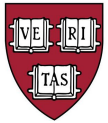


The Dataverse Project - Data sharing, Reproducibility, Research, Development and the Community

CNSTAT Expert meeting on Guidance on Data Sharing
for NIA Longitudinal Studies
May 10, 2021

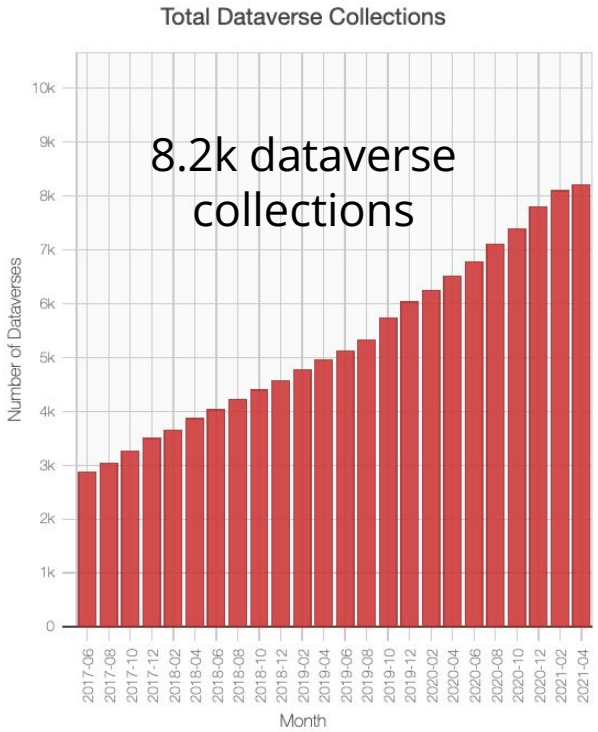
Ana Trisovic, Harvard University
on behalf of the Dataverse Project team

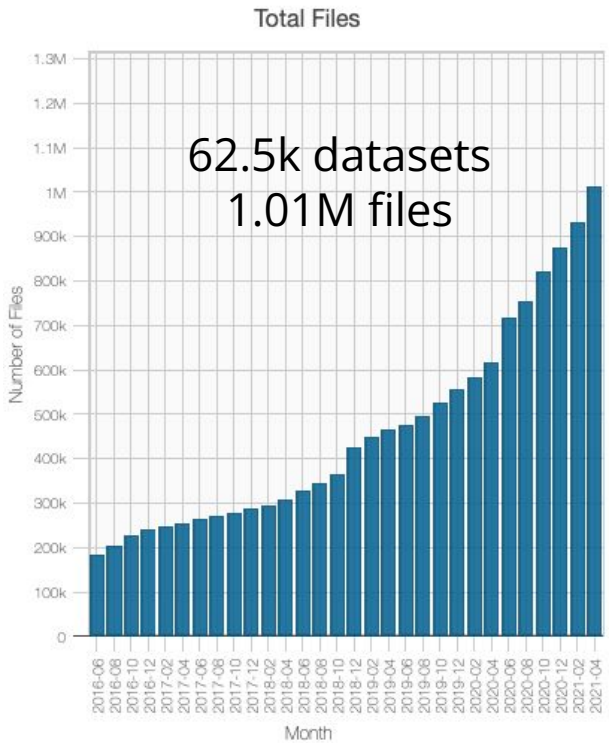
- A free and open-source software platform to archive, share, and cite research data
  - Focus on data sharing and making data available
- Provides data repository software that can be installed at institutions
  - Supports research communities for entire countries (NO, NL)
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community
  - 122 contributors to the software

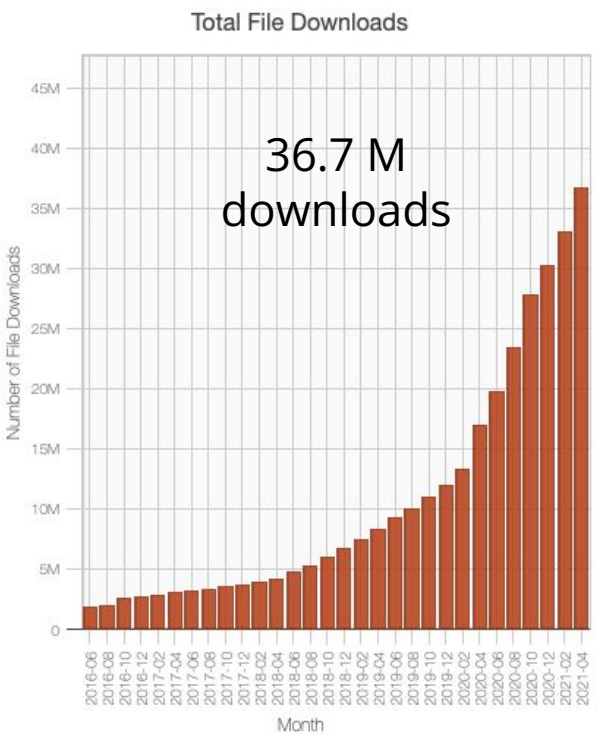
- A free and open-source software platform to archive, share, and cite research data
  - Focus on data sharing and making data available
- Provides data repository software that can be installed at institutions
  - Supports research communities for entire countries (NO, NL)
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) with contributions from the Dataverse community
  - 122 contributors to the software

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

Metrics







https://dataverse.org/metrics (29 Dataverse installations)

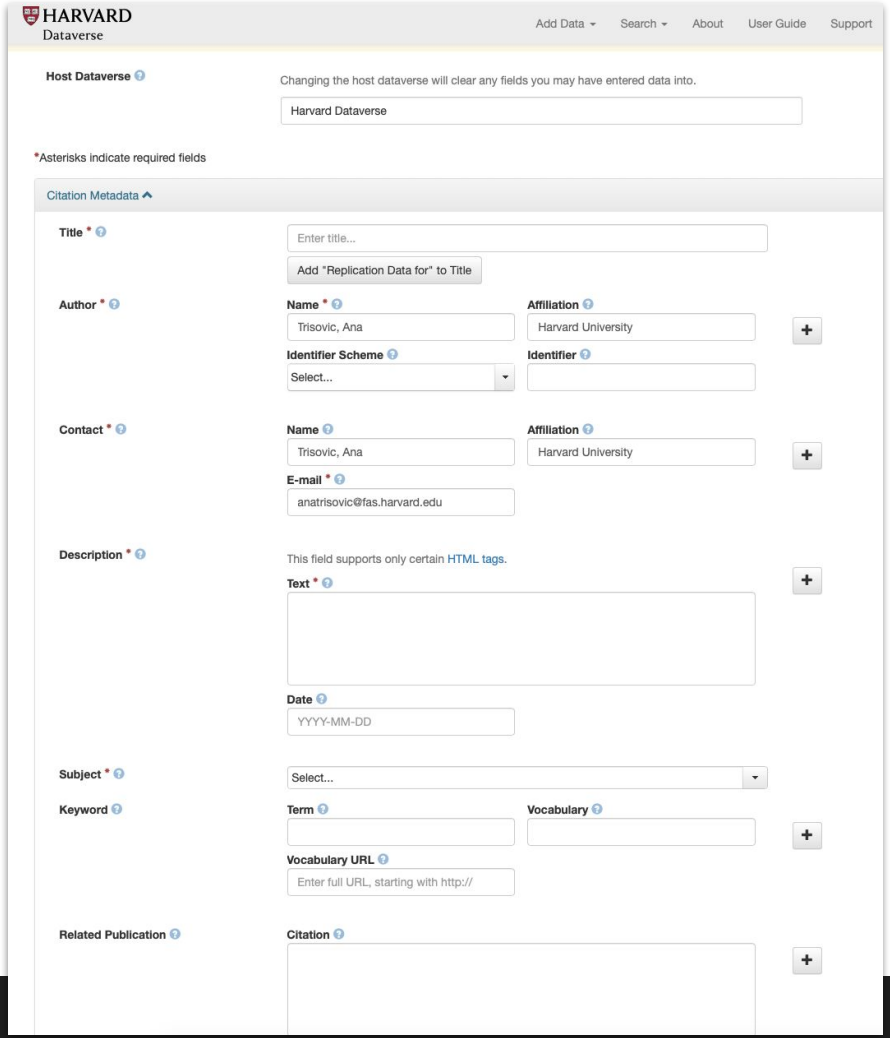
Presentation agenda

1. Data sharing at Dataverse repositories
2. Documentation and guidelines
3. Reproducibility and reuse
4. Open questions, research and development

Data sharing

Data sharing

- Stand-alone or institutional account for depositing data
  - Dataverse installation account
  - GitHub, Google, ORCID, University credentials
- Two approaches for data sharing:
  - Via UI in the web browser
  - Via API (command line or Dataverse Software clients)



Dataverse Software Clients

A client is an application that connects to and communicates with the remote Dataverse installation to transfer or manage data through the API.

pyDataverse

dataverse-client-javascript

dataverse-client- java

dataverse-client-python

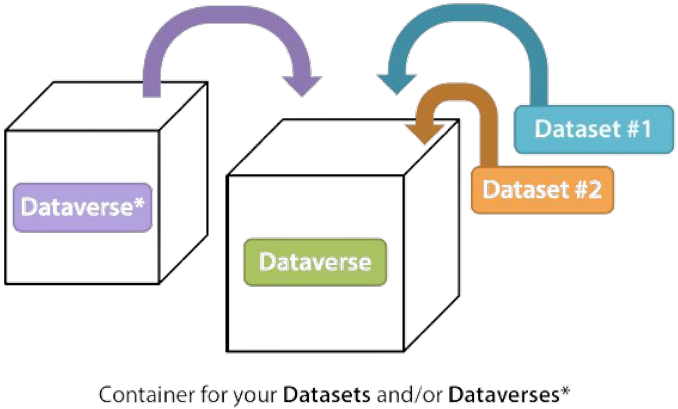
dataverse-client-r

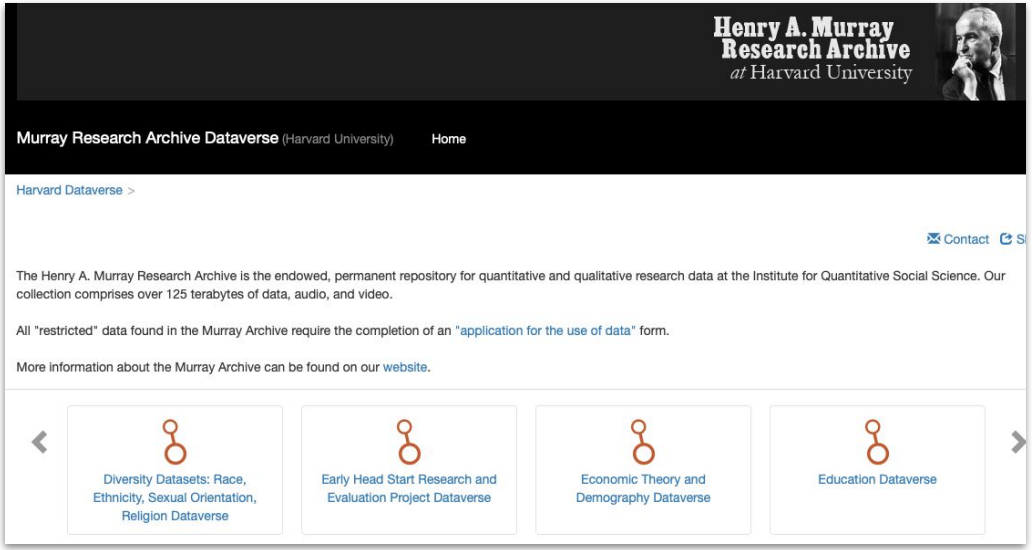
dataverse_api

All clients are open-source and available on GitHub.

Dataverse collection and datasets

- A dataverse collection is a collection of datasets and/or other collections
- Individuals, institutes or journals may have own dataverse collections

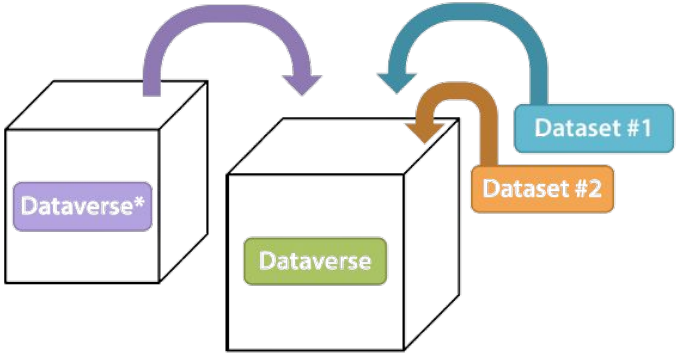


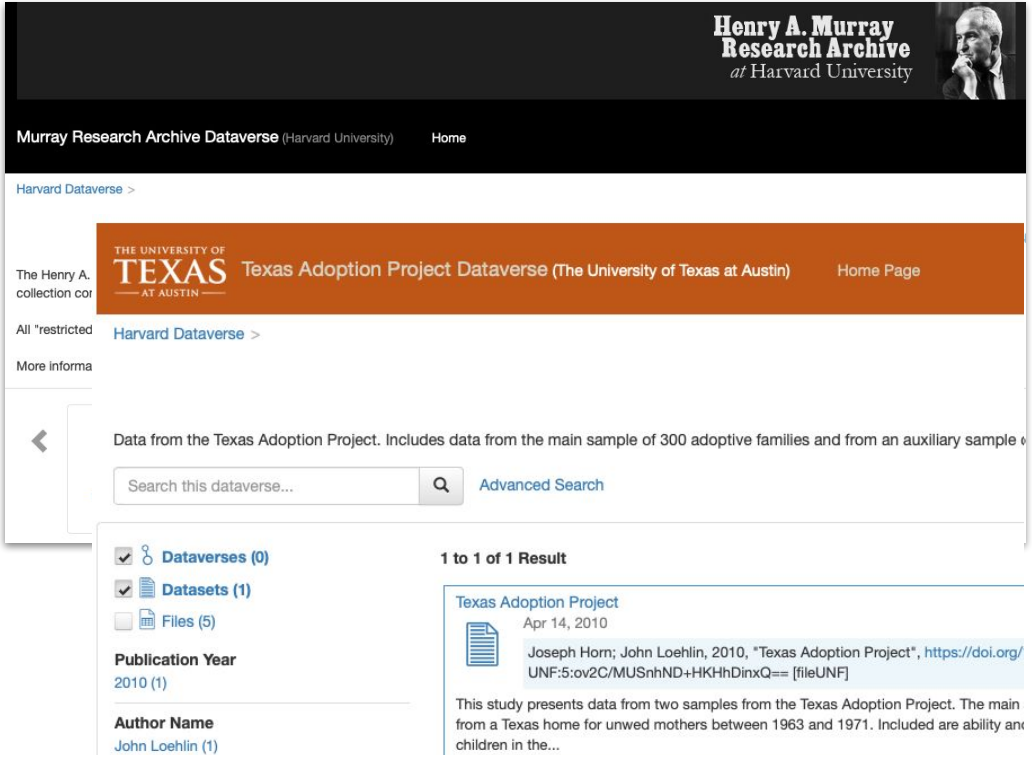


https://murray.harvard.edu

Dataverse collection and datasets

- A dataverse collection is a collection of datasets and/or other collections
- Individuals, institutes or journals may have own dataverse collections





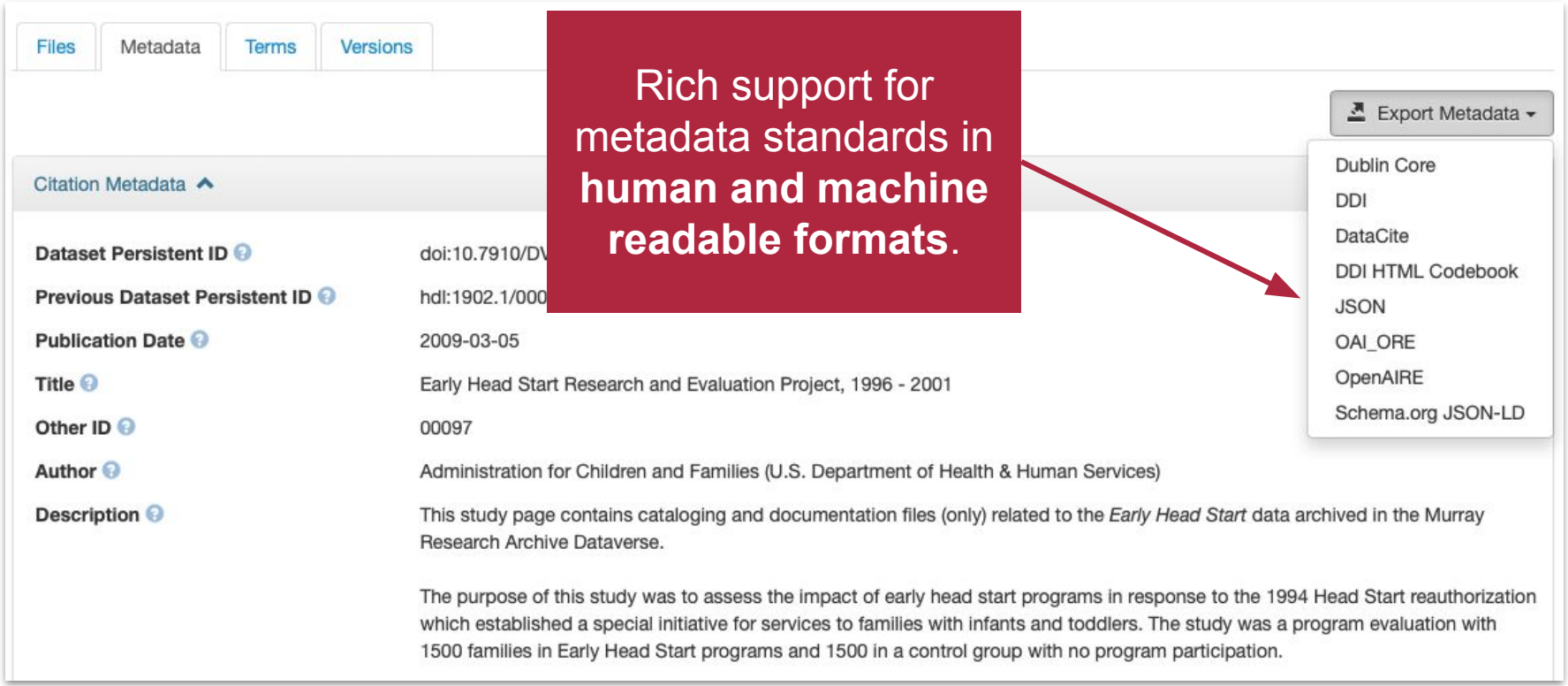
Replication dataset

- 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

Support for multiple metadata standards



Tabular files and variable metadata

ClimateRegressionData_150327.tab

This file is part of "Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.".

Version 1.0

File Citation

Albouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "ClimateRegressionData_150327.tab", *Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.*, https://doi.org/10.7910/DVN/QCE1XY/BNJLIA, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

Cite Data File ▾ Learn about Data Citation Standards.

Access File ▾ | Contact Owner | Share

File Metrics

105 Downloads

Dataset Citation

Albouy, David, Graf, Walter, Kellogg, Ryan, and Wolff, Hendrik, 2018, "Replication Data for: "Climate Amenities, Climate Change, and American Quality of Life" Journal of the Association of Environmental and Resource Economists 3, no. 1 (March 2016): 205-246.", https://doi.org/10.7910/DVN/QCE1XY, Harvard Dataverse, V1, UNF:6:CBIOoHJrG5/T6i+XjwBVwg== [fileUNF]

Cite Dataset ▾ Learn about Data Citation Standards.

Preview | Metadata | Versions

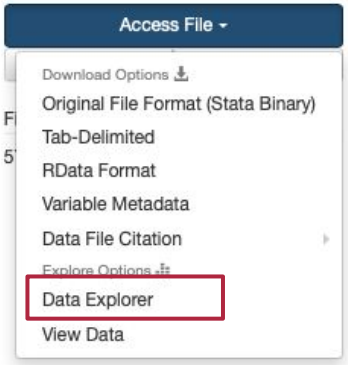
Explore on View Data

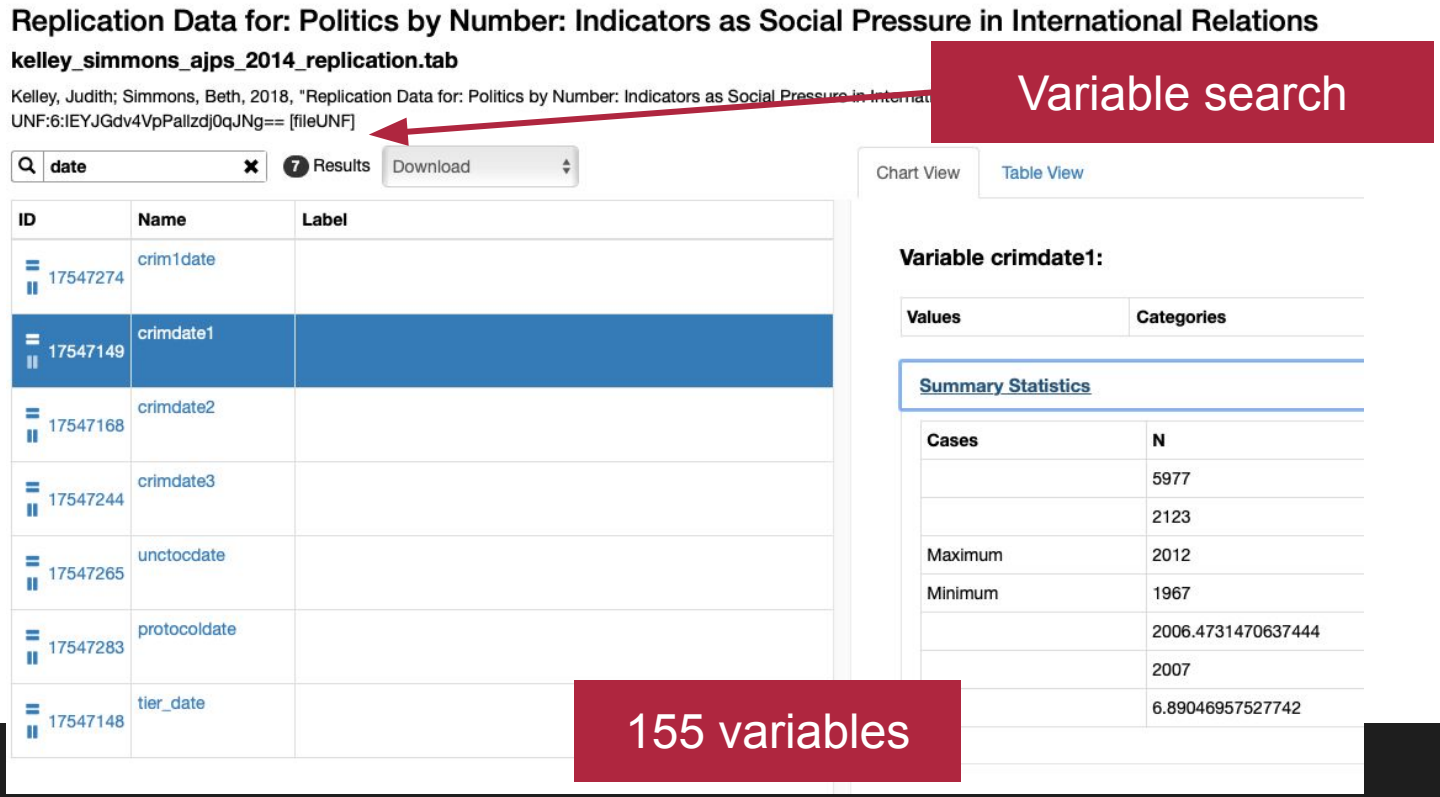
	msaname	Wage_orig	Wage	Price	QOL_orig	QOL_25_1	
47	Phoenix-Mesa, AZ	-0.0607973	0.014620701	-0.13433664	-0.013151387	0.027213141	-0
48	Phoenix-Mesa, AZ	-0.031059736	0.014049928	-0.20824265	-0.050257135	-0.021000927	-0
49	Phoenix-Mesa, AZ	0.13126019	0.025910802	0.1713764	-0.0038431883	-0.088416085	0.(
50	Phoenix-Mesa, AZ	0.083860636	0.024022257	0.08576128	-0.009372421	-0.062420316	0.(
51	Phoenix-Mesa, AZ	0.08740685	0.024419874	0.05565375	-0.020567477	-0.07349341	0.(

Variable metadata from tabular data files

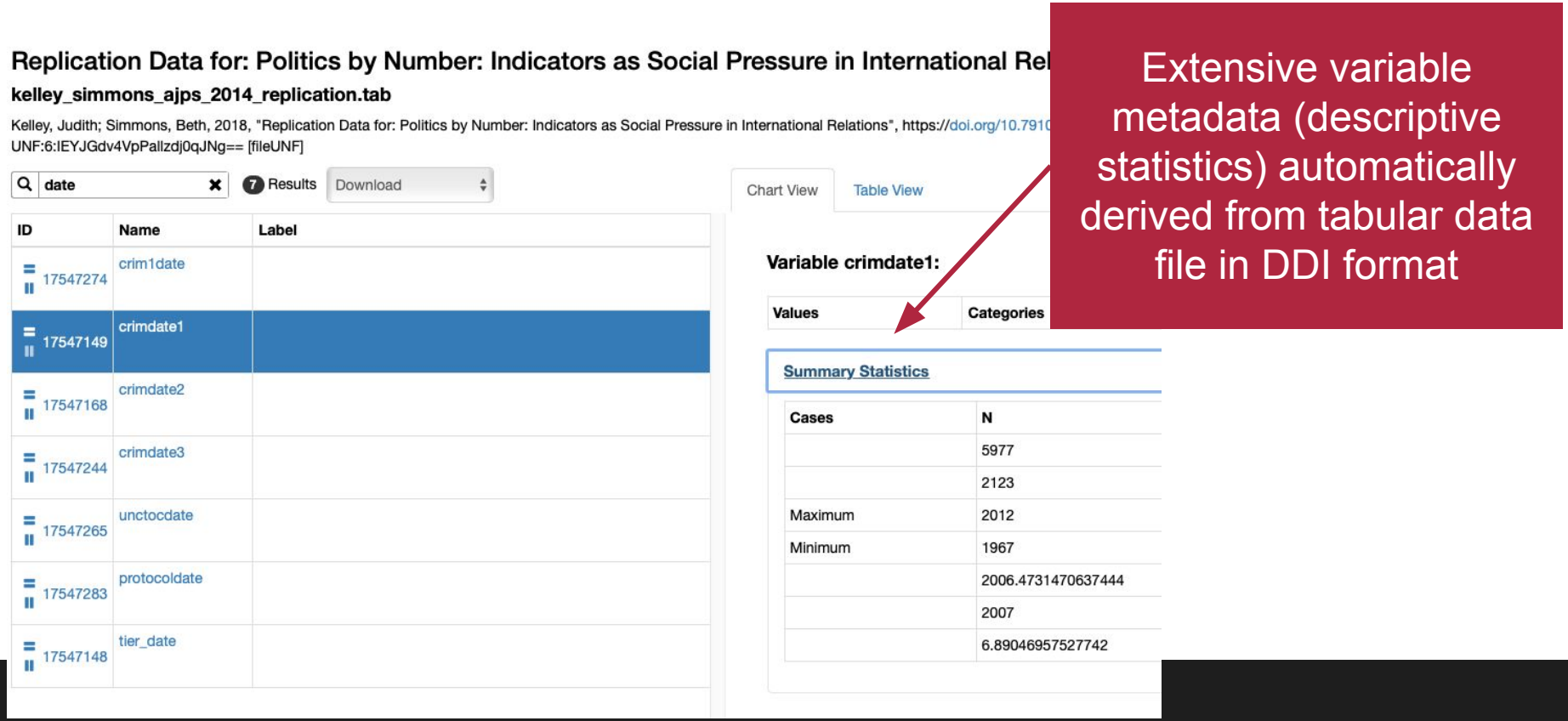
- var ..
  - @ID: v17547221
  - @name: gdpgrowthrate
  - @intrvl: contin
  - location
    - @fileid: f3002597
  - sumstat 106.27980041503906
    - @type: max
  - sumstat .
    - @type: mode
  - sumstat 5791.0
    - @type: vald
  - sumstat 6.563398040445294
    - @type: stdev
  - sumstat 3.8374838829040527
    - @type: medn
  - sumstat -51.030860900878906
    - @type: min
  - sumstat 2309.0
    - @type: invd
  - sumstat 3.6461661697240384
    - @type: mean
  - varformat
    - @type: numeric
    - @formatname: float
  - notes UNF:6:IWNGf1eqM40D8xQ
    - @subject: Universal Numeric
    - @level: variable
    - @type: VDC:UNF

Tabular files and variable metadata





Tabular files and variable metadata



Extensive variable metadata (descriptive statistics) automatically derived from tabular data file in DDI format

Data sharing

Summary

- Dataverse data repository software
- Versatile support for data sharing via UI and the clients
- Extensive support in the repository for data, variables and metadata

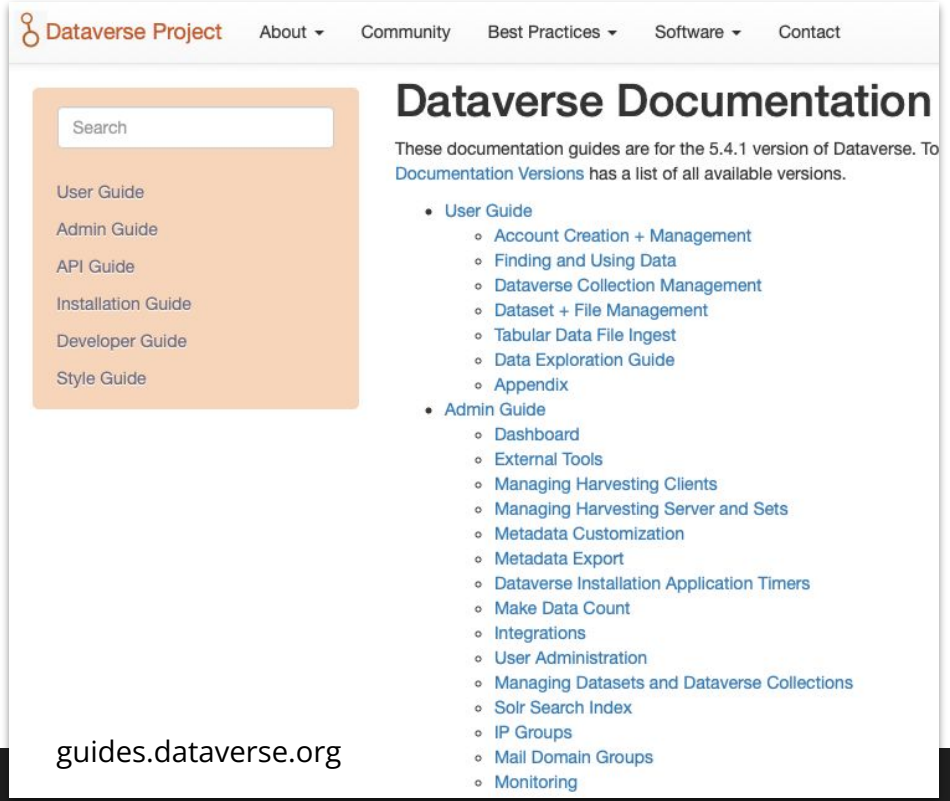
Documentation and guidelines

Documentation and guidelines

The Dataverse Project team maintains an extensive set of guidelines for repository managers, developers and users.

Dataverse Community meetings

- projects.iq.harvard.edu/dcm2021
- June 15-17, 2021



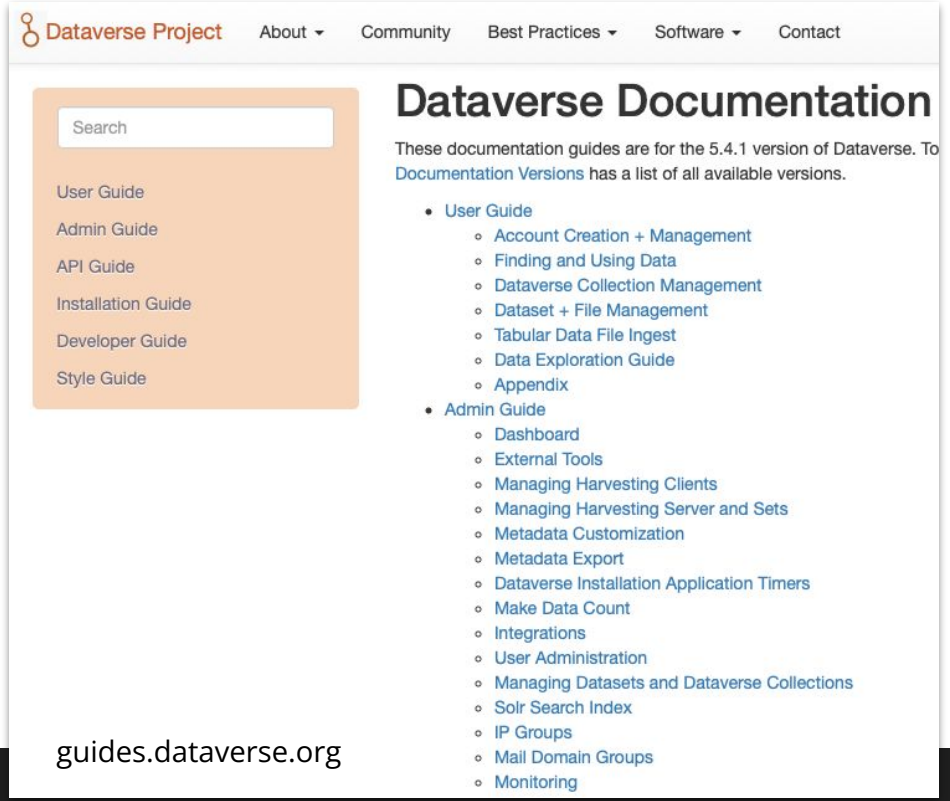
guides.dataverse.org

Documentation and guidelines

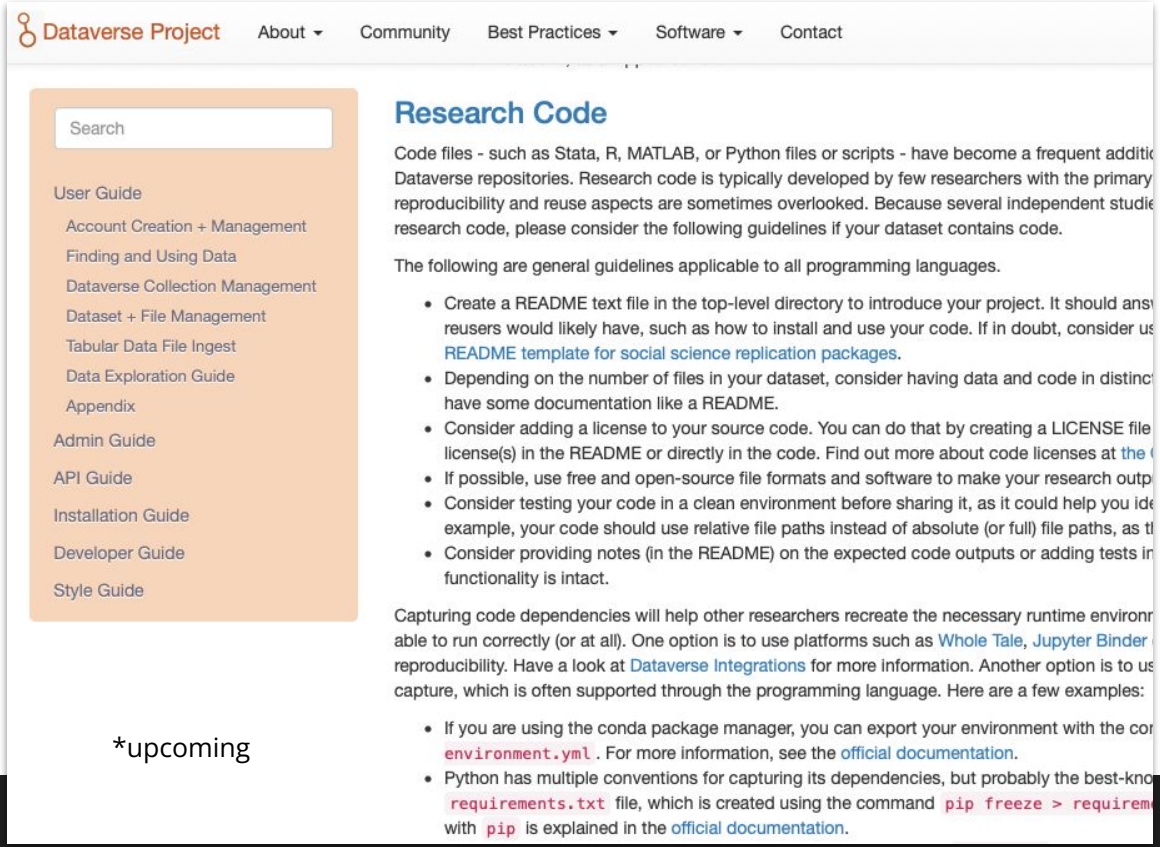
The Dataverse Project team maintains an extensive set of guidelines for repository managers, developers and users.

Dataverse Community meetings

- projects.iq.harvard.edu/dcm2021
- June 15-17, 2021



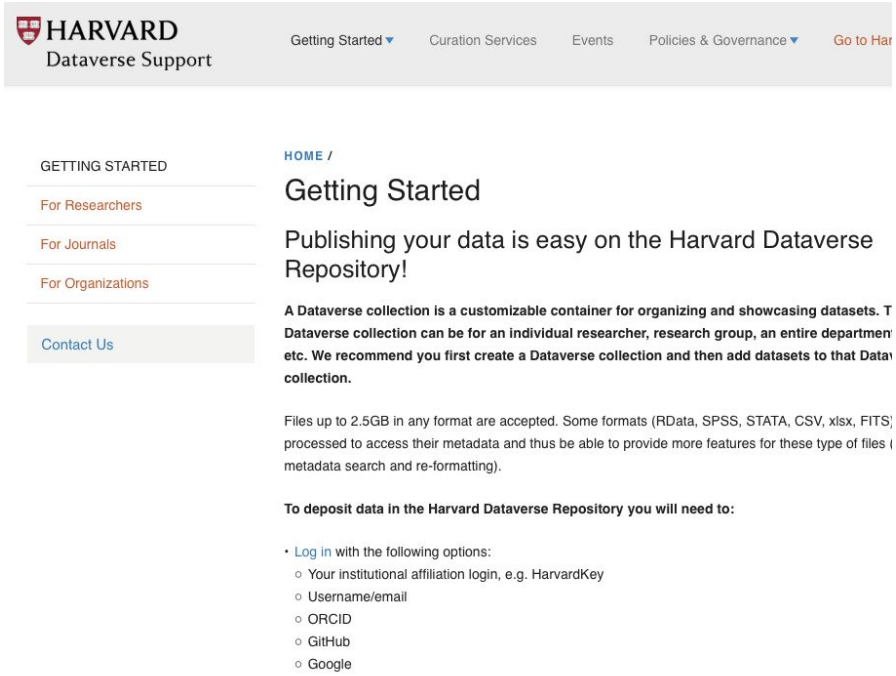
New guidelines



*upcoming

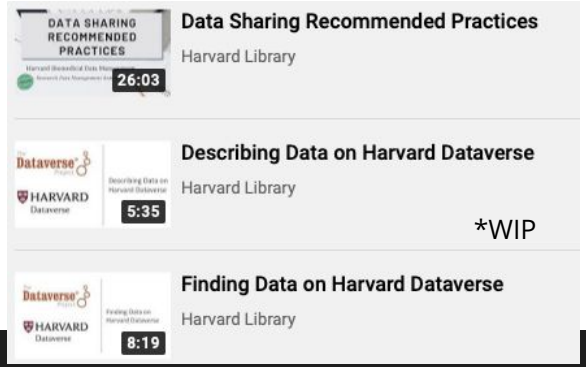
Documentation and guidelines

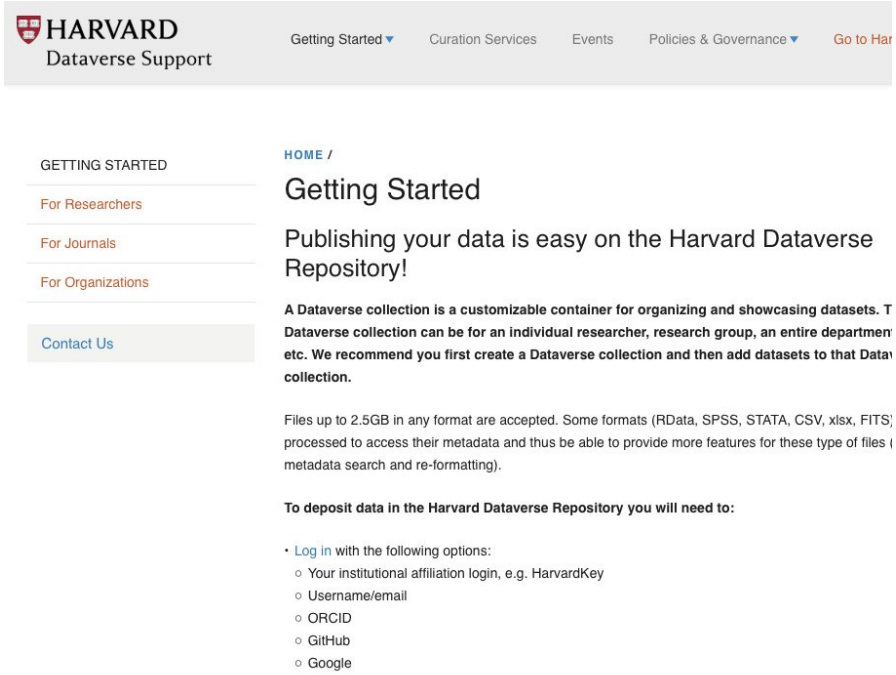
- Each Dataverse installation may have local documentation and support
- At Harvard, user training is organized by the Harvard Library and IQSS Data Curation Services



Documentation and guidelines

- Each Dataverse installation may have local documentation and support
- At Harvard, user training is organized by the Harvard Library and IQSS Data Curation Services
- Video guides on YouTube





Documentation and guidelines

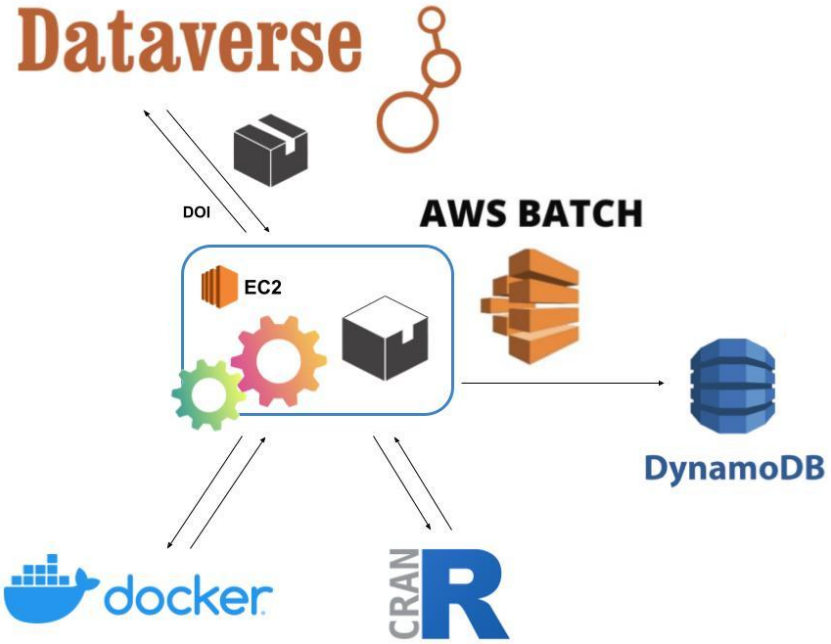
Summary

- Documentation and guidelines (global) are actively maintained
- Institutional guidelines, support and training

Reproducibility and reuse

Reproducibility and reuse

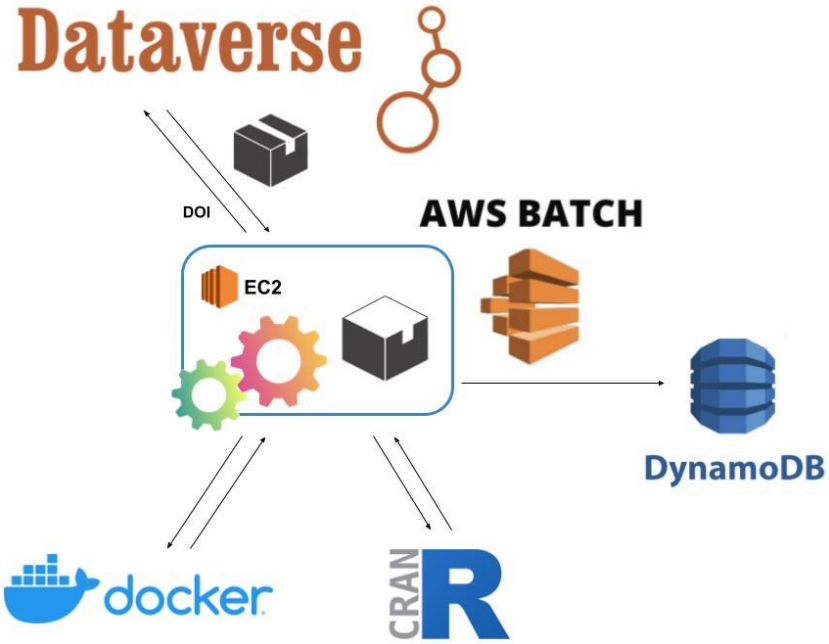
- Reproducibility: “obtaining consistent computational results using the same input data, steps, code, and conditions of analysis”
- Conducted large scale analysis to examine execution and quality of R and Python code
- We retrieved replication datasets from Harvard Dataverse repository and R code was re-executed in a pre-installed Docker container



National Academies of Sciences, Engineering, and Medicine. Reproducibility and replicability in science. National Academies Press, 2019.
Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv preprint arXiv:2103.12793 (2021).

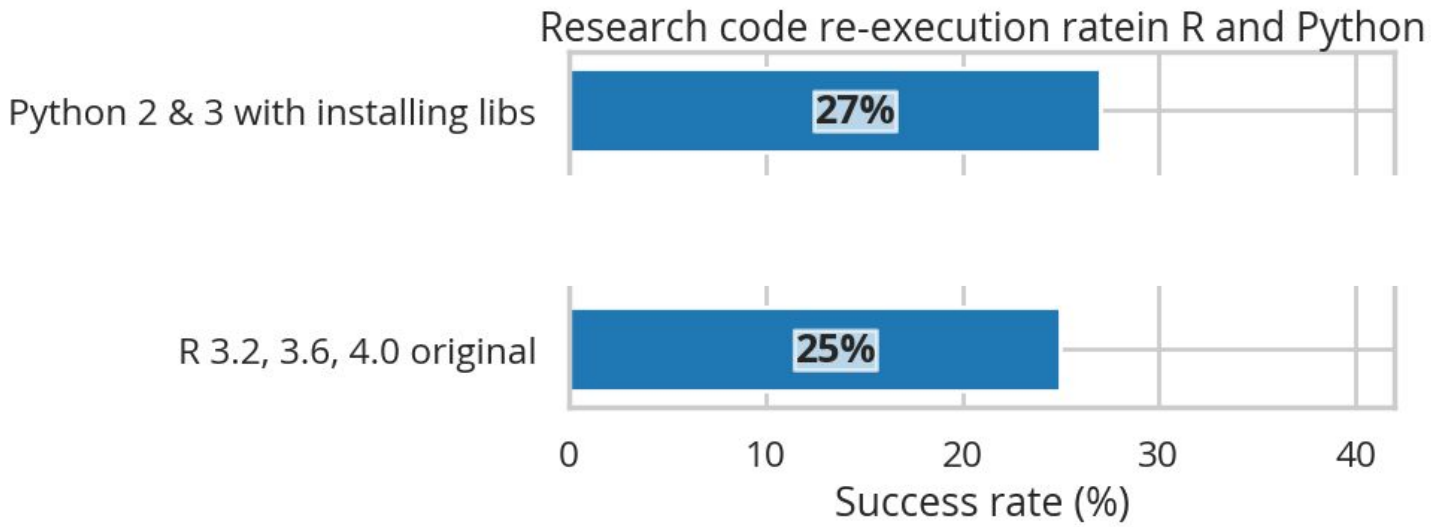
Reproducibility and reuse

- Reproducibility: “obtaining consistent computational results using the same input data, steps, code, and conditions of analysis”
- Conducted large scale analysis to examine execution and quality of R and Python code
- We retrieved replication datasets from Harvard Dataverse repository and R code was re-executed in a pre-installed Docker container



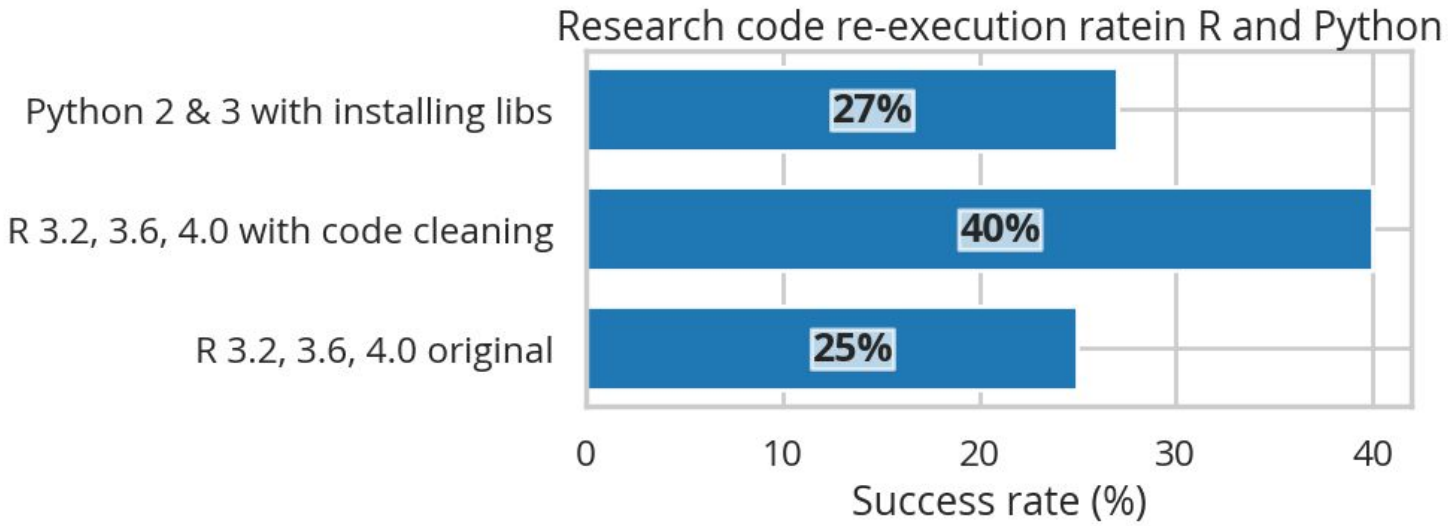
National Academies of Sciences, Engineering, and Medicine. Reproducibility and replicability in science. National Academies Press, 2019.
Trisovic, Ana, et al. "A large-scale study on research code quality and execution." arXiv preprint arXiv:2103.12793 (2021).

Reproducibility and reuse



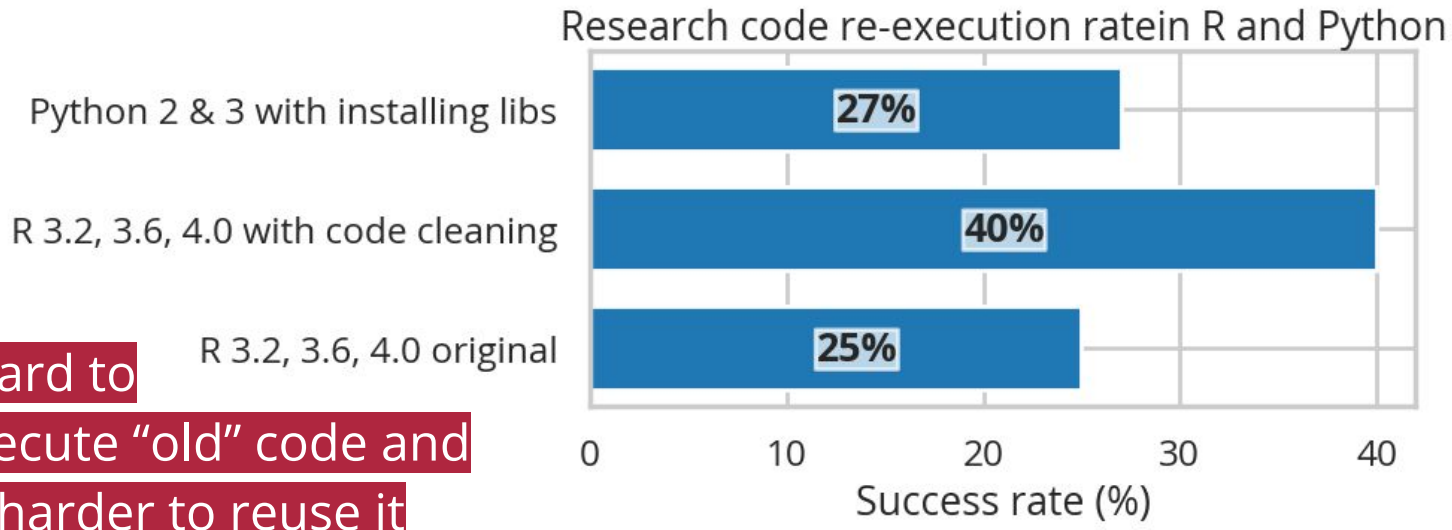
Trisovic, Ana, et al. "Repository Approaches to Improving Quality of Shared Data and Code." Data 6.2 (2021): 15.

Reproducibility and reuse



Trisovic, Ana, et al. "Repository Approaches to Improving Quality of Shared Data and Code." Data 6.2 (2021): 15.

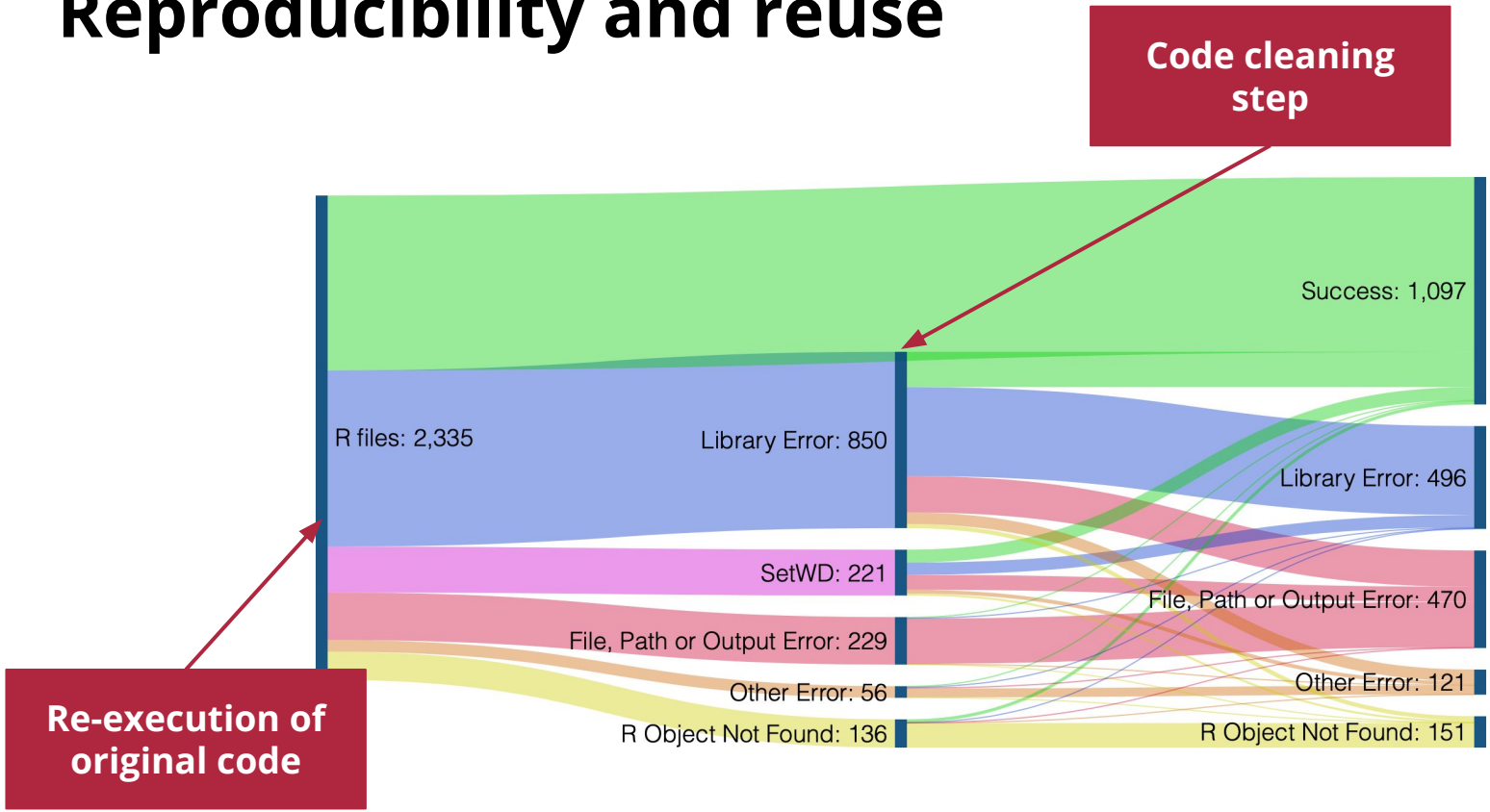
Reproducibility and reuse



It is hard to re-execute “old” code and even harder to reuse it

Trisovic, Ana, et al. "Repository Approaches to Improving Quality of Shared Data and Code." Data 6.2 (2021): 15.

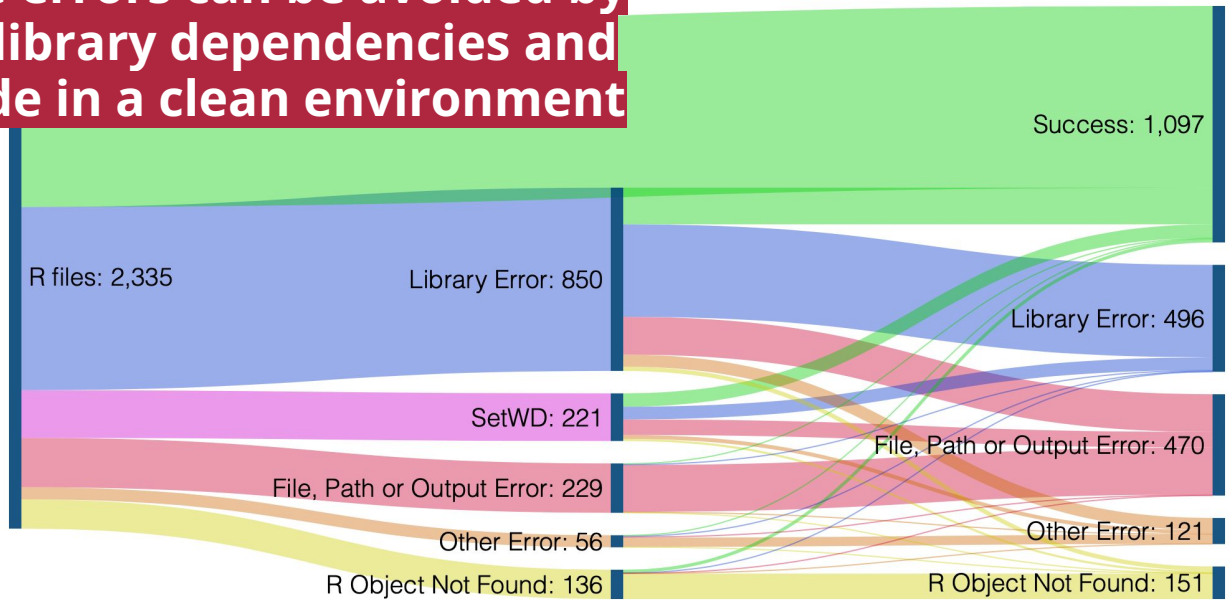
Reproducibility and reuse



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

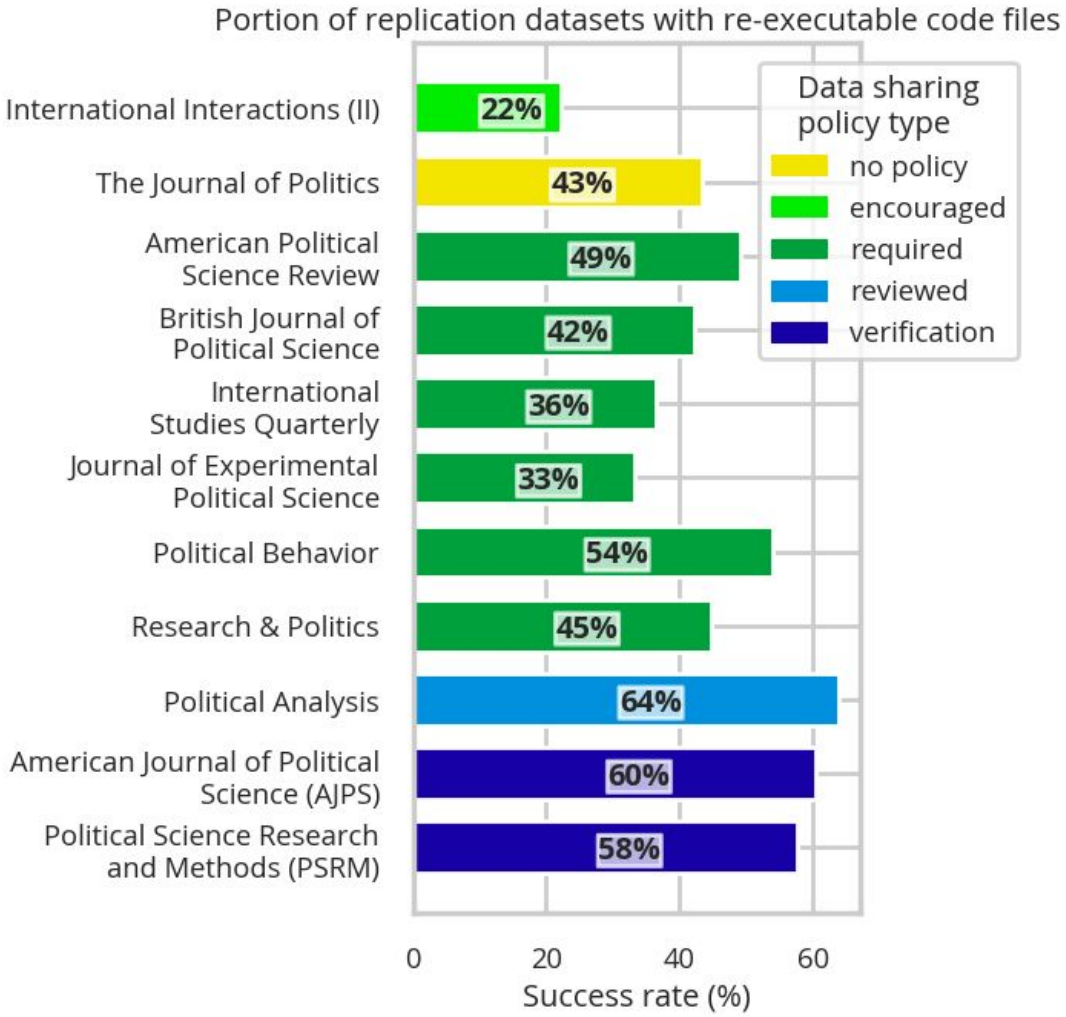
Reproducibility and reuse

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

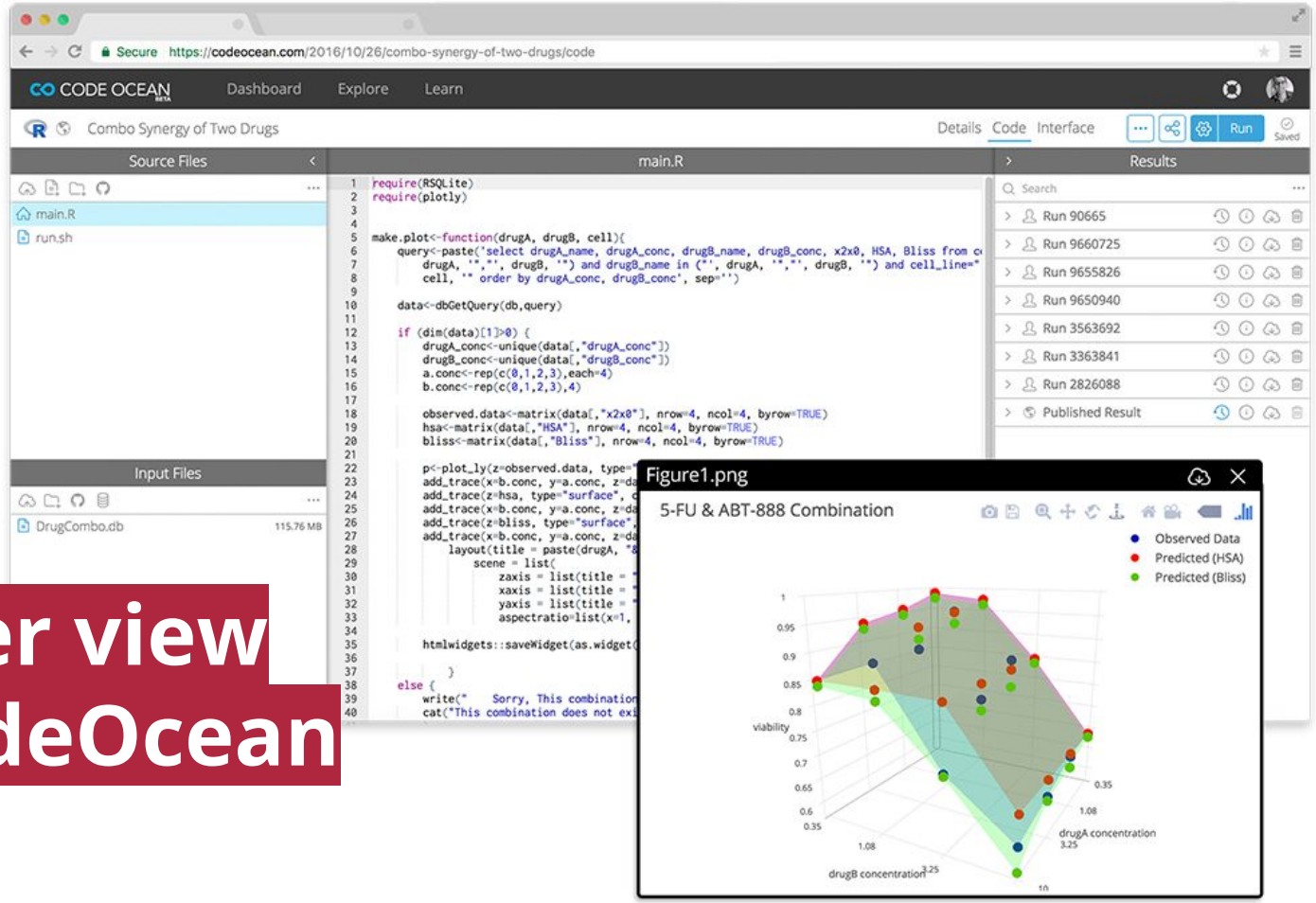


Integration with reproducibility platforms

- New cloud tools have emerged to support collaborative work and research reproducibility by capturing code dependencies inside a web browser
- Dataverse software integration with reproducibility platforms allows:
  - Importing new research data with necessary libraries
  - Exporting and reusing the existing data

Trisovic, Ana, et al. "Advancing computational reproducibility in the Dataverse data repository platform." P-RECS'20.

In-browser view of the CodeOcean platform



codeocean.com

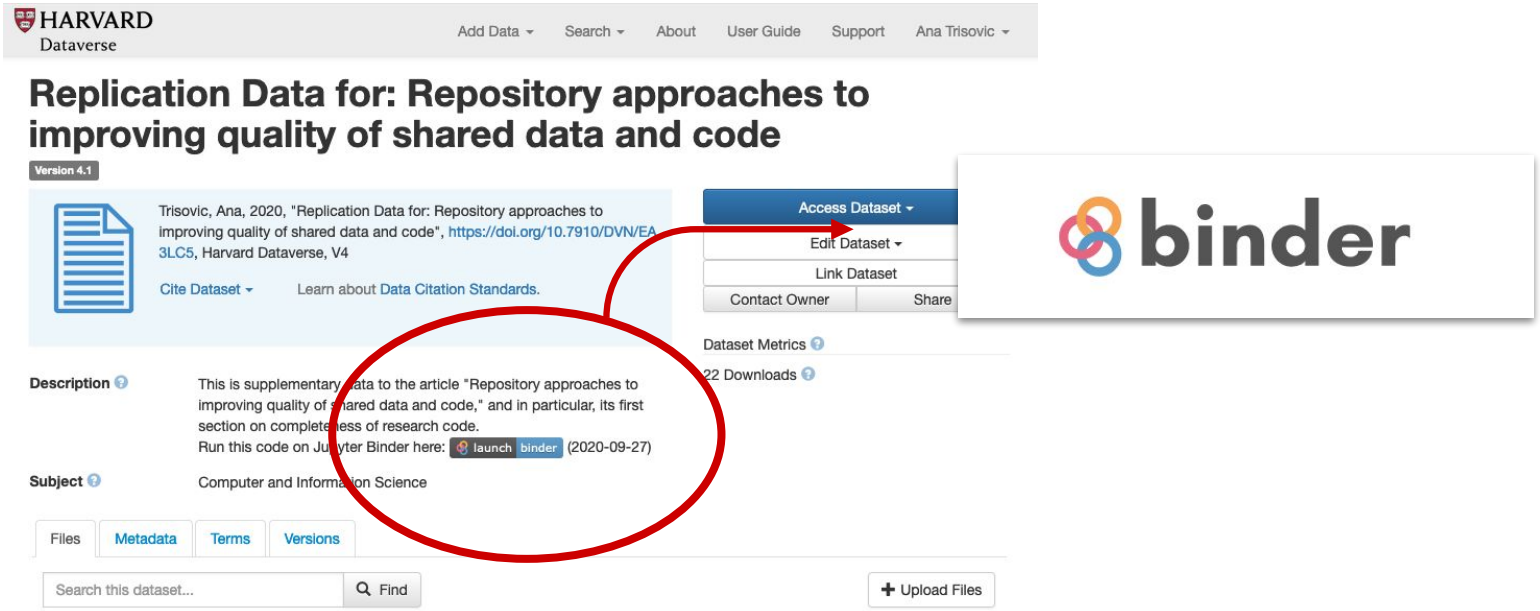
Figure credit: tech.cornell.edu/news

Integration with reproducibility platforms

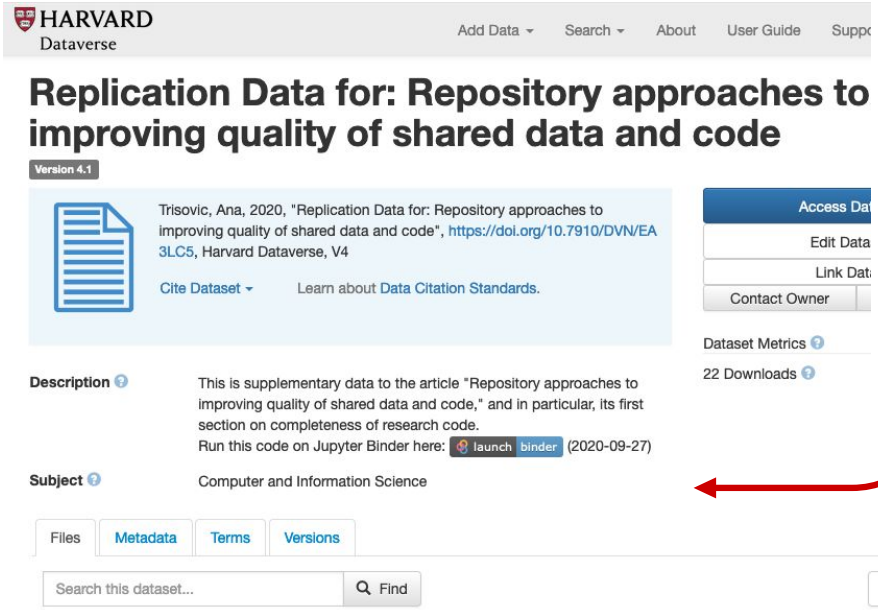
- New cloud tools have emerged to support collaborative work and research reproducibility by capturing code dependencies inside a web browser
- Dataverse software integration with reproducibility platforms allows:
  - Importing new research data with necessary libraries
  - Exporting and reusing the existing data

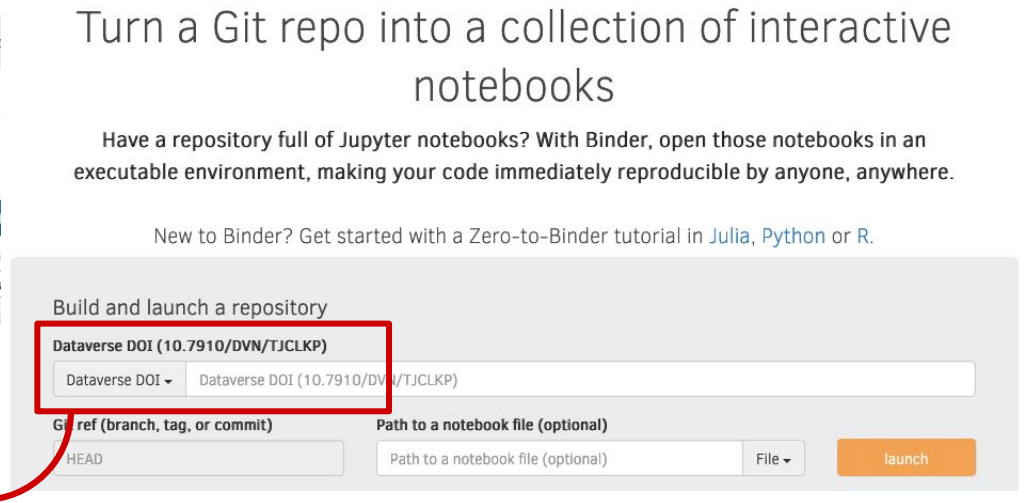
Trisovic, Ana, et al. "Advancing computational reproducibility in the Dataverse data repository platform." P-RECS'20.

Example: Jupyter Binder and Harvard Dataverse repository



Example: Jupyter Binder and Harvard Dataverse repository

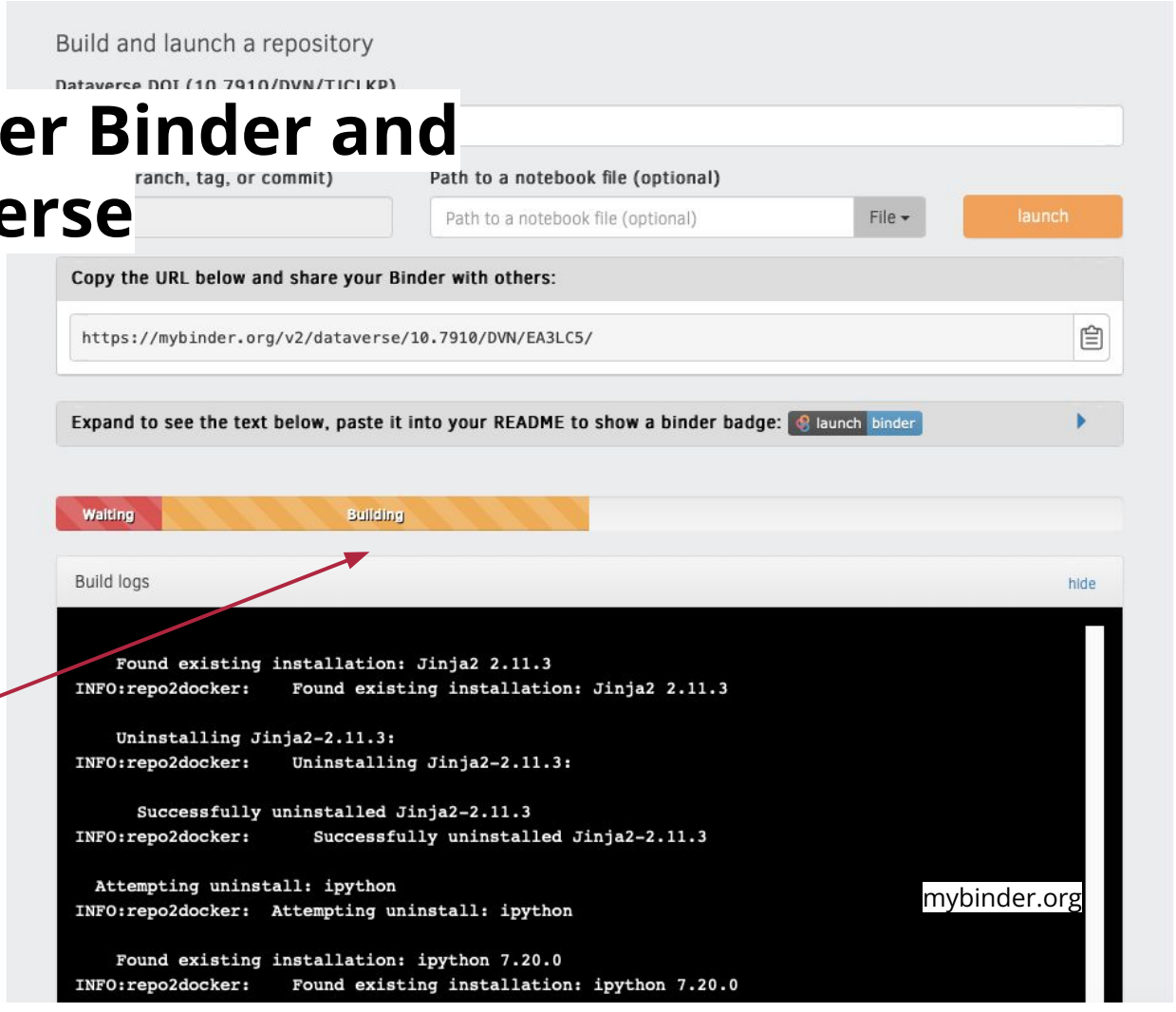




mybinder.org

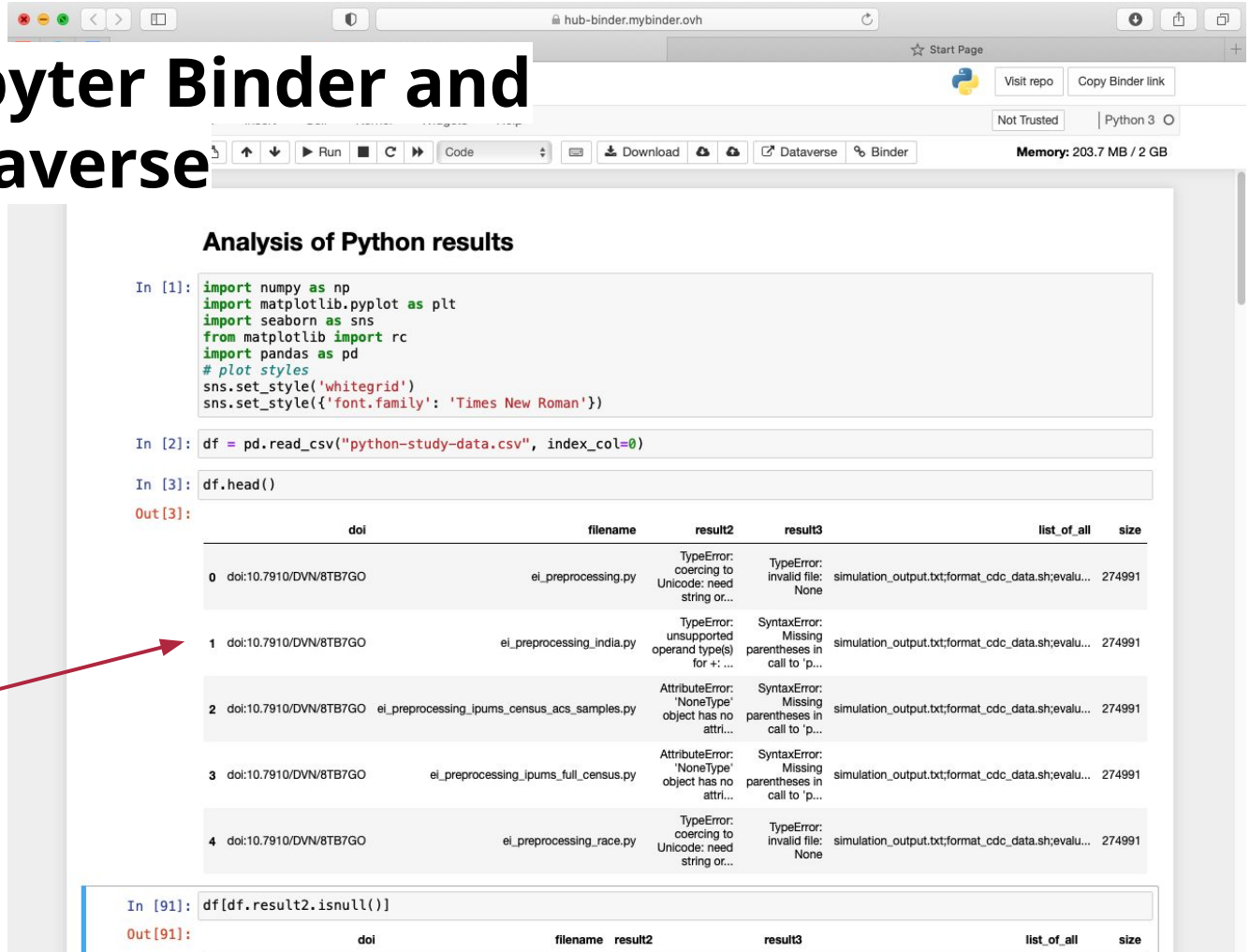
Example: Jupyter Binder and Harvard Dataverse repository

Jupyter Binder creates runtime environment for research code

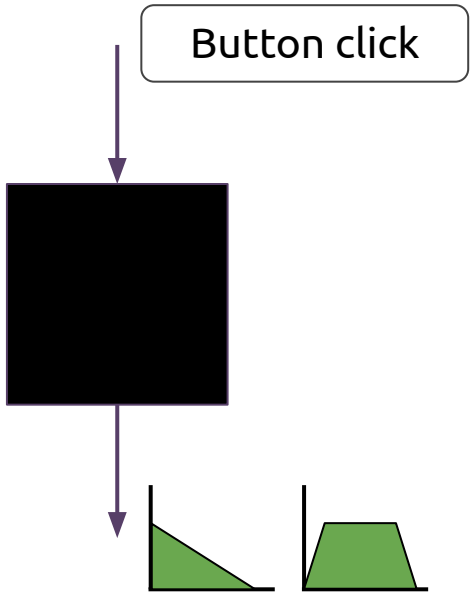


Example: Jupyter Binder and Harvard Dataverse repository

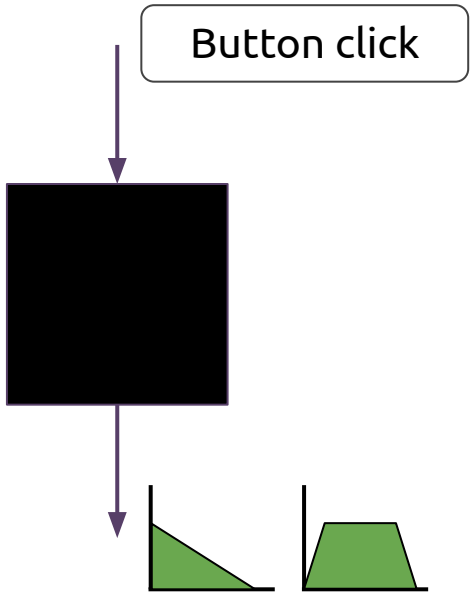
Anyone can explore data, make direct changes in the code or data

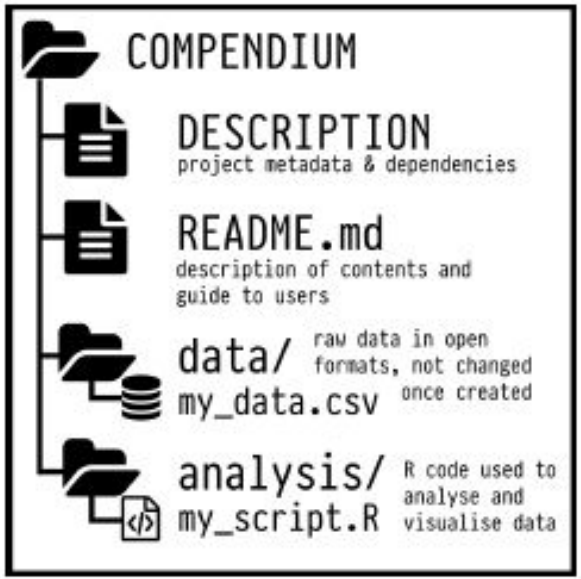


Reproducibility versus reuse

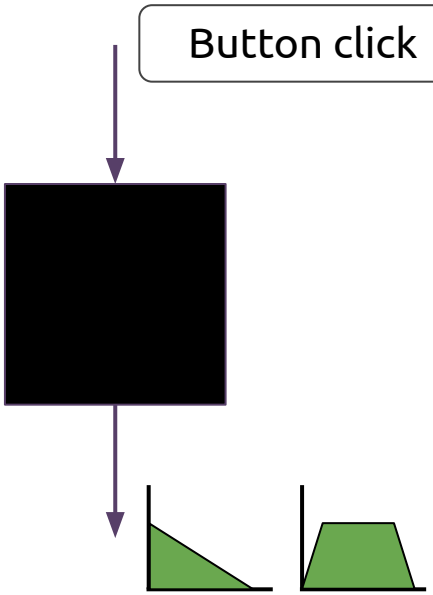


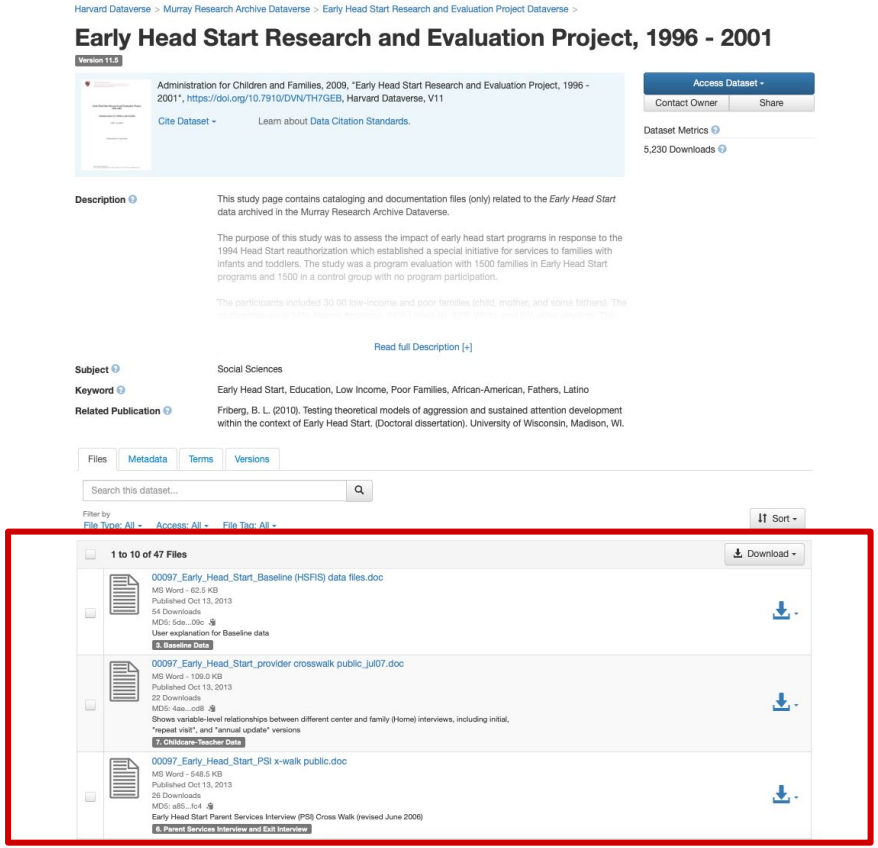
Reproducibility versus reuse





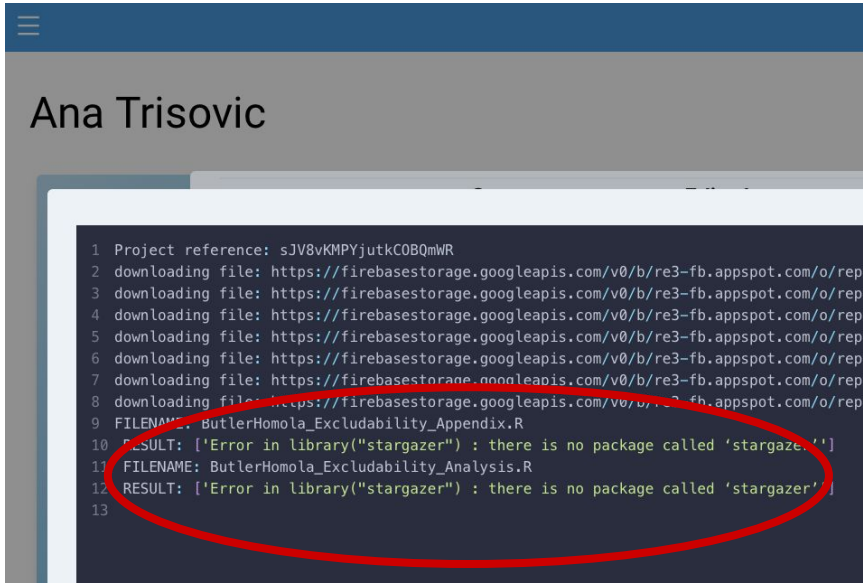
Reproducibility versus reuse





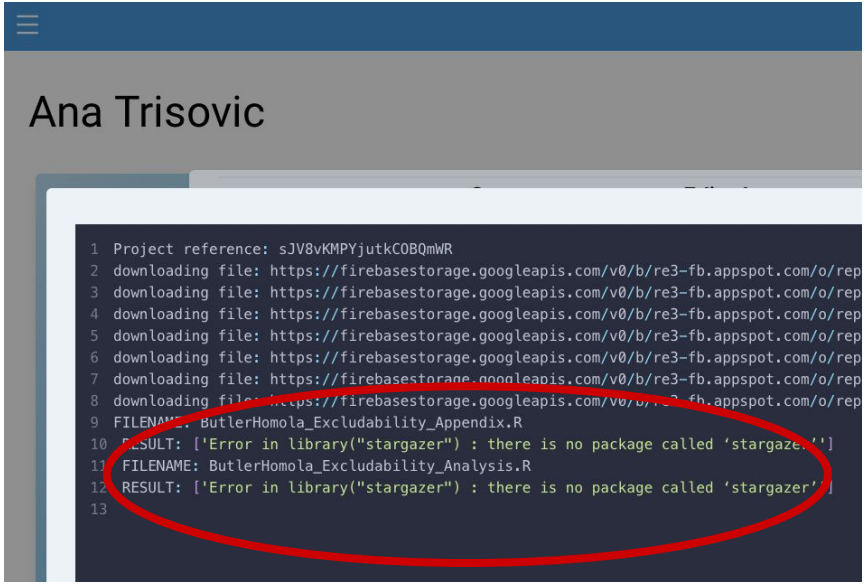
Research and prototype projects

Research reproducibility test at upload

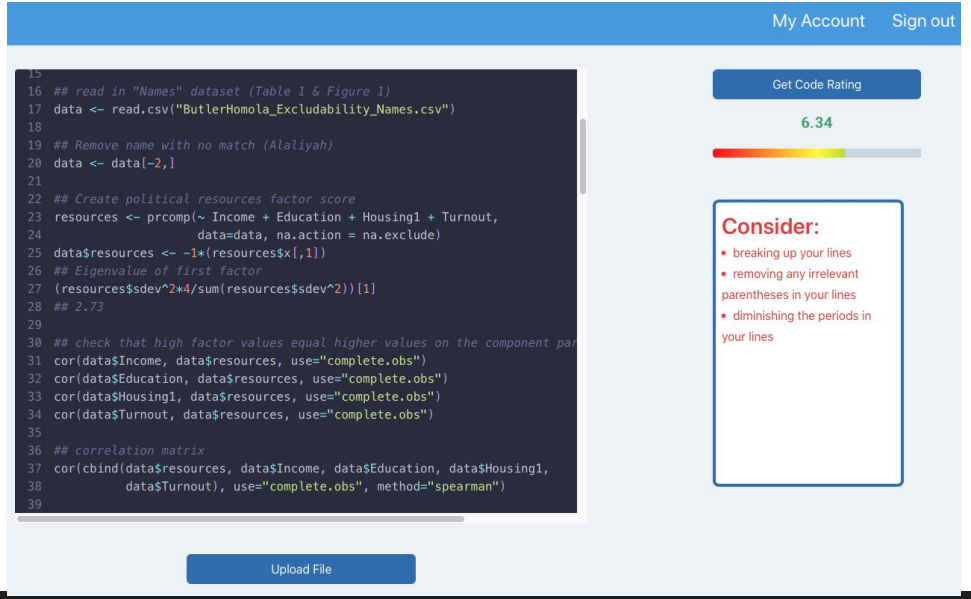


Research and prototype projects

Research reproducibility test at upload



Code readability assessment



Bahaidarah, Layan, et al. "Toward Reusable Science with Readable Code and Reproducibility." (to appear)

Reproducibility and reuse

Summary

- A large-scale study sheds light on reproducibility challenges
- Integrations with cloud tools facilitate software capture
- Further development focuses on reuse

Open questions, research and development

Open questions, research and development

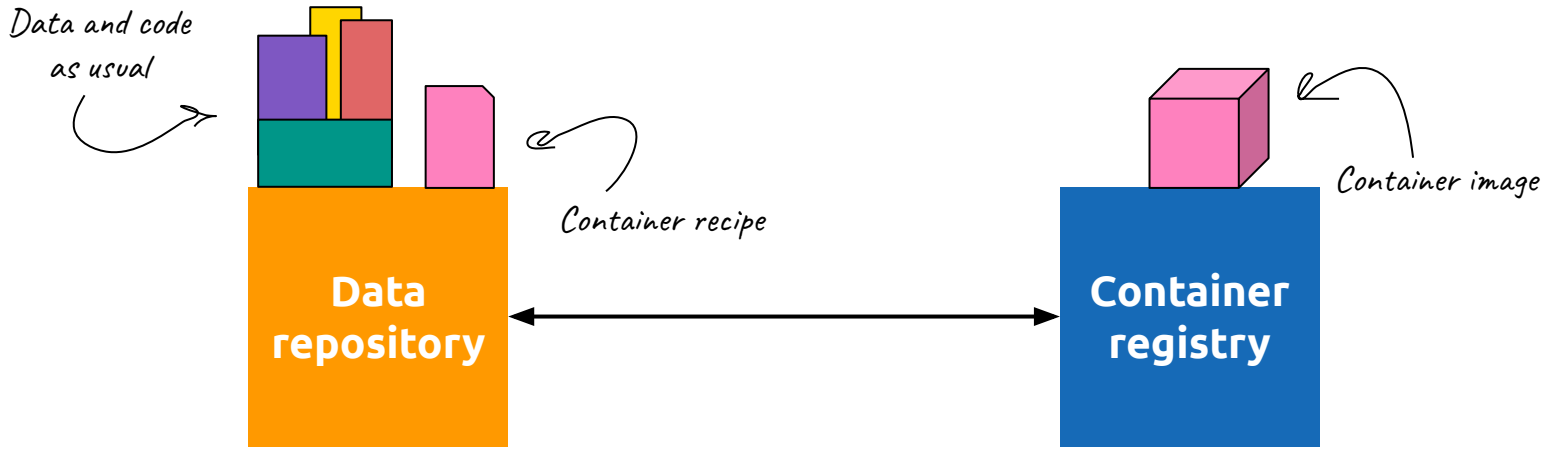
- DDI – Cross Domain Integration (DDI-CDI)
  - Planning to implement cross-domain metadata to facilitate merging data files
  - Controlled vocabularies for each variable
- Will incorporate software metadata
  - Enabling licencing, attribution and software dependency
- Global Dataverse Community Consortium (GDCC) - facilitates Dataverse community efforts, and supports Dataverse repositories around the world
  - GDCC working groups are formed to tackle the open questions

Open questions, research and development

- DDI – Cross Domain Integration (DDI-CDI)
  - Planning to implement cross-domain metadata to facilitate merging data files
  - Controlled vocabularies for each variable
- Will incorporate software metadata
  - Enabling licencing, attribution and software dependency
- Global Dataverse Community Consortium (GDCC) - facilitates Dataverse community efforts, and supports Dataverse repositories around the world
  - GDCC working groups are formed to tackle the open questions

Software, workflows and containers working group

- SWC (swc.gdcc.io) - identifies requirements and undertakes developments to support research software and reproducibility at the Dataverse Project
  - We welcome new members!
  - One solution for containers:

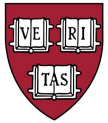


Conclusion

- The Dataverse Project creates **research data repository** software
- Ongoing support and ongoing enhancements for descriptive metadata for datasets and variable metadata for tabular files
- Research is conducted to:
  - **examine** common practices,
  - **identify** shortcomings **of existing approaches and**
  - new software developments
- Further development include supporting dissemination of emerging computational components, such as **workflows** and **containers**, with specific metadata, tools and infrastructure

Conclusion

- The Dataverse Project creates **research data repository** software
- Ongoing support and ongoing enhancements for descriptive metadata for datasets and variable metadata for tabular files
- Research is conducted to:
  - **examine** common practices,
  - **identify** shortcomings **of existing approaches and**
  - new software developments
- Further development include supporting dissemination of emerging computational components, such as **workflows** and **containers**, with specific metadata, tools and infrastructure



Thank you!

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