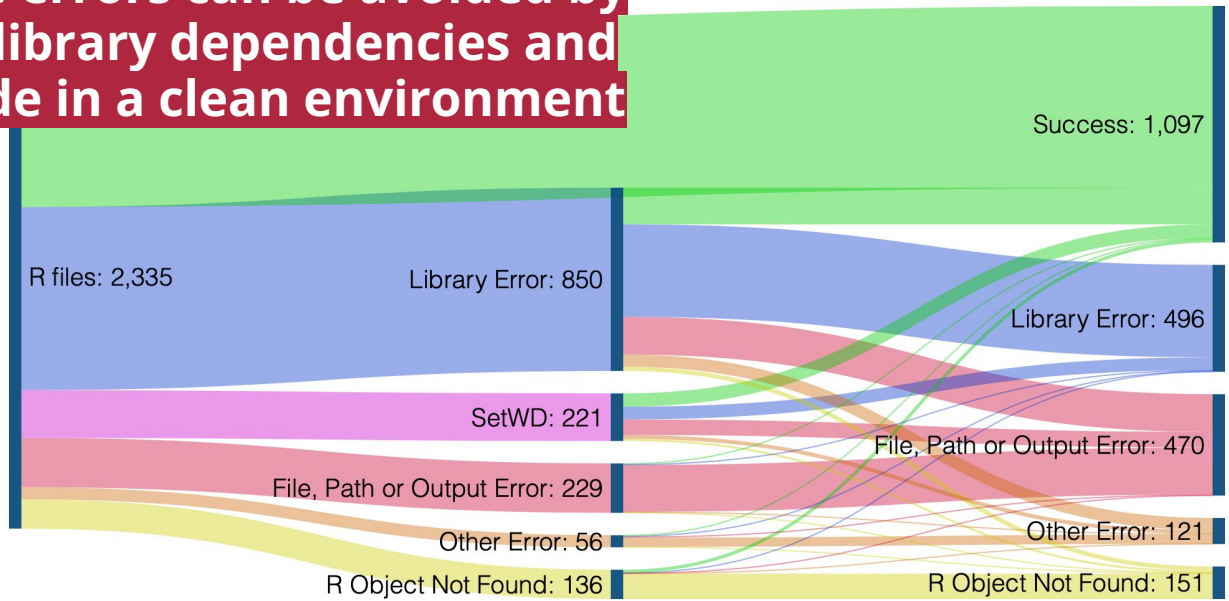
Code re-execution



Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv preprint arXiv:2103.12793 (2021).

Code re-execution

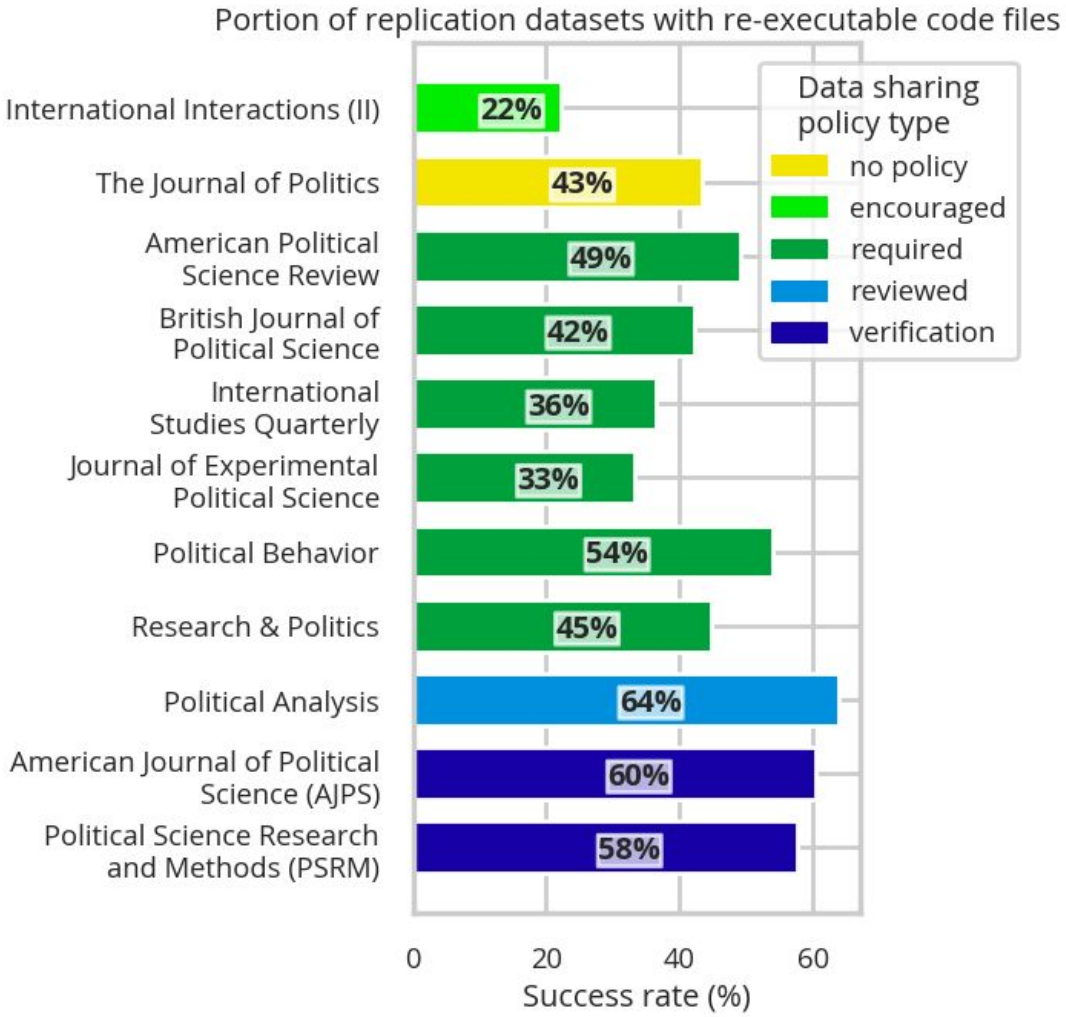
Many code errors can be avoided by capturing library dependencies and testing code in a clean environment



Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv preprint arXiv:2103.12793 (2021).

Journals with stricter data policies have higher rate of executable code

Journal average: 47%
Total average: 45%

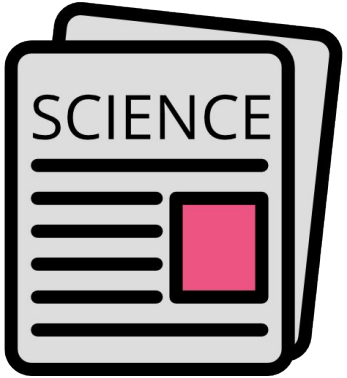


Summary

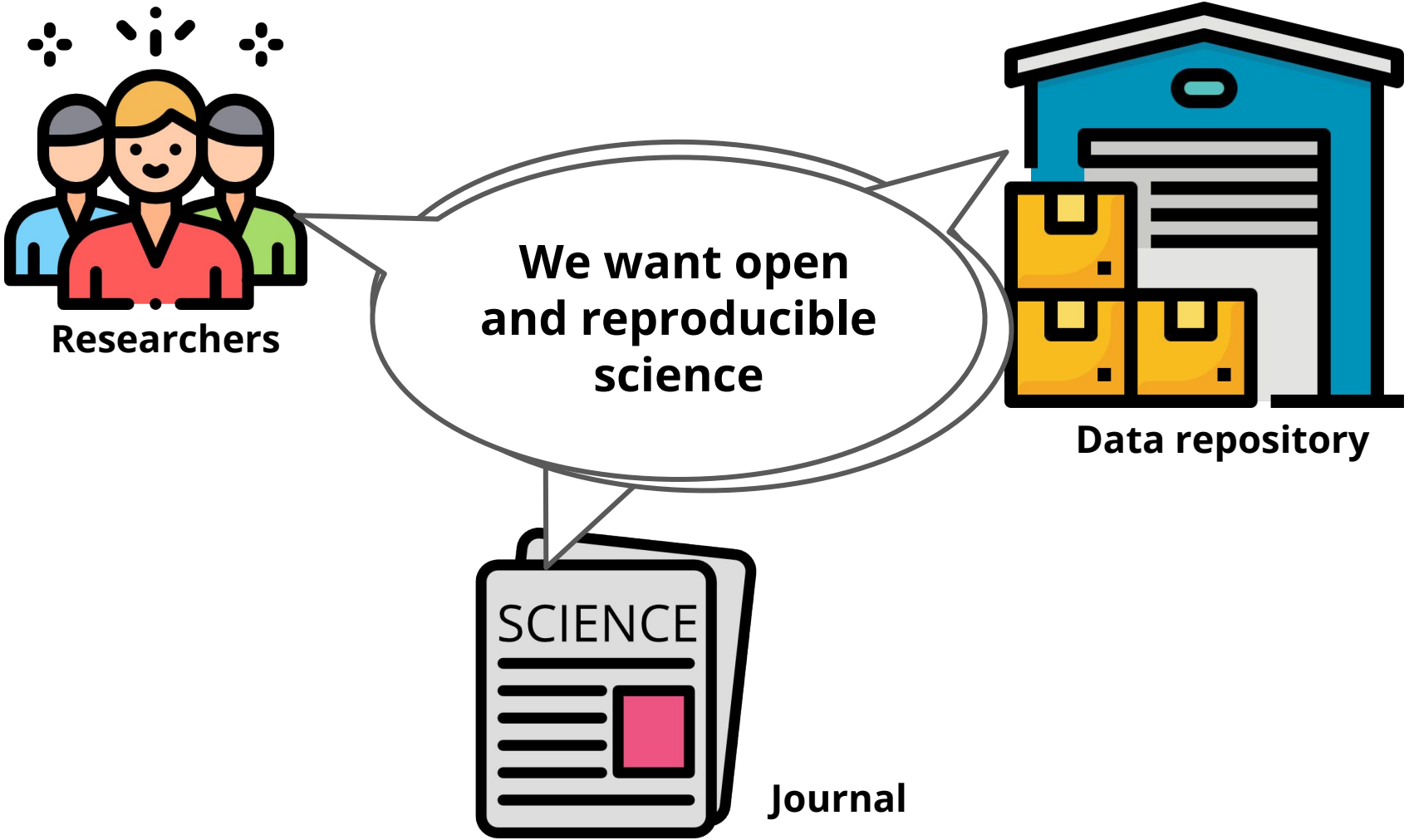
- No libraries for code testing, provenance or workflows
- Simple code cleaning resulted in substantial improvement in re-execution
- Re-execution correlates with journals' data sharing policy strictness

Researchers

Data repository



Journal


We want open and reproducible science
Researchers
Data repository
SCIENCE
Journal

What can researchers do?

Best practices and recommendations

1. Library versions should be captured by, minimally, the output of `sessionInfo()` from the researcher's R session, or a `DESCRIPTION` file, or `install.R`, or by using the `renv` package to track the libraries and their version number.
2. When referring to data, code or other files, use relative file paths, as full paths will cause an error when the code is executed on other systems.

Best practices and recommendations

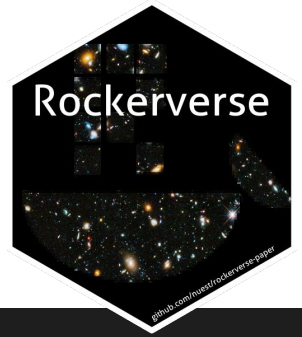
1. Library versions should be captured by, minimally, the output of `sessionInfo()` from the researcher's R session, or a `DESCRIPTION` file, or `install.R`, or by using the `renv` package to track the libraries and their version number.
2. When referring to data, code or other files, use relative file paths, as full paths will cause an error when the code is executed on other systems.

Best practices and recommendations

3. Workflow capture and managementment methods such as `R Markdown`, `targets` (or `drake`) will help to automate your code and specify the correct execution sequence.
4. Use Docker to document your runtime environment in a machine-readable format, and to ensure others can recreate your computing environment.

Best practices and recommendations

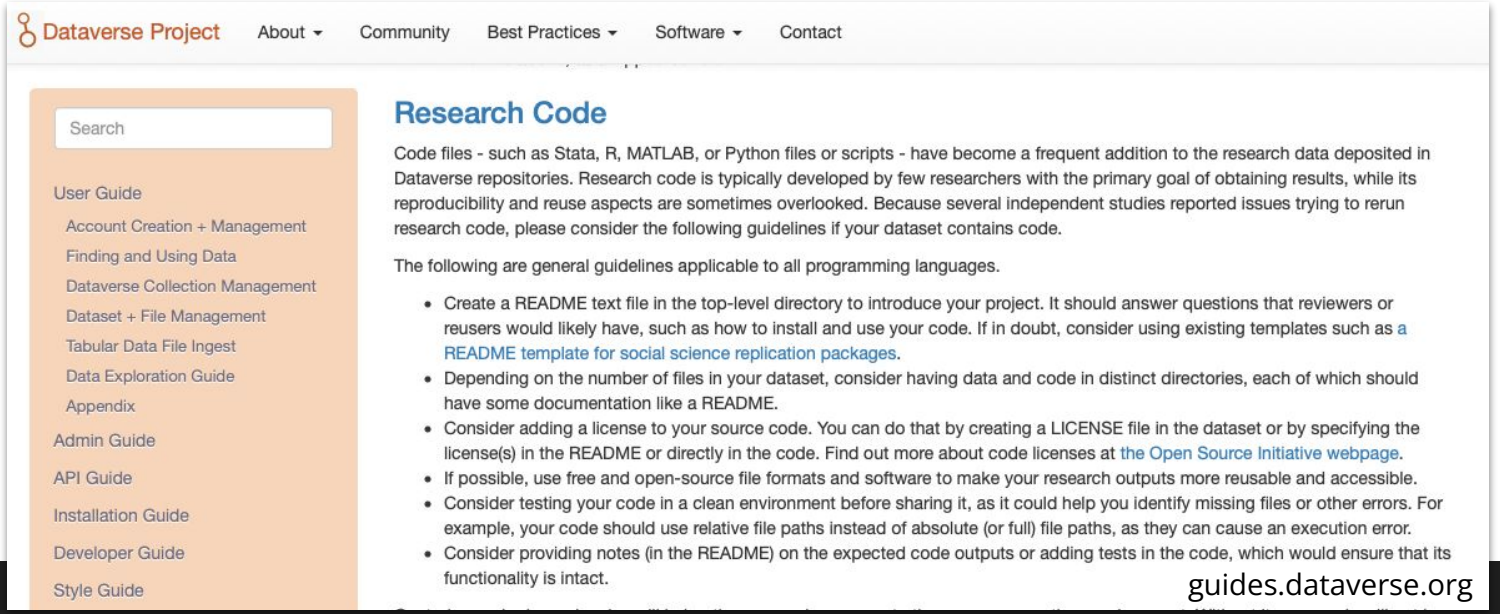
3. Workflow capture and managementment methods such as `R Markdown`, `targets` (or `drake`) will help to automate your code and specify the correct execution sequence.
4. Use Docker to document your runtime environment in a machine-readable format, and to ensure others can recreate your computing environment.



What can repositories do?

Best practices and recommendations

1. Create and maintain documentation on adequate deposit of research code.



guides.dataverse.org

Best practices and recommendations

2. Integrations with reproducibility platforms such as CodeOcean, WholeTale, Jupyter Binder and Renku will facilitate environment capture and encapsulation of research code.

Best practices and recommendations

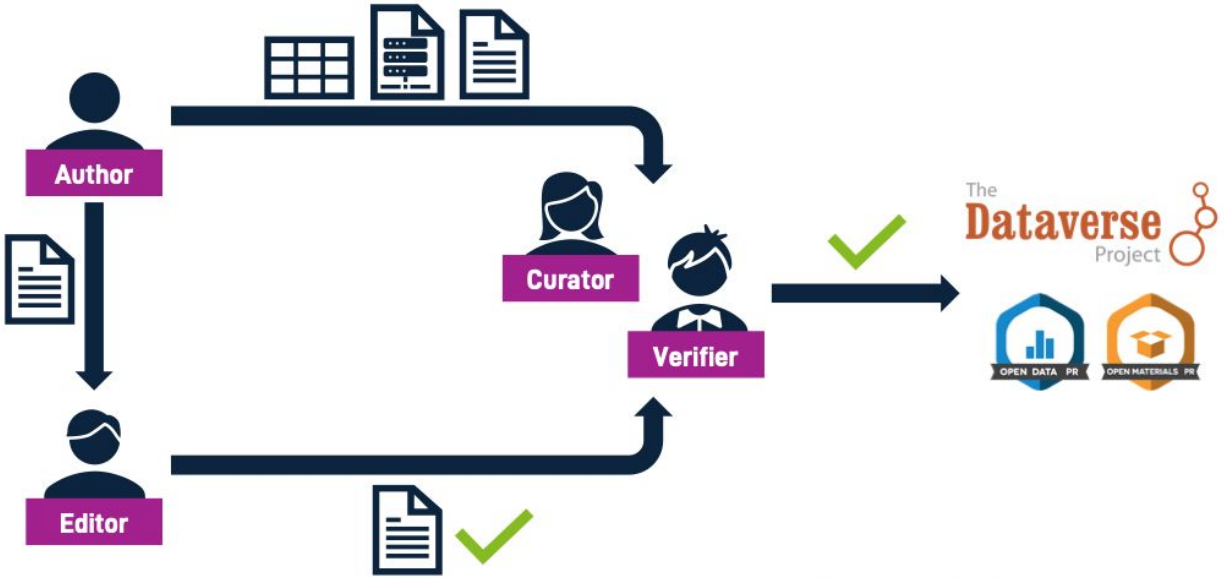
3. An internal working group will help identify community-wide problems, prioritize them, and implement solutions.

https://swc.wgs.gdcc.io

What can journals do?

Best practices and recommendations

1. Encourage a simple review of all deposited material if a code verification is infeasible



Best practices and recommendations

2. Create reproducibility checklist or templates for authors

A template README for social science replication packages.

The template README provided on this website is in a form that follows best practices as defined by a number of data editors at social science journals.

Template README and Guidance

> *INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see Endorsers. It is available as Markdown/txt, Word, LaTeX, and PDF. In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with > INSTRUCTIONS). Please ensure that a PDF is submitted in addition to the chosen native format.*

Overview

> *INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.*

Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al, 2018; Inglehart et al, 2019; BEA, 2016) using Stata and Julia. Two main files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

Data Availability and Provenance Statements

https://social-science-data-editors.github.io

@larsvil

Best practices and recommendations

3. Integrations with reproducibility platforms

Conclusions

Conclusion

- We’ve seen evidence of both good and bad coding and dissemination practices (documentation, commenting, convention files rarely used)
- It is hard to re-execute “old” code and even harder to reuse it
  - Curated replication datasets have higher re-execution rates.
  - Things are looking up!
- Employing proposed recommendations would help researchers, repositories and journals contribute to research transparency and reproducibility.

Conclusion

- We’ve seen evidence of both good and bad coding and dissemination practices (documentation, commenting, convention files rarely used)
- It is hard to re-execute “old” code and even harder to reuse it
  - Curated replication datasets have higher re-execution rates.
  - Things are looking up!
- Employing proposed recommendations would help researchers, repositories and journals contribute to research transparency and reproducibility.

References

- This presentation was based on findings at Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv:2103.12793 (2021).

A large-scale study on
research code quality and execution

Ana Trisovic[1*], Matthew K. Lau[2], Thomas Pasquier[3], Mercè Crosas[1]

March 25, 2021

1. Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA
2. CAS Key Laboratory of Forest Ecology and Management, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang, China
3. Department of Computer Science, University of British Columbia, Vancouver, BC, Canada
*corresponding author(s): Ana Trisovic (anatrisovic@g.harvard.edu)

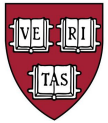
Abstract

This article presents a study on the quality and execution of research code from publicly-available replication datasets at the Harvard Dataverse repository. Research code is typically created by a group of scientists and published together with academic papers to facilitate research transparency and reproducibility. For this study, we define ten questions to address aspects impacting research reproducibility and reuse. First, we retrieve and analyze more than 2000 replication datasets with over 9000 unique R files published from 2010 to 2020. Second, we execute the code in a clean runtime environment to assess its ease of reuse. Common coding errors were identified, and some of them were solved with automatic code cleaning to aid code execution. We find that 74% of R files crashed in the initial execution, while 56% crashed when code cleaning was applied, showing that many errors can be prevented with good coding practices. We also analyze the replication datasets from journals' collections and discuss the impact of the journal policy strictness on the code re-execution rate. Finally, based on our results, we propose a set of recommendations for code dissemination aimed at researchers, journals, and repositories.

1 Introduction

Researchers increasingly publish their data and code to enable scientific transparency, reproducibility, reuse, or compliance with funding bodies, journals, and academic institutions [1]. Reusing data and code should propel new research and save researchers' time, but in practice, it is often easier to write new code

Acknowledgements:
icons by Freepik, Pixel perfect, Nhor Phai, Darius Dan (flaticon.com)



Thank you!

Email: anatrisovic@g.harvard.edu
GitHub & Twitter: atrisovic
Dataverse Project: https://dataverse.org/contact