ALFRED P. SLOAN
FOUNDATION

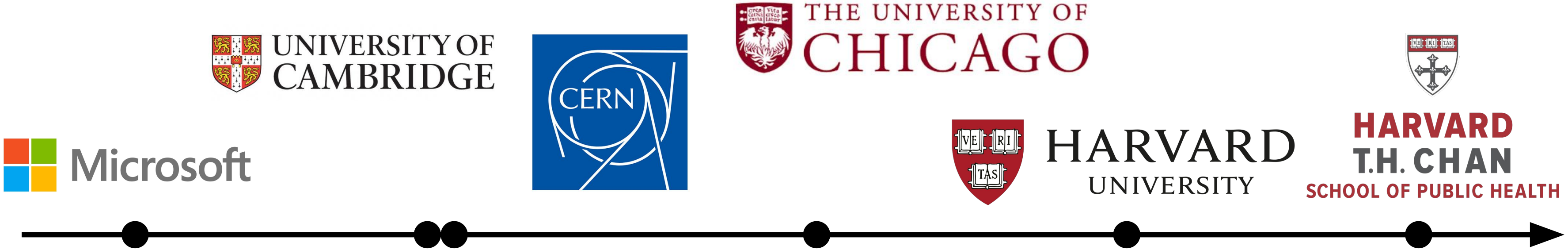
The importance of research data preservation and reproducibility

Belle II Data Preservation Workshop,
The Dipartimento di Ingegneria of the Roma Tre University
October 6, 2022

Ana Trisovic, Harvard University

About me

- PhD on “Data preservation and reproducibility at LHCb”
- Focus on data science, data engineering and reproducibility


Microsoft
UNIVERSITY OF CAMBRIDGE
CERN
THE UNIVERSITY OF CHICAGO
HARVARD UNIVERSITY
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Data Science Associate at Microsoft in 2013

Started my PhD at University of Cambridge and CERN in 2014

Started a postdoc at the University of Chicago in 2018

Started a postdoc at Harvard University in 2019

Research Associate at HSPH in 2022

Outline

- The potential of data preservation
- The power of open data
- The reproducibility challenge
- Summary & recommendations

The potential of data preservation

Ancel Keys

- **Minnesota Coronary Experiment** by Ancel Keys and Ivan Frantz
  - Randomized and blinded control trial in late 1960s
  - No positive effects of the altered dietary intake

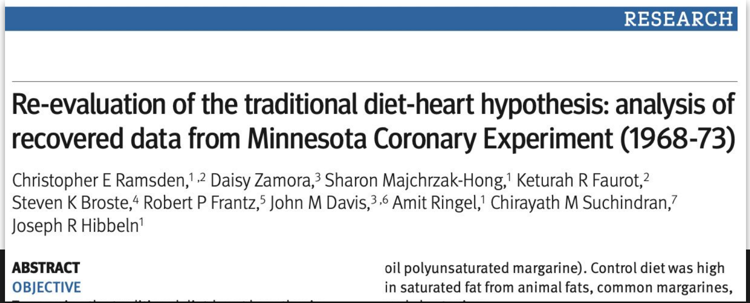
The potential of data preservation

Ancel Keys

- **Minnesota Coronary Experiment** by Ancel Keys and Ivan Frantz
  - Randomized and blinded control trial in late 1960s
  - No positive effects of the altered dietary intake
- The raw data and analysis were discovered in 2013

The potential of data preservation

Ancel Keys



RESEARCH

Re-evaluation of the traditional diet-heart hypothesis: analysis of recovered data from Minnesota Coronary Experiment (1968-73)

Christopher E Ramsden,[1,2] Daisy Zamora,[3] Sharon Majchrzak-Hong,[1] Keturah R Faurot,[2] Steven K Broste,[4] Robert P Frantz,[5] John M Davis,[3,6] Amit Ringel,[1] Chirayath M Suchindran,[7] Joseph R Hibbeln[1]

ABSTRACT

OBJECTIVE

oil polyunsaturated margarine). Control diet was high in saturated fat from animal fats, common margarines,

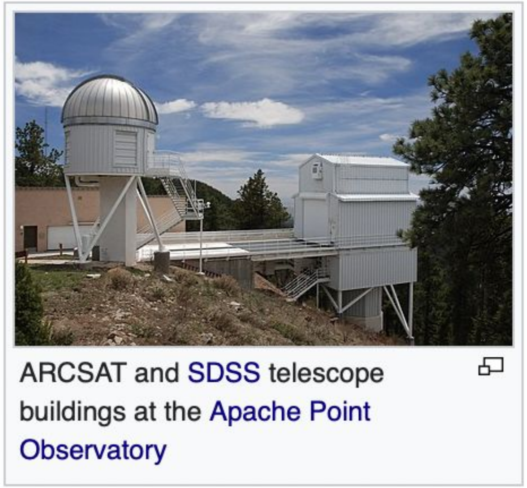
- **Minnesota Coronary Experiment** by Ancel Keys and Ivan Frantz
  - Randomized and blinded control trial in late 1960s
  - No positive effects of the altered dietary intake
- The raw data and analysis were discovered in 2013
- New knowledge with the analysis of recovered data

Takeaway

The potential of data preservation lies in research verification and data reuse

- Primary vs. Secondary data analysis

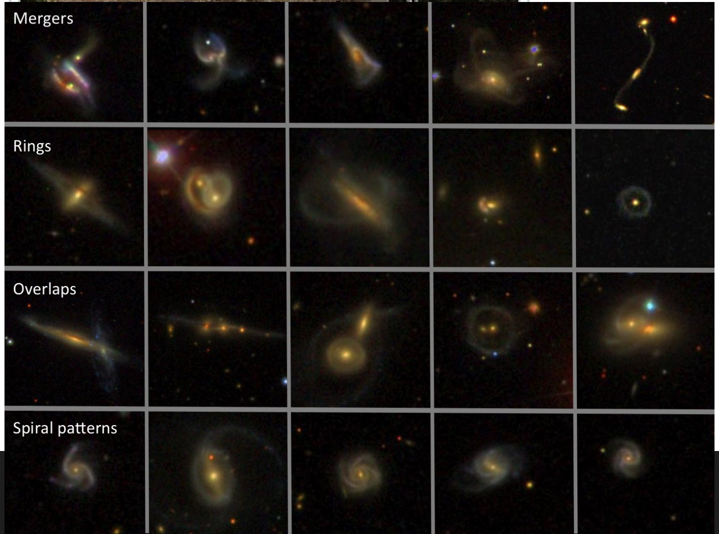
The power of open data



ARCSAT and SDSS telescope buildings at the Apache Point Observatory

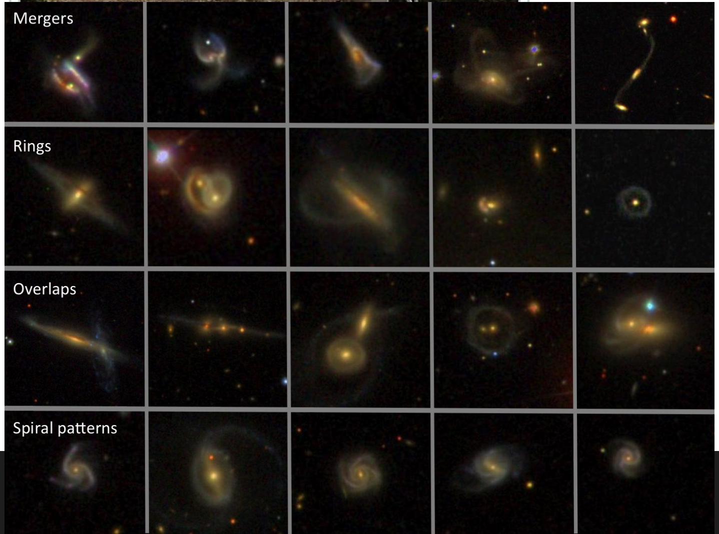
- The Sloan Digital Sky Survey or SDSS

The power of open data



- The Sloan Digital Sky Survey or SDSS
- Galaxy Zoo - a crowdsourced astronomy project that invites general public to assist in the classification of large number of galaxies

The power of open data

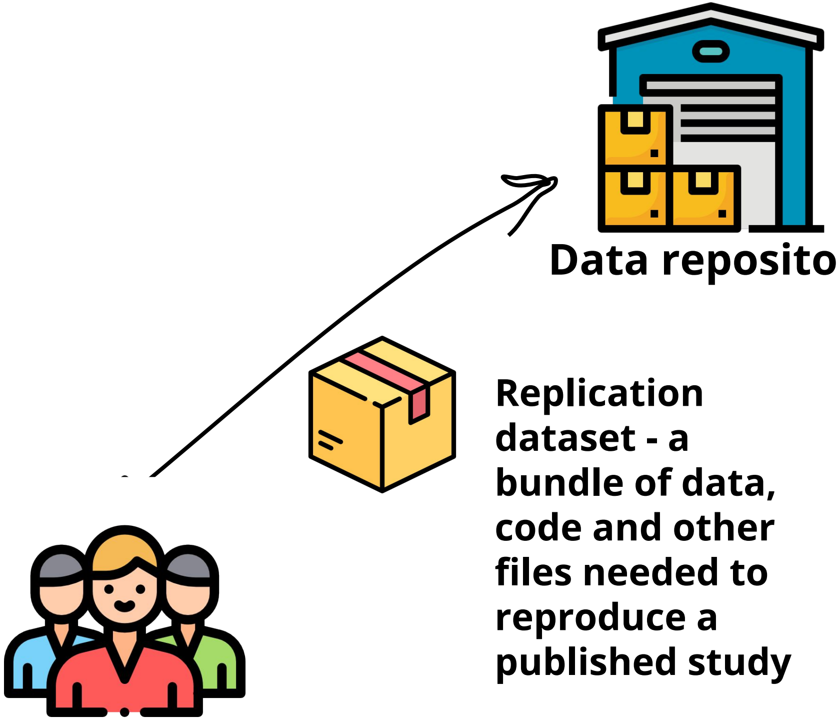


- The Sloan Digital Sky Survey or SDSS
- Galaxy Zoo - a crowdsourced astronomy project that invites general public to assist in the classification of large number of galaxies
    - Over 500k people contributed to the effort
    - The team published over 50 papers
    - Inspired similar citizen science projects across astrophysics and beyond

Takeaway

This potential of data is maximized when the data is shared as open data

The challenge of reproducibility

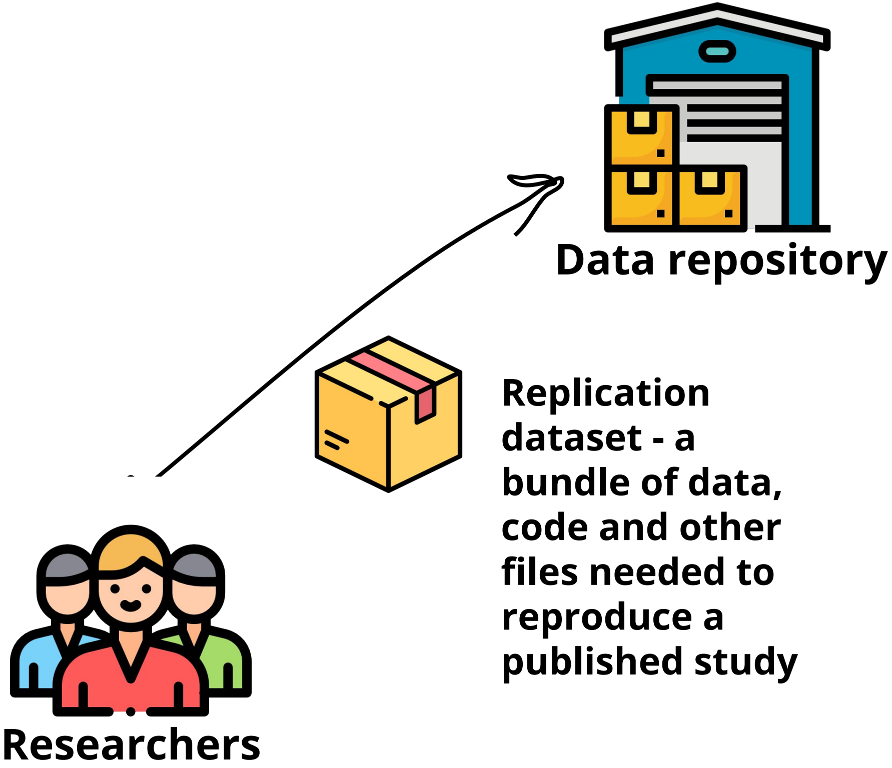


Data repository

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

Researchers

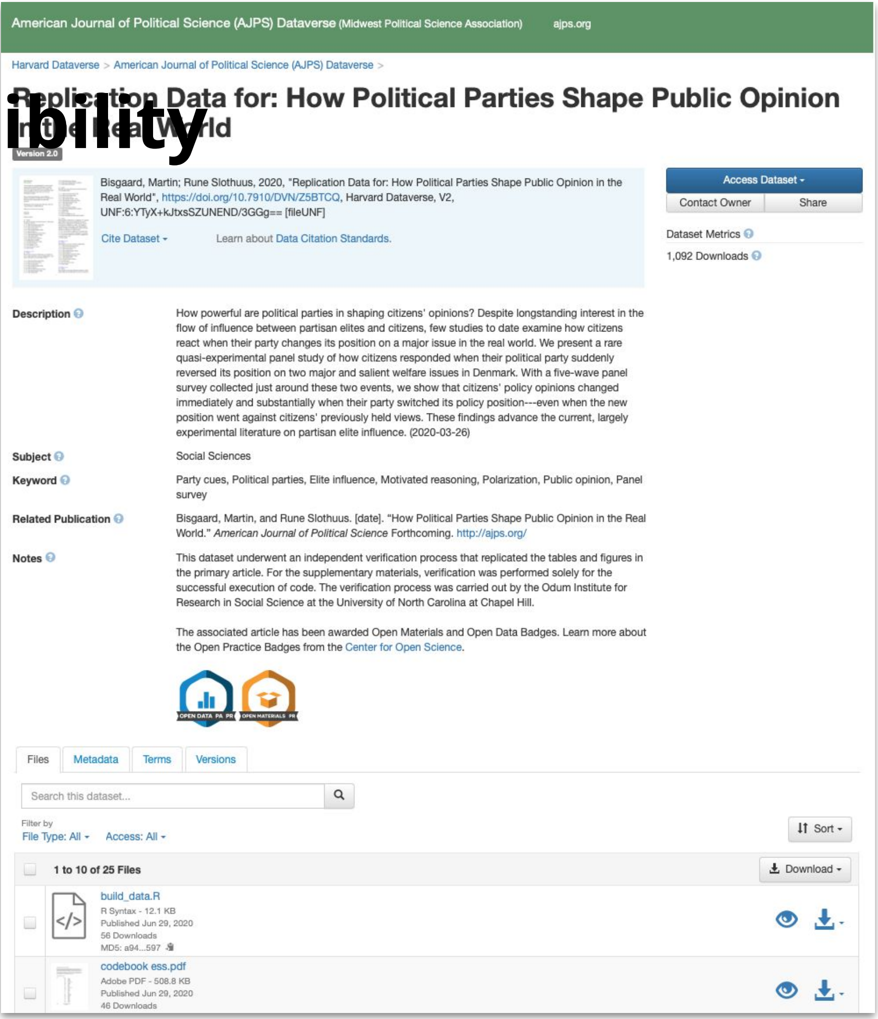
The challenge of reproducibility



Data repository

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

Researchers



American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", https://doi.org/10.7910/DVN/Z5BTCQ, Harvard Dataverse, V2, UNF:6:YTyX+kJtxsSZUNEND/3GGg== [fileUNF]

Cite Dataset ▾ Learn about Data Citation Standards.

Dataset Metrics

1,092 Downloads

Description How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy position---even when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)

Subject Social Sciences

Keyword Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. http://ajps.org/

Notes This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.

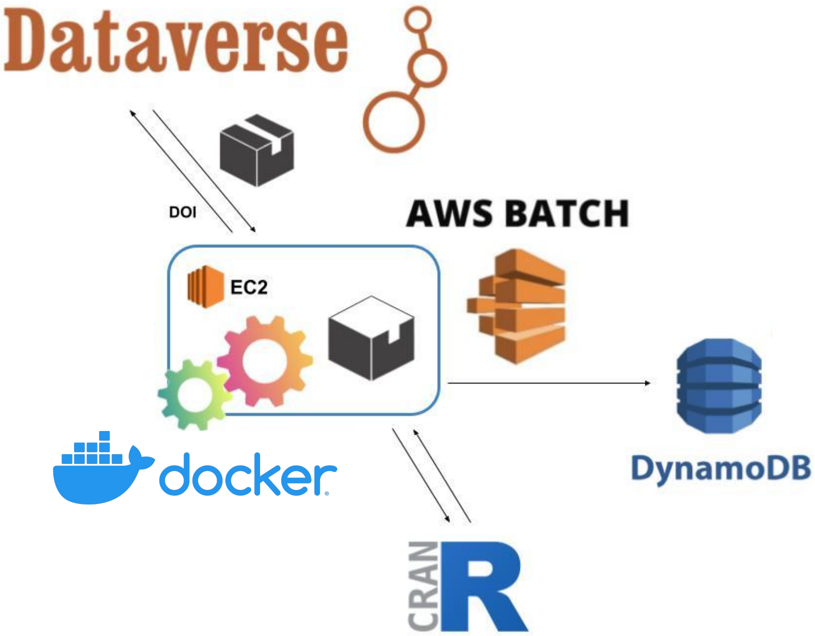
□ 1 to 10 of 25 Files

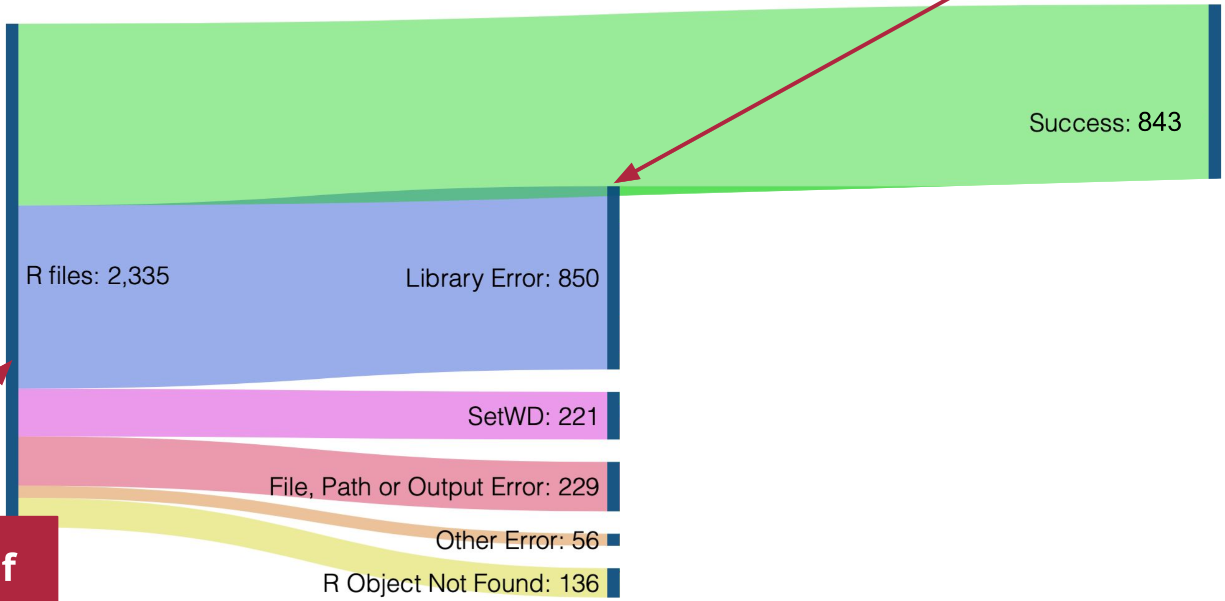
□ build_data.R
R Syntax - 12.1 KB
Published Jun 29, 2020
56 Downloads
MD5: a94...597

□ codebook ess.pdf
Adobe PDF - 508.8 KB
Published Jun 29, 2020
46 Downloads

Our data collection workflow

1. Replication dataset is retrieved from Harvard Dataverse to AWS
2. We collect data on the content, install used libraries and attempt automatic code re-execution
3. The re-execution result and other collected data are passed to the backend database for analysis

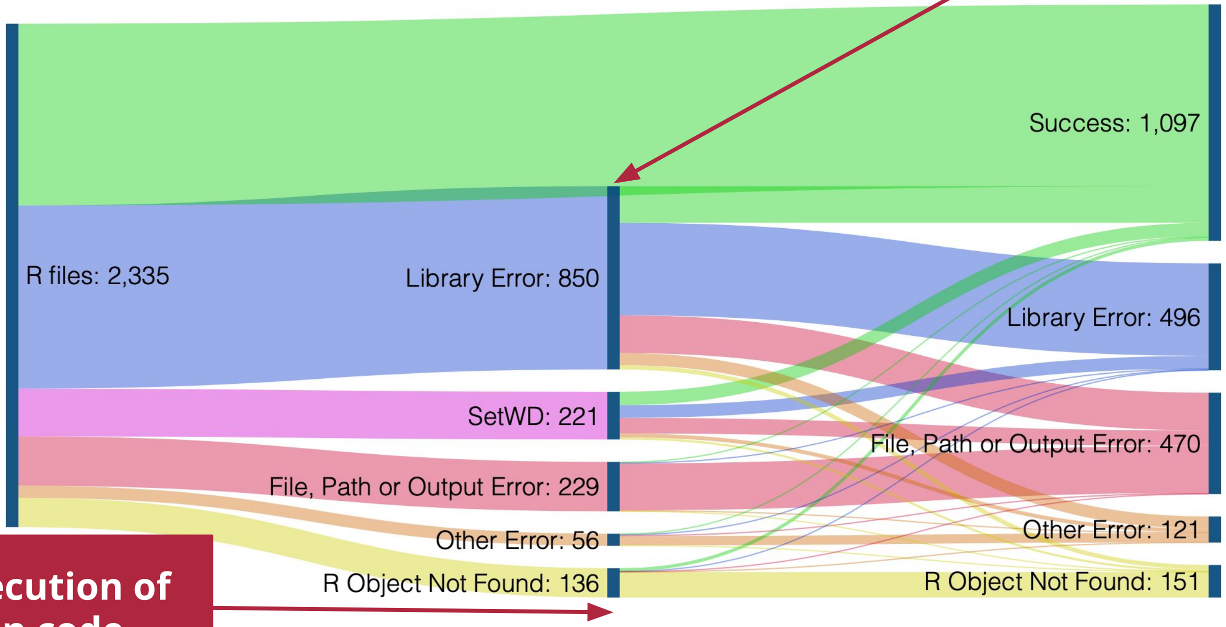




Automatized code cleaning

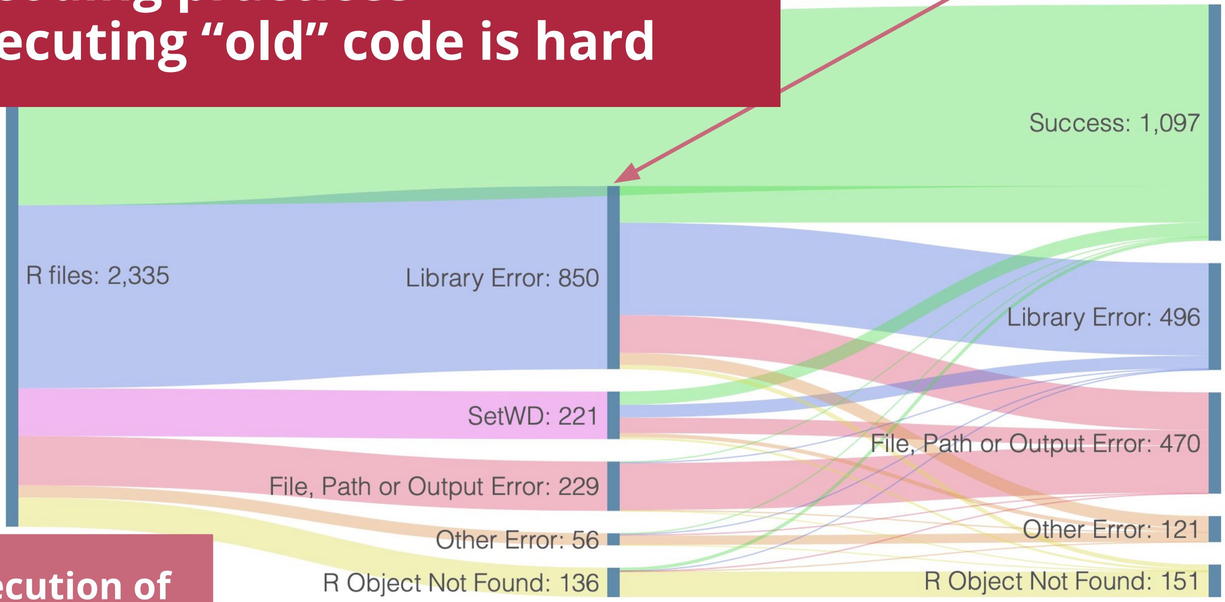
Re-execution of original code

Automatized code cleaning



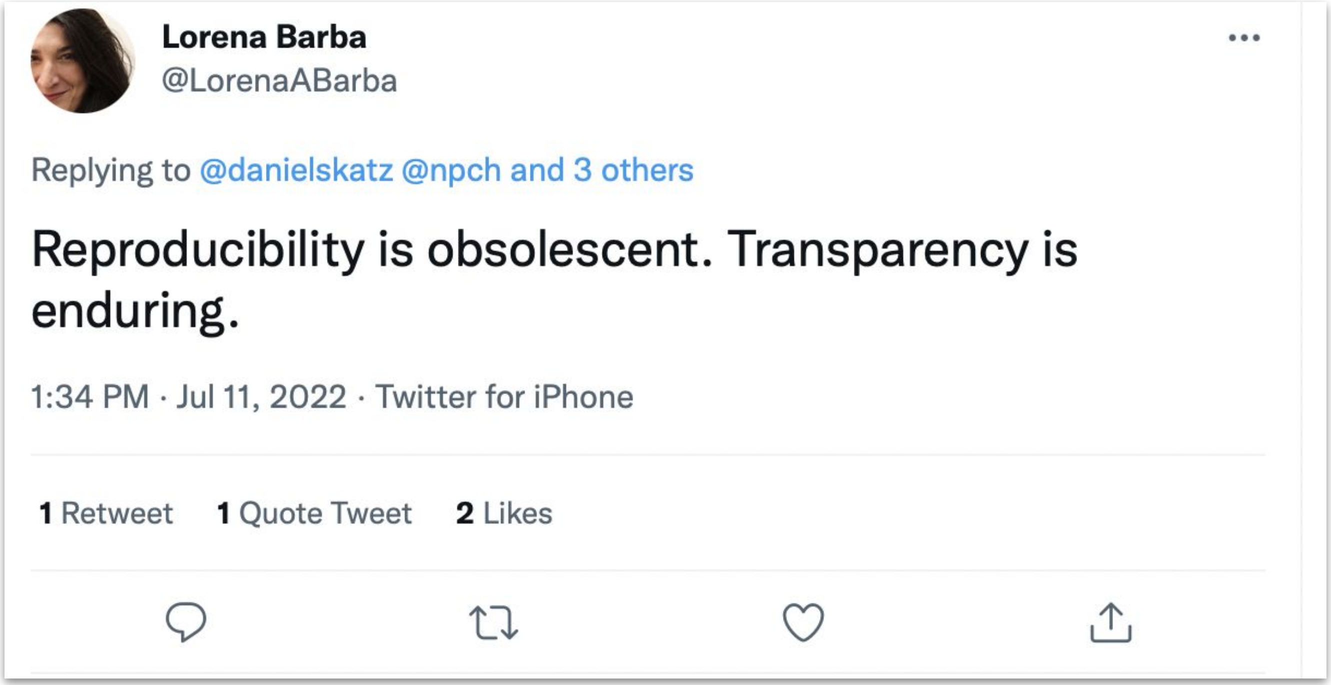
Re-execution of clean code

- **Many errors can be avoided with good coding practices**
- **Re-executing “old” code is hard**



Automatized code cleaning

Re-execution of clean code



Lorena Barba

@LorenaABarba

Replying to @danielskatz @npch and 3 others

Reproducibility is obsolescent. Transparency is enduring.

1:34 PM · Jul 11, 2022 · Twitter for iPhone

1 Retweet **1** Quote Tweet **2** Likes

Takeaway

Reproducibility as a realistic goal for the short term, transparency for the long term

Summary & recommendations

So why is data preservation important?

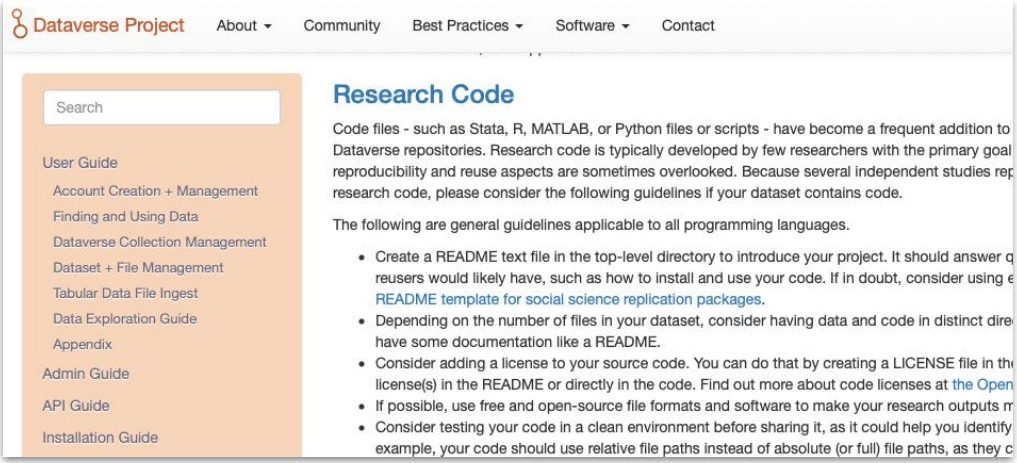
- Reuse allows for new scientific results
- Data sharing facilitates collaboration

So why is data preservation important?

- Reuse allows for new scientific results
- Data sharing facilitates collaboration
- Result verification & troubleshooting
- Education and training

Recommendations

- Guidance



Research Code

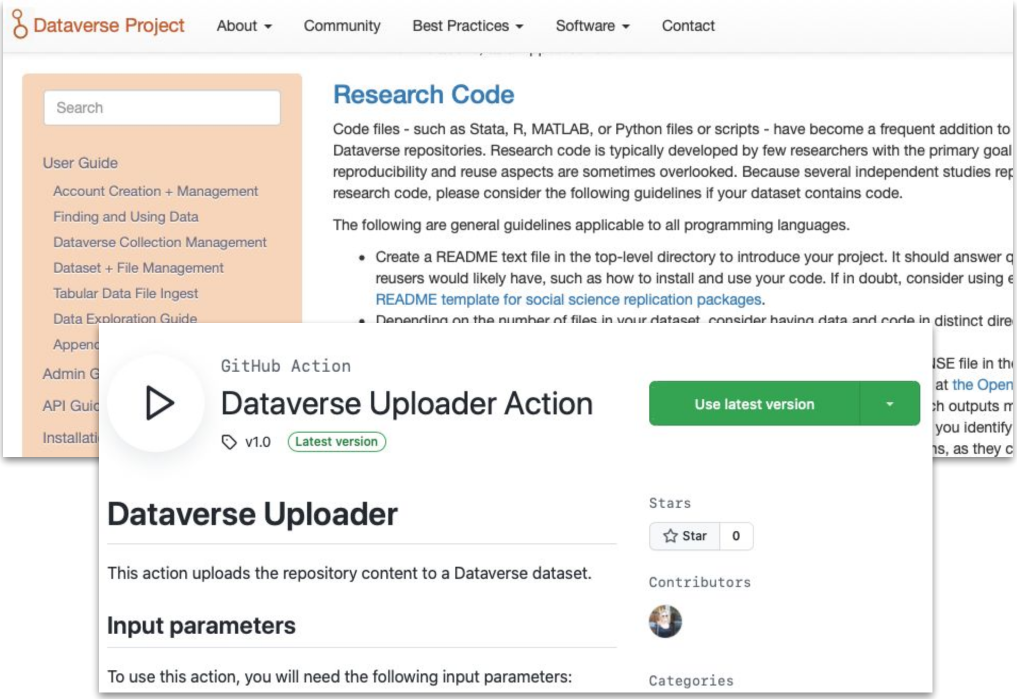
Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to
Dataverse repositories. Research code is typically developed by few researchers with the primary goal
reproducibility and reuse aspects are sometimes overlooked. Because several independent studies rep
research code, please consider the following guidelines if your dataset contains code.

The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer q
reusers would likely have, such as how to install and use your code. If in doubt, consider using e
README template for social science replication packages.
- Depending on the number of files in your dataset, consider having data and code in distinct dire
have some documentation like a README.
- Consider adding a license to your source code. You can do that by creating a LICENSE file in th
license(s) in the README or directly in the code. Find out more about code licenses at the Open
- If possible, use free and open-source file formats and software to make your research outputs m
- Consider testing your code in a clean environment before sharing it, as it could help you identify
example, your code should use relative file paths instead of absolute (or full) file paths, as they c

Recommendations

- Guidance
- Helper tools



Research Code

Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to Dataverse repositories. Research code is typically developed by few researchers with the primary goal reproducibility and reuse aspects are sometimes overlooked. Because several independent studies rep research code, please consider the following guidelines if your dataset contains code.

The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer q reusers would likely have, such as how to install and use your code. If in doubt, consider using e README template for social science replication packages.
- Depending on the number of files in your dataset, consider having data and code in distinct dire

SE file in th
at the Open
h outputs m
you identify
s, as they c

Dataverse Uploader Action

Dataverse Uploader

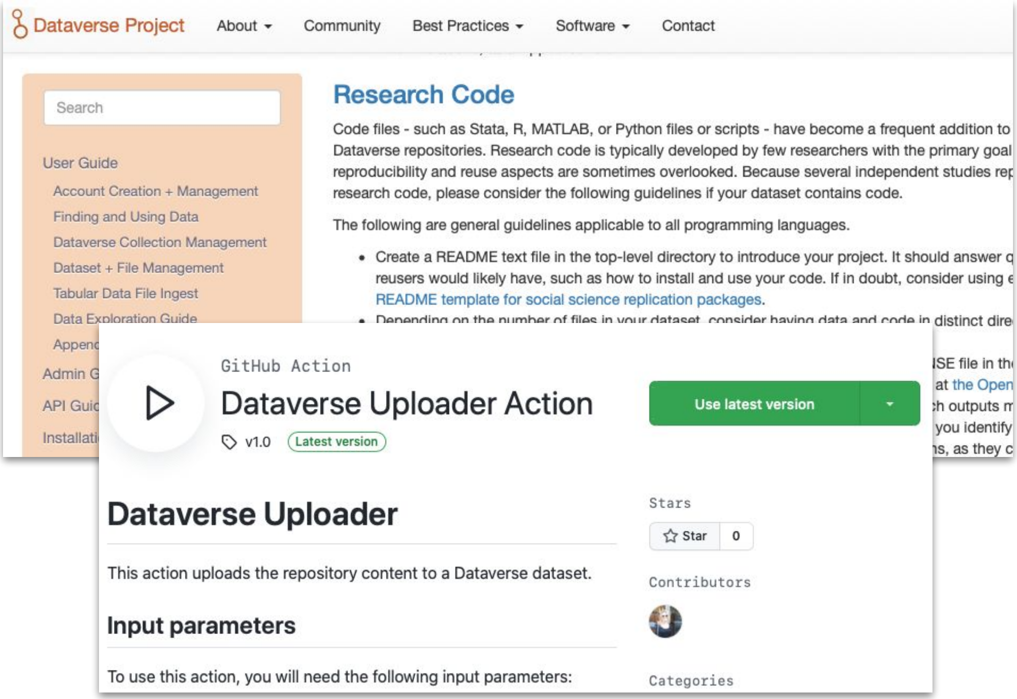
This action uploads the repository content to a Dataverse dataset.

Input parameters

To use this action, you will need the following input parameters:

Recommendations

- Guidance
- Helper tools



Research Code

Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to
Dataverse repositories. Research code is typically developed by few researchers with the primary goal
reproducibility and reuse aspects are sometimes overlooked. Because several independent studies rep
research code, please consider the following guidelines if your dataset contains code.

The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer q
reusers would likely have, such as how to install and use your code. If in doubt, consider using e
README template for social science replication packages.
- Depending on the number of files in your dataset, consider having data and code in distinct dire

SE file in th
at the Open
h outputs m
you identify
s, as they c

Dataverse Uploader Action

Dataverse Uploader

This action uploads the repository content to a Dataverse dataset.

Input parameters

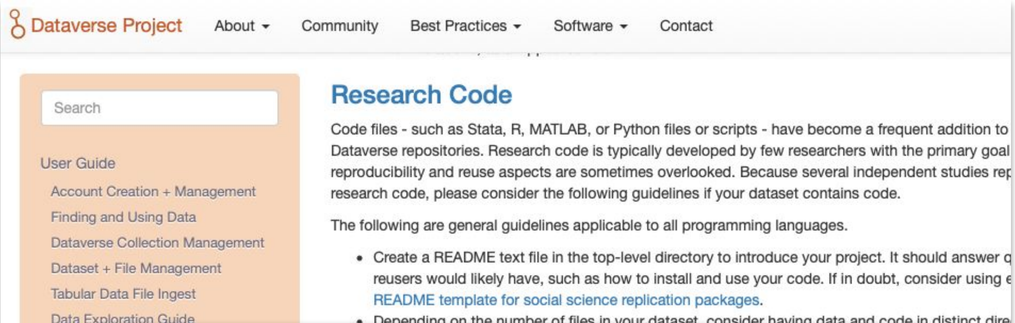
To use this action, you will need the following input parameters:

Stars
Star
0
Contributors
Categories

Dataverse®

Recommendations

- Guidance
- Helper tools
- Policy



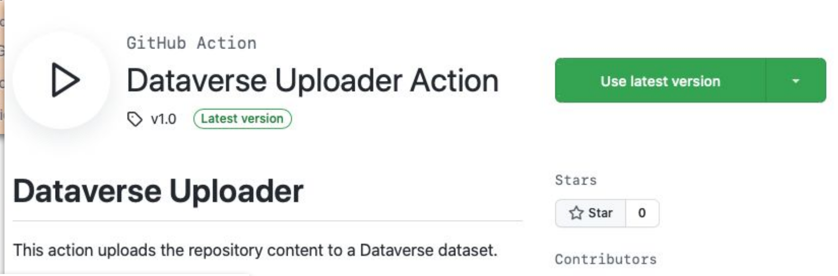
Research Code

Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to Dataverse repositories. Research code is typically developed by few researchers with the primary goal reproducibility and reuse aspects are sometimes overlooked. Because several independent studies rep research code, please consider the following guidelines if your dataset contains code.

The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer q reusers would likely have, such as how to install and use your code. If in doubt, consider using e README template for social science replication packages.
- Depending on the number of files in your dataset, consider having data and code in distinct dire

SE file in th at the Open h outputs m you identify s, as they c

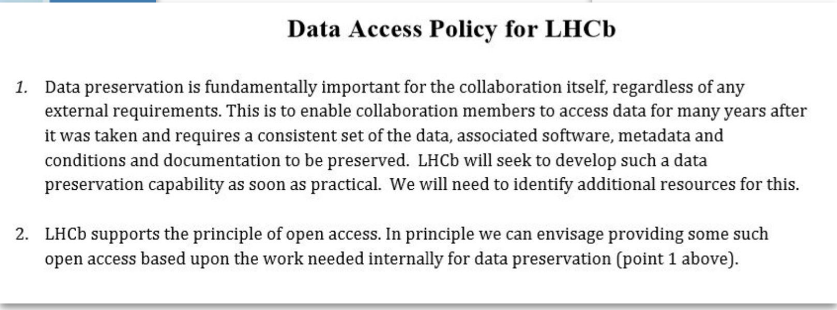


Dataverse Uploader Action

Dataverse Uploader

This action uploads the repository content to a Dataverse dataset.

the following input parameters:



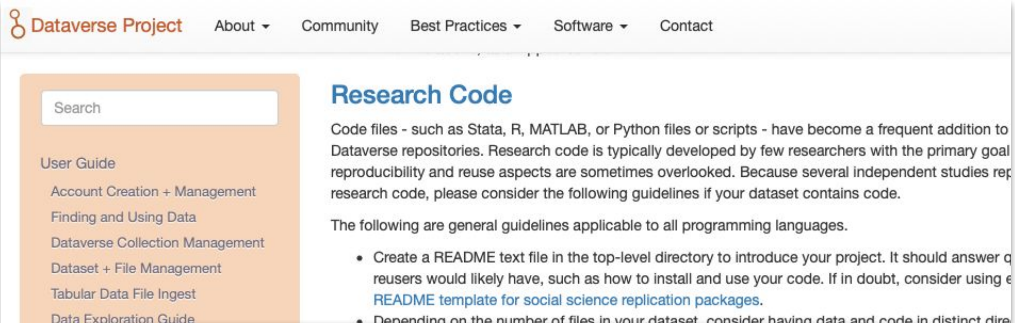
Data Access Policy for LHCb

1. Data preservation is fundamentally important for the collaboration itself, regardless of any external requirements. This is to enable collaboration members to access data for many years after it was taken and requires a consistent set of the data, associated software, metadata and conditions and documentation to be preserved. LHCb will seek to develop such a data preservation capability as soon as practical. We will need to identify additional resources for this.

2. LHCb supports the principle of open access. In principle we can envisage providing some such open access based upon the work needed internally for data preservation (point 1 above).

ALFRED P. SLOAN FOUNDATION

- Guidance
- Helper tools
- Policy



Research Code

Code files - such as Stata, R, MATLAB, or Python files or scripts - have become a frequent addition to Dataverse repositories. Research code is typically developed by few researchers with the primary goal reproducibility and reuse aspects are sometimes overlooked. Because several independent studies rep research code, please consider the following guidelines if your dataset contains code.

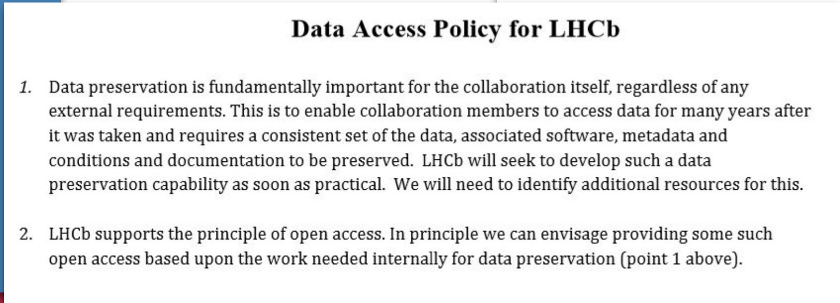
The following are general guidelines applicable to all programming languages.

- Create a README text file in the top-level directory to introduce your project. It should answer q reusers would likely have, such as how to install and use your code. If in doubt, consider using a README template for social science replication packages.
- Depending on the number of files in your dataset, consider having data and code in distinct dire

ISE file in th
at the Open
h outputs m
you identify
ns, as they c

a Dataverse dataset.

input parameters:



Data Access Policy for LHCb

1. Data preservation is fundamentally important for the collaboration itself, regardless of any external requirements. This is to enable collaboration members to access data for many years after it was taken and requires a consistent set of the data, associated software, metadata and conditions and documentation to be preserved. LHCb will seek to develop such a data preservation capability as soon as practical. We will need to identify additional resources for this.

2. LHCb supports the principle of open access. In principle we can envisage providing some such open access based upon the work needed internally for data preservation (point 1 above).

Email: anatrisovic@g.harvard.edu
GitHub & Twitter: atrisovic