Data & Materials Sharing[1]

Why, when, where, & how

Rick Gilmore   Bruce Desmarais

Psychology/CSC

Political Science/SoDA

Preliminaries

Agenda

- Why share data & materials
- When to share...
- Where to share...
- How to share...

Why share data

Transparency

The Mind of a Con Man

Diederik Stapel, a Dutch social psychologist, perpetrated an audacious academic fraud by making up studies that told the world what it wanted to hear about human nature. Koos Breukel for The New York Times

Figure 1: (Bhattacharjee 2013)

Figure 2: (Marcus and Oransky, n.d.)

Accelerate discovery

Pfizer   Science   Products   Stories   Newsroom   About

Pfizer and BioNTech Announce Vaccine Candidate Against COVID-19 Achieved Success in First Interim Analysis from Phase 3 Study

Monday, November 09, 2020 - 06:45am

Figure 3: (Pfizer 2020)

Improve rigor

Figure 4: (Cho 2012)

Strengthen reproducibility

▶ What's your project's "bus number"?
▶ Could someone else on your team reproduce analyses done by X?
▶ *Methods* reproducibility (Goodman, Fanelli, and Ioannidis 2016)

Meet a requirement

- Funder
- Journal
- Institutional
- Professional organization

8.14 Sharing Research Data for Verification

(a) After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release. This does not preclude psychologists from requiring that such individuals or groups be responsible for costs associated with the provision of such information.

(b) Psychologists who request data from other psychologists to verify the substantive claims through reanalysis may use shared data only for the declared purpose. Requesting psychologists obtain prior written agreement for all other uses of the data.

Figure 5: (American Psychological Association, n.d.)

Ethical duty

Whereas regulators of human subjects research often view data sharing solely in terms of potential risks to subjects, we argue that the principles of human subject research require an analysis of both risks and benefits, and that such an analysis suggests that **researchers may have a positive duty to share data** *in order to maximize the contribution that individual participants have made. (Brakewood and Poldrack 2013)*

When to share

When NOT to share...
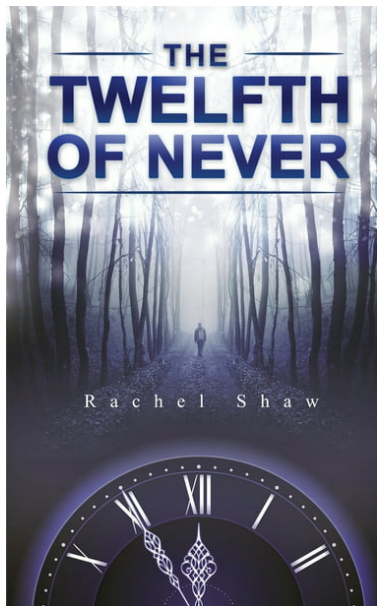
- When you've milked every last ounce of information out of a dataset, and all that remains is a lifeless, shriveled husk

On submission

▶ Some journals require that authors submit data for review. Ex, *Nature Scientific Data*

2 Select a repository for your data ⬆

When submitting a **Data Descriptor**, authors must deposit all relevant datasets in an appropriate public repository prior submission, and the completeness of these datasets will be considered during editorial evaluation and peer-review. Datasets must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (except reasonable controls related to human privacy issues or public safety).

▶ Reviewers have the option to review the data along with the manuscript.

▶ Bruce had experience of reviewer at NSD pointing out that numbers from the manuscript did not match data repository.

After acceptance

▶ Pre-publication verification (separate from review) increasingly common

GUIDELINES FOR DATA REPLICATION

The Journal of Politics operates a strict data replication policy effective January 1, 2021. It is the policy of the *JOP* to publish papers only if the data used in the analysis are clearly documented, available, and the empirical findings reproducible. Authors of accepted papers that contain empirical as well as simulation-based analyses are required to provide datasets, codes, and other relevant information necessary to facilitate replication. All manuscripts are accepted contingent on their replicability, which will initially be assessed by the *JOP* replication analyst(s) who will be assigned to your manuscript at the conditional acceptance stage. Manuscripts that are not replicable will be rejected for publication. This document provides guidelines on how to prepare data replication materials.

▶ Can take up to six months of back-and-forth with journal assistant.

▶ Leads to more careful organization pre-submission.

On publication

▶ Post and disseminate data and code with paper.
▶ Creates independent way to get cites.

SNAP judgments into the digital age: Reporting on food stamps varies significantly with time, publication type, a…
Benjamin W. Chrisinger, Eliza W. Kinsey, Ellie Pavlick, Chris Callison-Burch

Abstract

Introduction

Materials and methods

Results

Discussion

20. Bakshy E, Messing S, Adamic L. Replication Data for: Exposure to Ideologically Diverse News and Opinion on Facebook. Harvard Dataverse, V2; 2015.
View Article • Google Scholar

21. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science. 2014;58: 1064–1082.
View Article • Google Scholar

End of grant period

▶ Major funders require data/material/publication sharing

NSF's data sharing policy

NSF-funded investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF awards.

Where to share data

Data repository

- Dataverse (linked to journal article)
- Databrary.org (esp. video/audio, identifiable data)
- Inter-University Consortium for Political and Social Research (ICPSR)
- Open Science Framework (OSF)

- PSU ScholarSphere
- PSU Data Commons
- NIH Data Archive
- Zenodo.org

Why repositories?

- Long-term storage (preservation, in FAIR forms)
- Data "communities"
- Access
- Specialized expertise (e.g., curation)
- Persistent identifiers (e.g., DOIs) enable citation & tracking

Alternatives to repositories

▶ Supplemental materials
▶ Lab/project website

Data journal

List of data journals

Kindling, Maxi; Strecker, Dorothea

This document describes a dataset that aggregates information about 135 data journals.
Data journals focus on the publication of data papers -- a specialized publication type describing datasets, their collection and reuse potential that is peer-reviewed, citable and indexed.
This dataset includes a comprehensive list of data journals that was compiled by aggregating existing sources, as well as an overview of these sources.

The list is continually updated on GitHub, where additional information on data journals (URLs of data journal homepages) is provided: https://github.com/MaxiKi/data-journals

Preview

issn	journal_title	publisher	data_journal_type
2363-4952	RIDE - A review journal for digital editions and resources	Institut für Dokumentologie und Editorik	pure
2059-481X	Journal of Open Humanities Data	Ubiquity Press	pure
2296-7745	Frontiers in Marine Science	Frontiers	mixed
2054-345X	Human Genome Variation	Springer Nature	mixed
2032-6378	Journal of Astronomical Data	Vrije Universiteit Brussel	pure
2603-	Viticulture Data Journal	Pensoft Publishers	pure

scientific **data**

Two decades of fumigation data from the Soybean Free Air Concentration Enrichment facility

Elise Kole Aspray, Timothy A. Mies ... Elizabeth A. Ainsworth

Data Descriptor | 20 April 2023
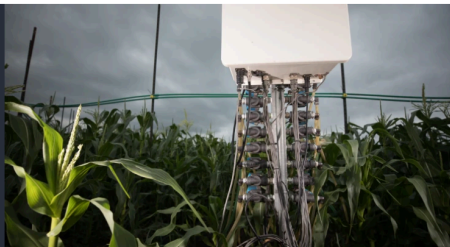
Announcements

Collection open for submissions

Scientific Data is open to submissions for this special collection: Meteorology and hydroclimate observations and models

Open for submissions

www.nature.com/sdata/

n.b.– Gilmore is on editorial board.

- Could be extra pub
- Especially good for early-career researchers

How to share data

SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

FAIR principles

F indable   A ccessible   I nteroperable   R eusable
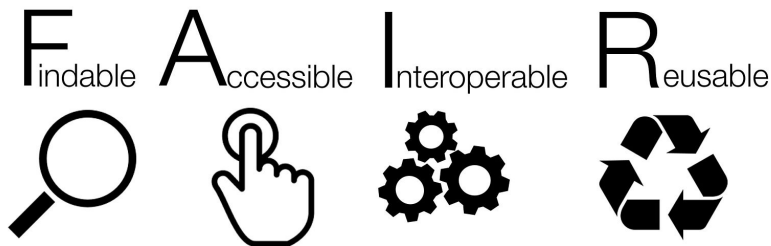
(Wilkinson et al. 2016)

FAIR in-practice

- **F**indable
    - Repository or journal with persistent identifier
- **Accessible**
    - Not behind paywall
- **I**nteroperable
    - {.txt, .csv} vs. {.docx, .xlsx}
- **R**eusable
    - Data definitions/dictionary
    - Code to visualize/analyze

CARE principles

Figure 7: (Carroll et al. 2021)

Seek permission to share

▶ Sensitive & identifiable data can (often) be shared with *restricted audiences* (i.e., researchers)
  ▶ e.g., Databrary (databrary.org)
▶ Qualitative Data Repository (QDR; qdr.syracuse.edu)
▶ Sharing $\neq$ public access

Final thoughts

- Plan for sharing as early as possible
- Data + code $\rightarrow$ reproducible analyses
- Internally reproducible $\rightarrow$ easier to share

Your thoughts?

Resources

About

This talk was produced using Quarto.

The source files are in R and R Markdown, then rendered to PDF slides in the beamer format. The PDF slides are hosted in a GitHub repo:

https://github.com/penn-state-open-science/bootcamp-2023-data-sharing

References

American Psychological Association. n.d. "Ethical Principles of Psychologists and Code of Conduct."
https://www.apa.org/ethics/code.
https://www.apa.org/ethics/code.

Bhattacharjee, Yudhijit. 2013. "The Mind of a Con Man." *The New York Times*, April.
https://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html.

Brakewood, Beth, and Russell A Poldrack. 2013. "The Ethics of Secondary Data Analysis: Considering the Application of Belmont Principles to the Sharing of Neuroimaging Data." *NeuroImage* 82 (November): 671–76.
https://doi.org/10.1016/j.neuroimage.2013.02.040.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, and Shelley Stall. 2021. "Operationalizing the CARE and FAIR Principles for Indigenous Data Futures." *Scientific Data* 8 (1): 108.
https://doi.org/10.1038/s41597-021-00892-0.

Cho, Adrian. 2012. "Once Again, Physicists Debunk