Repository Approaches to Improving Quality of Shared Data and Code

Ana Trisovic, Katherine Mika, Ceilyn Boyd, Sebastian Feger and Mercè Crosas

Introduction

- Researchers share data, code and other materials to enable research transparency, reproducibility and verification
  - More importantly, this material should be reusable for students, early-career researchers and others who want to build on it
- In practice, this is often not the case
  - Shared material is not always well-documented, understandable, and reusable
  - A reproducibility crisis has been reported

Our Idea

- Data repositories as a primary venue for research data sharing
- How can data repositories improve data and code quality, and how can they signal data and code quality to external researchers?

Our Idea

- Data repositories as a primary venue for research data sharing
- How can data repositories improve data and code quality, and how can they signal data and code quality to external researchers?
- Defining data quality
- Diverse approaches
  - Approach #1: curation features
  - Approach #2: code completeness
  - Approach #3: gamified design elements
- Approaches in practice: a Dataverse use-case

Dataverse

- Harvard Dataverse is a multi-disciplinary research data repository that allows members of the worldwide scientific community to deposit, publish, and share their datasets
- Dataverse repositories are based on the open-source software

Dataverse

- Harvard Dataverse is a multi-disciplinary research data repository that allows members of the worldwide scientific community to deposit, publish, and share their datasets
- Dataverse repositories are based on the open-source software

Defining Data Quality

Cai & Zhu	Martin et al.
Availability: accessibility, timeliness, authorization	Accessibility, timeliness, representational consistency, visibility, user-friendliness, platform functionality
Usability: definition/documentation, credibility, metadata	Intended use, subject matter expertise, technical skills, metadata quality (standards & consistency)
Reliability: accuracy, integrity, consistency, completeness, auditability	Data accuracy, validity, reliability, completeness, missing data, collection methods, format & layout, size etc.
Relevance: fitness	Relevancy, value added
Presentation quality: readability & structure	Concise representation, ease of understanding, ease of manipulation
	Platform promotion and user training: availability of information, responding to feedback, financial resources

Defining Data Quality

Cai & Zhu	Martin et al.	Examples of data repository features & functionalities
Availability: accessibility, timeliness, authorization	Accessibility, timeliness, representational consistency, visibility, user-friendliness, platform functionality	Capturing data citation information, minting DOIs
Usability: definition/documentation, credibility, metadata	Intended use, subject matter expertise, technical skills, metadata quality (standards & consistency)	Supporting documentation, reuse licensing, terms of access/restrictions
Reliability: accuracy, integrity, consistency, completeness, auditability	Data accuracy, validity, reliability, completeness, missing data, collection methods, format & layout, size etc.	Metadata standards, variable level metadata indexing
Relevance: fitness	Relevancy, value added	Reuse metrics, preview options, granular description
Presentation quality: readability & structure	Concise representation, ease of understanding, ease of manipulation	CURE or services that advise on data publishing
	Platform promotion and user training: availability of information, responding to feedback, financial resources	Support services, preservation policies, governance and legal policies

Defining Data Quality

- While some of the quality dimensions refer to intrinsic qualities of data files such as accuracy, integrity, and completeness, several important features can be improved by data repositories.
- These features include:
  - presentation quality,
  - documentation,
  - metadata, and
  - accessibility

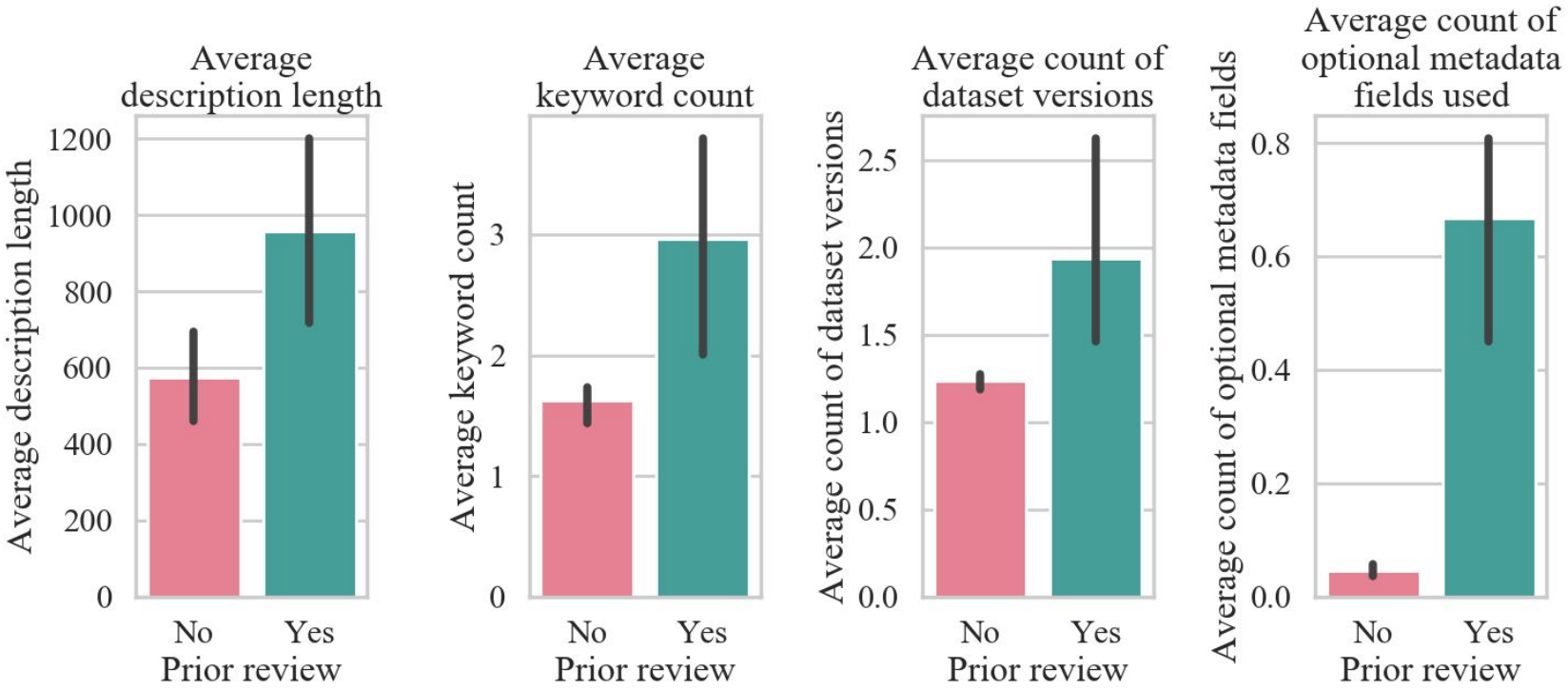
Approach #1: Encourage Use of Curation Features

- Groups such as journals, laboratories, and project teams curate their data collections on Dataverse
- The Dataverse software supports review workflows that allow curators to ensure that deposited datasets meet group-defined expectations

Approach #1: Encourage Use of Curation Features

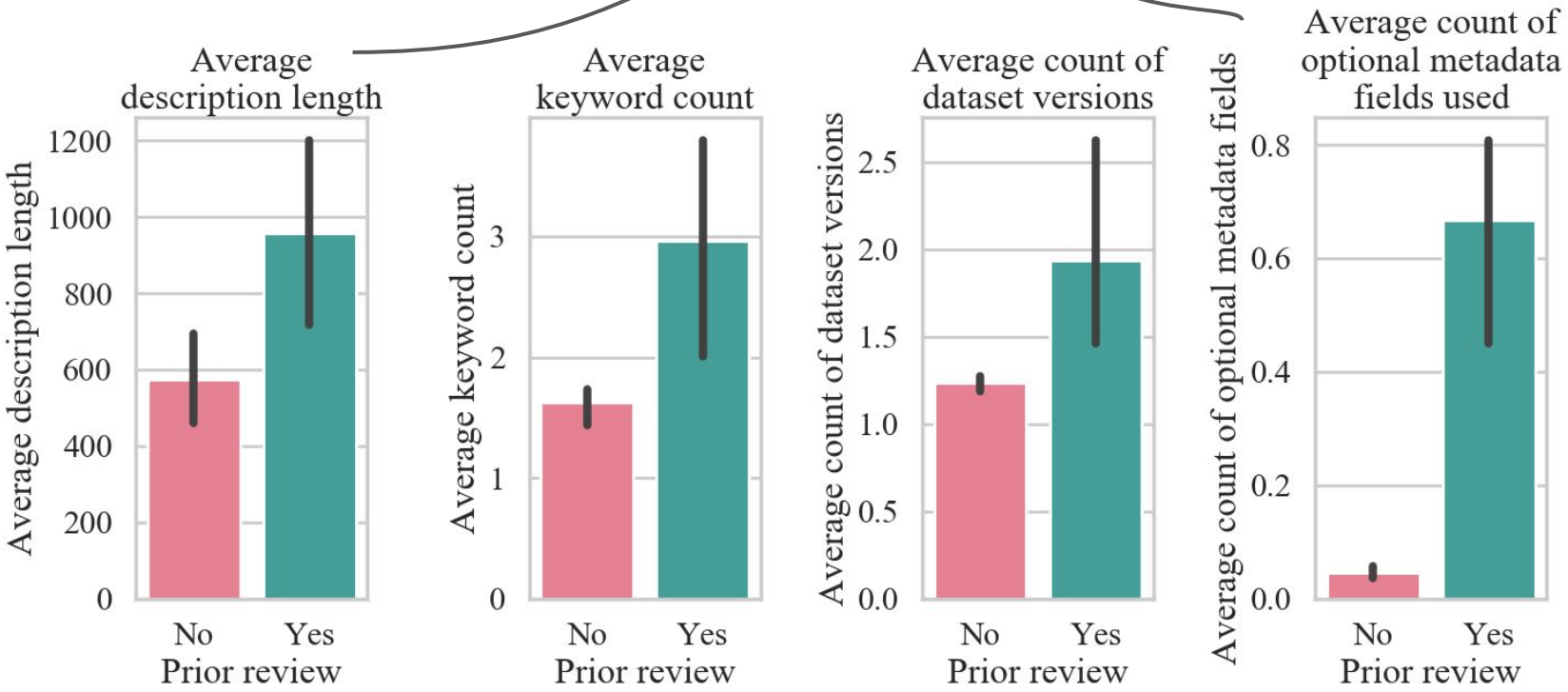
- Groups such as journals, laboratories, and project teams curate their data collections on Dataverse
- The Dataverse software supports review workflows that allow curators to ensure that deposited datasets meet group-defined expectations
- Data quality influencers (or researchers' perception of data quality):
  - reputation and reliability of a repository
  - use optional repository features
  - The more extensive the use of optional features, the more FAIR (Findable, Accessible, Interoperable, and Reusable) the dataset is likely to be

Dataset characteristics with and without prior review in Harvard Dataverse



Dataset characteristics with and without Harvard Dataverse

inputs to indexers and web crawlers used by search engines like Google's Dataset Search



Approach #1: Encourage Use of Curation Features

- We find that data depositors often do not adequately document their datasets. Prior review and mandatory fields can improve the quality of curation and the quality of deposited datasets.

Approach #1: Encourage Use of Curation Features

- We find that data depositors often do not adequately document their datasets. Prior review and mandatory fields can improve the quality of curation and the quality of deposited datasets.
- Finally, we find that articles linked to published data often include contextual information that metadata cannot sufficiently capture.
    - Academic literature remains the primary avenue through which researchers find and evaluate secondary data.

Approach #1: Encourage Use of Curation Features

- We find that data depositors often do not adequately document their datasets. Prior review and mandatory fields can improve the quality of curation and the quality of deposited datasets.
- Finally, we find that articles linked to published data often include contextual information that metadata cannot sufficiently capture.
  - Academic literature remains the primary avenue through which researchers find and evaluate secondary data.
  - Repositories can encourage bi-directional linking between publications and datasets to facilitate direct access between them. Therefore, citing datasets across the scholarly record makes them both more **findable** and **better documented**.

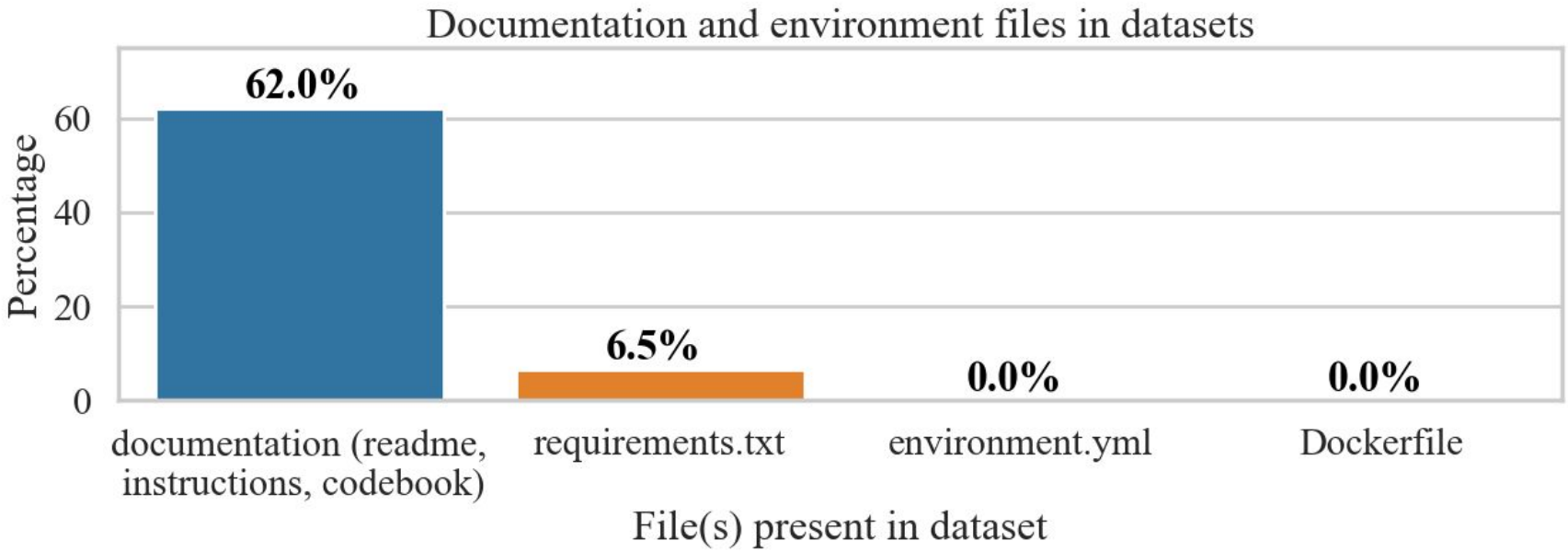
Approach #2: Ensure Code Completeness

- Research code as a common element in many datasets
- Challenges of sharing code
  - Research code dependent on software, OS, hardware
  - Computing methods are often not sufficiently documented

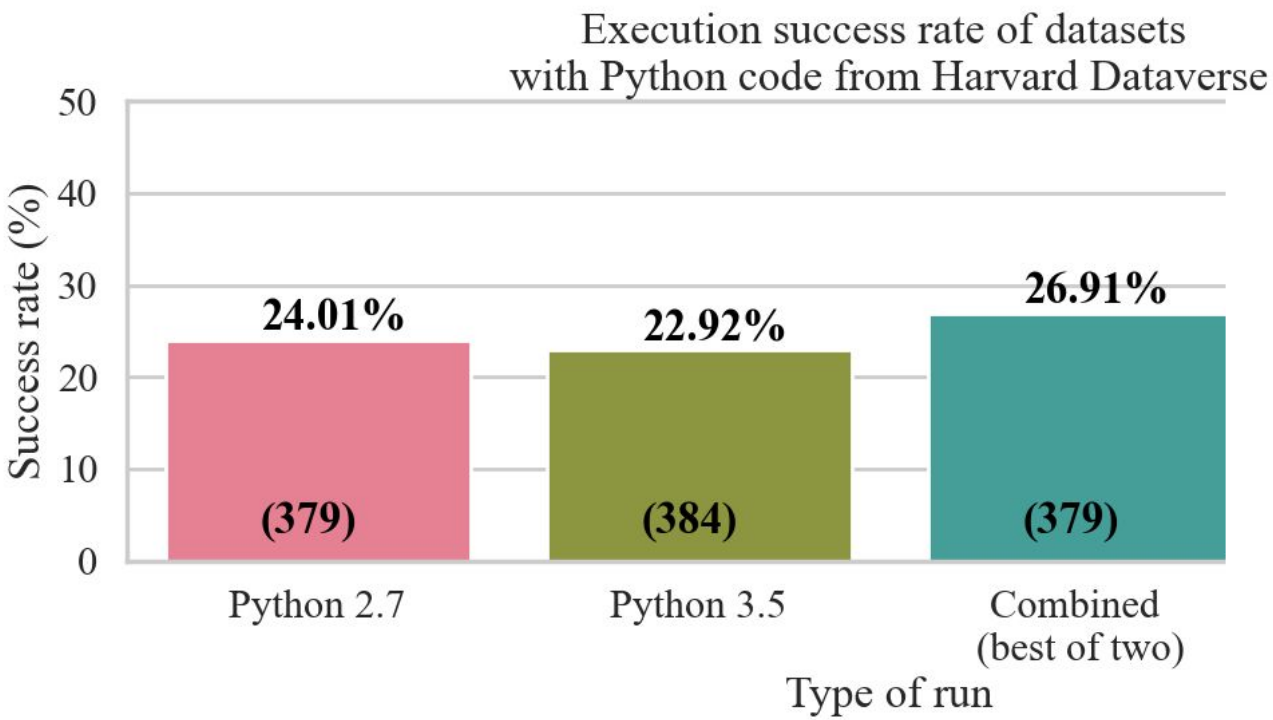
Approach #2: Ensure Code Completeness

- Research code as a common element in many datasets
- Challenges of sharing code
  - Research code dependent on software, OS, hardware
  - Computing methods are often not sufficiently documented
- To illustrate this challenge, we conducted a study where we retrieved 92 publicly available replication datasets that contain Python code
  - We examined the datasets, looking for files such as requirements.txt, which are common conventions for documenting needed code dependencies
  - We automatically (naively) re-executed Python files with Python 2.7 and Python 3.5 with a time limit of 10 minutes per file

Do datasets contain documentation and a record of code dependencies?

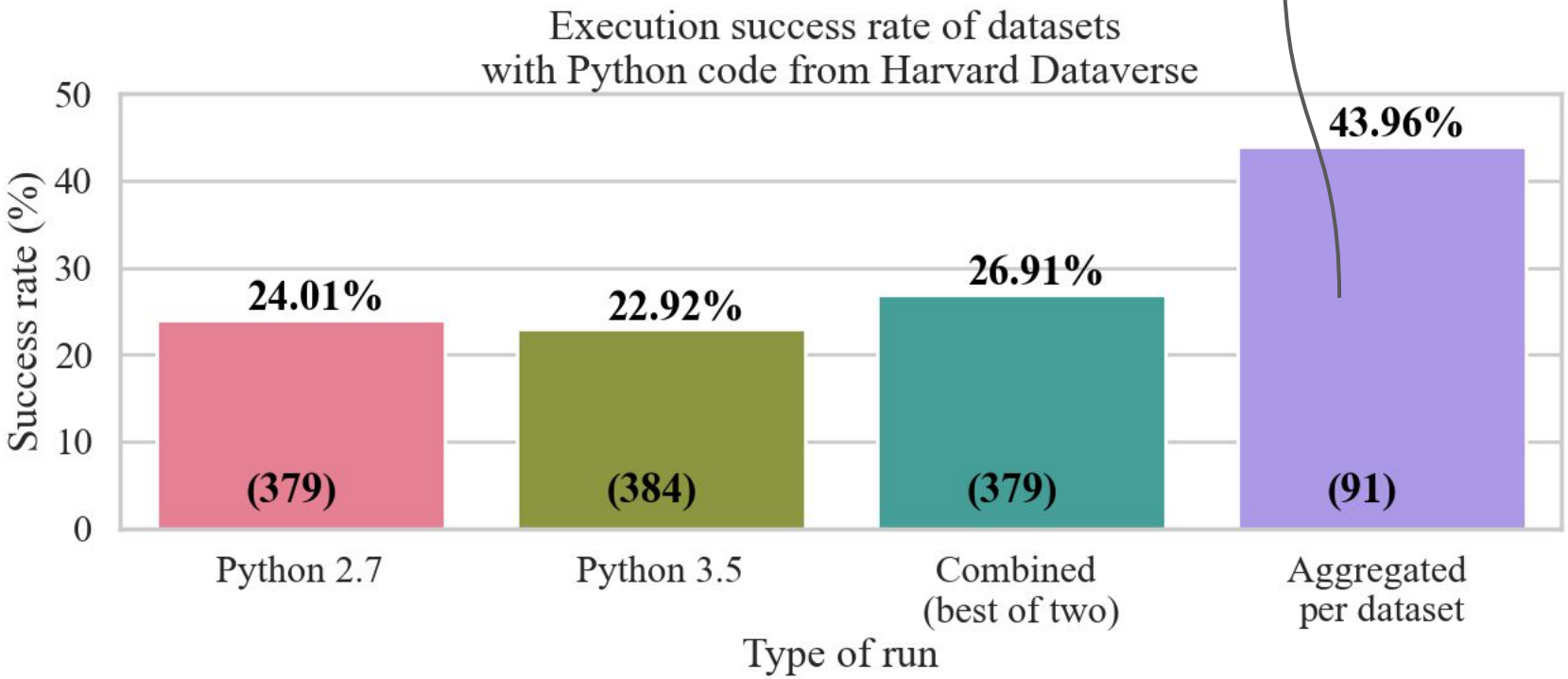


Are Python files re-executable?



Are Python files re-executable?

Success if one Python file from a dataset executes with success



Approach #2: Ensure Code Completeness

- The results show that further support is needed to document research code in data repositories
- Encourage storing environment files in datasets:
  - These are for example requirement.txt for Python and install.R for R
  - Pop-up windows or user instructions

Approach #2: Ensure Code Completeness

- The results show that further support is needed to document research code in data repositories
- Encourage storing environment files in datasets:
  - These are for example requirement.txt for Python and install.R for R
  - Pop-up windows or user instructions
- Integration with reproducibility platforms that use virtual containers and encapsulation
  - Whole Tale, Code Ocean, Binder and Renku
  - Considered within Dataverse open-source community

Approaches in practice: a Dataverse use-case

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Ana Trisovic

Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1

Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", https://doi.org/10.7910/DVN/EA3LC5, Harvard Dataverse, V4

Cite Dataset | Learn about Data Citation Standards.

Access Dataset
Edit Dataset
Link Dataset
Contact Owner | Share

Dataset Metrics
22 Downloads

Description

This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on completeness of research code.
Run this code on Jupyter Binder here: launch binder (2020-09-27)

Subject

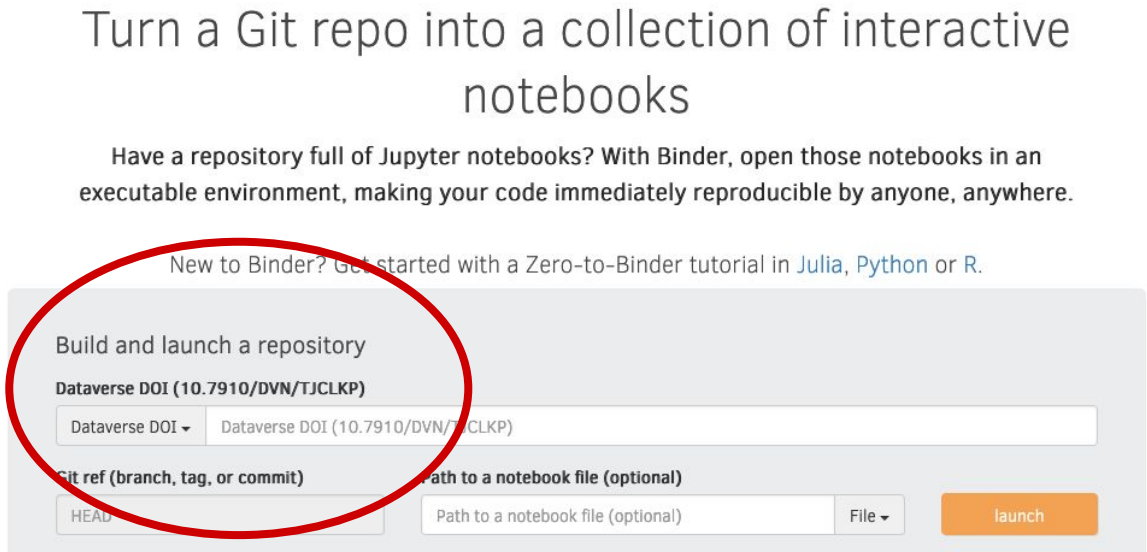
Computer and Information Science

Files | Metadata | Terms | Versions

Search this dataset... | Find

+ Upload Files

Approaches in practice: a Dataverse use-case



Approach #3: Incorporate Gamified Design Elements

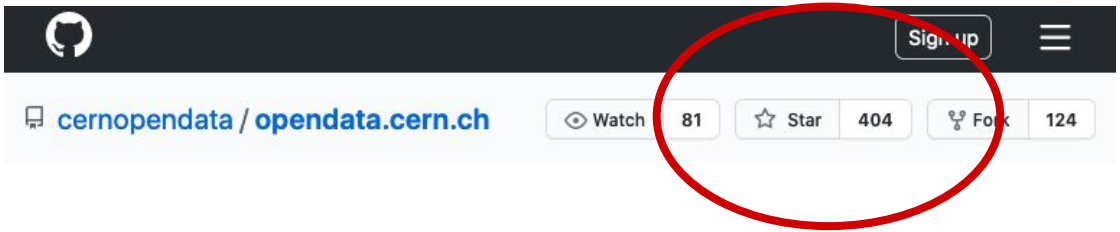
- Gamification is a use of game design elements in non-game context
- Badges, points and leaderboards are some of the most common game design elements
- Used in teaching, citizen science, health applications etc.

Approach #3: Incorporate Gamified Design Elements

- Gamification is a use of game design elements in non-game context
- Badges, points and leaderboards are some of the most common game design elements
- Used in teaching, citizen science, health applications etc.
- Gamified badges are identified as the most suitable element to incentive dataset sharing. They are seen as achievable goal while also improving visibility.
    - CERN study
    - Open Science Badges (OSB) incentive data sharing for medical journals

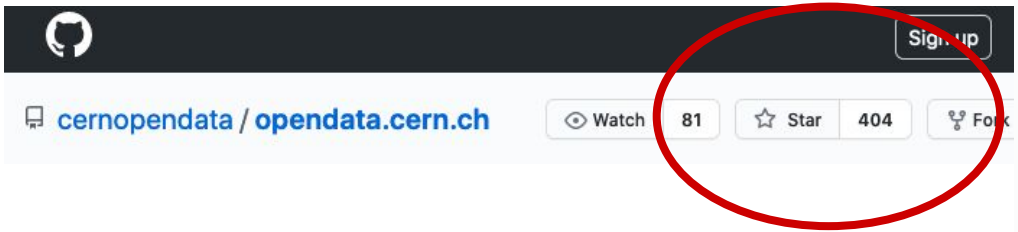
Approach #3: Incorporate Gamified Design Elements

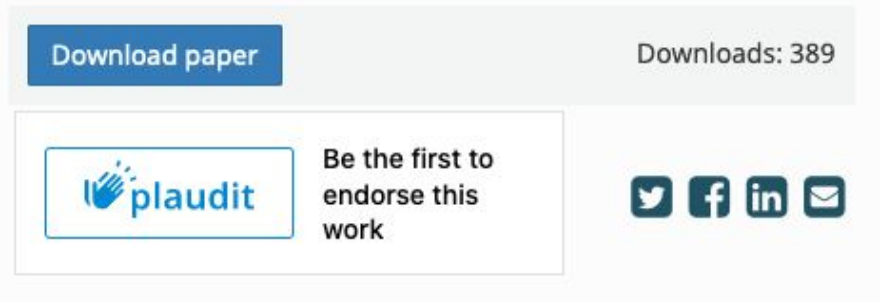
- In addition to motivating researchers, gamified elements can be used to identify resources of high quality within an available resource pool
- Examples:
  - Github,

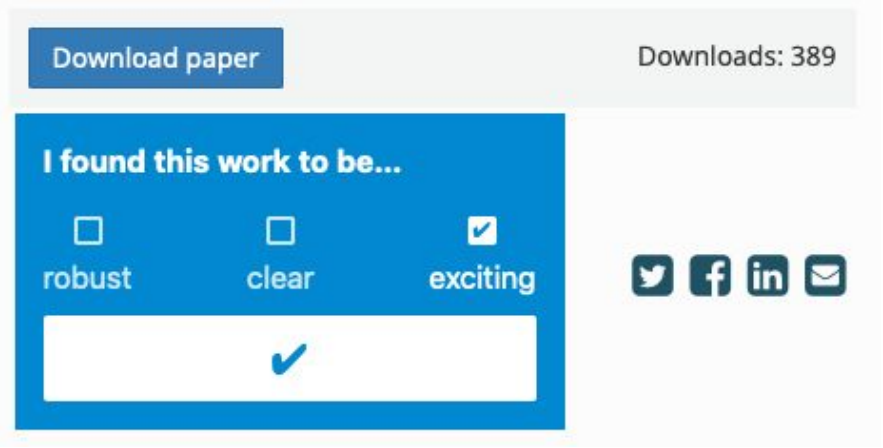


Approach #3: Incorporate Gamified Design Elements

- In addition to motivating researchers, gamified elements can be used to identify resources of high quality within an available resource pool
- Examples:
  - Github,
  - SocArXiv







Approaches in practice: a Dataverse use-case

HARVARD Dataverse

Add Data | Search | About | User Guide | Support | Ana Trisovic

Replication Data for: Repository approaches to improving quality of shared data and code

Version 4.1

Trisovic, Ana, 2020, "Replication Data for: Repository approaches to improving quality of shared data and code", https://doi.org/10.7910/DVN/EA3LC5, Harvard Dataverse, V4

Cite Dataset | Learn about Data Citation Standards.

Access Dataset
Edit Dataset
Link Dataset
Contact Owner | Share

Dataset Metrics
22 Downloads

Description This is supplementary data to the article "Repository approaches to improving quality of shared data and code," and in particular, its first section on completeness of research code.
Run this code on Jupyter Binder here: launch binder (2020-09-27)

Subject Computer and Information Science

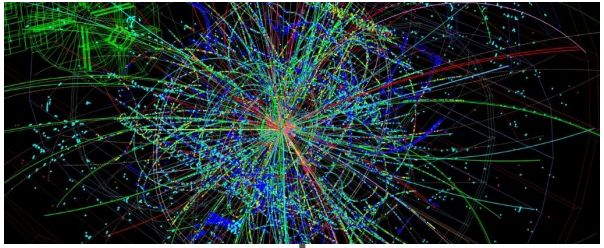
Files | Metadata | Terms | Versions

Search this dataset... | Find | Upload Files

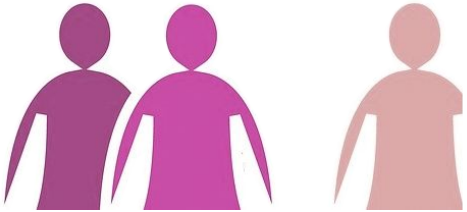
A high number of downloads signals the popularity of a dataset

Long-tail of science

Big Science data products



Small but much more common data



Approaches in practice: a Dataverse use-case

- Center for Open Science badges

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: Brokers, Social Networks, Re Clientelism

Version 1.0

Ravanilla, Nico, 2021, "Replication Data for: Brokers, Social Networks, Reciprocity, and Clientelism", https://doi.org/10.7910/DVN/UTRXT7, Harvard Dataverse, V1, UNF:6:ogxX+qke1Dd6xsG8ctEYcA== [fileUNF]

Cite Dataset Learn about Data Citation Standards.

Description

Although canonical models of clientelism argue that brokers use dense social networks to monitor and enforce vote buying, recent evidence suggests that brokers can instead target intrinsically reciprocal voters and reduce the need for active monitoring and enforcement. Combining a trove of survey data on brokers and voters in the Philippines with an experiment-based measure of reciprocity, and relying on local naming conventions to build social networks, we demonstrate that brokers employ both strategies conditional on the underlying social network structure. We show that brokers are chosen for their central position in networks and are knowledgeable about voters, including their reciprocity levels. We then show that, where village social networks are dense, brokers prefer to target voters that have many ties in the network because their votes are easiest to monitor. Where networks are sparse, brokers target intrinsically reciprocal voters whose behavior they need not monitor. (2020-07-14)

Subject Social Sciences

Keyword Brokers, Social networks, Reciprocity, Vote buying, Clientelism

Related Publication Ravanilla, Nico, Dotan Haim, and Allen Hicken. [date]. "Brokers, Social Networks, Reciprocity, and Clientelism." *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.

OPEN DATA PR OPEN MATERIALS PR

Conclusions

- Data repository features and services can contribute significantly to the quality and reusability of shared datasets
  - Runtime environment components for code can be encouraged
  - Repositories can support a deposit workflow with prior review, which often results in better-curated datasets
  - Including gamification elements promote data sharing by providing recognition for authors and useful metrics for data reusers

Conclusions

- Data repository features and services can contribute significantly to the quality and reusability of shared datasets
  - Runtime environment components for code can be encouraged
  - Repositories can support a deposit workflow with prior review, which often results in better-curated datasets
  - Including gamification elements promote data sharing by providing recognition for authors and useful metrics for data reusers

Open Access | Feature Paper | Article

Repository Approaches to Improving Quality of Shared Data and Code

by Ana Trisovic [1,*], Katherine Mika [1], Ceilyn Boyd [1], Sebastian Feger [2,3] and Mercè Crosas [1]

[1] Institute for Quantitative Social Science, Harvard University, 1737 Cambridge St, Cambridge, MA 02138, USA
[2] European Organization for Nuclear Research (CERN), 1, Esplanade des Particules, CH-1217 Meyrin, Switzerland
[3] LMU Munich, 1, Geschwister-Scholl-Platz, 80539 Munich, Germany
[*] Author to whom correspondence should be addressed.

Academic Editor: Maurizio Lenzerini

Data **2021**, *6*(2), 15; https://doi.org/10.3390/data6020015

Received: 22 December 2020 / Revised: 27 January 2021 / Accepted: 28 January 2021 / Published: 3 February 2021

(This article belongs to the Special Issue Data Quality and Data Access for Research)

View Full-Text | Download PDF | Browse Figures | Citation Export

Abstract

Sharing data and code for reuse have become increasingly important in scientific work over the past decade. However, in practice, shared data and code may be unusable, or published results obtained from them may be irreproducible. Data repository features and services contribute significantly to the quality, longevity, and reusability of datasets. This paper presents a combination of original and secondary data analysis studies focusing on computational reproducibility, data curation, and gamified design elements that can be employed to indicate and improve the quality of shared data and code. The findings of these studies are sorted into three approaches that can be valuable to data repositories, archives, and other research dissemination platforms. View Full-Text

Keywords: data quality; data repository; digital libraries; data curation; fair principles; open data; open code; gamification

▼ Show Figures

10.3390/data6020015

Thank you for your attention!