

Reproducibility, Transparency, and Data Ethics

How and when to share data, and when not to

Lars Vilhuber

Invalid Date



Follow along



labordynamicsinstitute.github.io/reproducibility-ethics/presentation/ (PDF)

Reproducibility, Transparency, and Data Ethics

How and when to share data, and when not to



Overview

The question

How do you create **transparency** and **credibility** when you cannot share the data you use?

- What kind of data **can** you share?
- Strategies for data **collection**, data **use**, and data **sharing**
- Relevant if you **collect your own data**, or if you use **confidential data**

Roadmap

1. When can you share data – and how do you know you cannot?
2. How do you know you **have** data you cannot share?
3. Methods to make non-shareable data shareable
4. Methods to publish data securely, with controls

When can you share data?

When can you share data?

- How do you know you **can** share data?

How do you know you cannot share data?

- **Licenses and data use agreements**
 - Example 1: PSID
 - Example 2: Demographic and Health Surveys (DHS)
- **Institutional review board** (IRB) restrictions
- **Confidentiality agreements**
- **Legal restrictions** ( GDPR,  LGPD)

Data ethics

- Discussion on data ethics (video)
- Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, 1(1), 131-144.
<https://doi.org/10.1177/2515245917747656>

**Do you have data
you cannot share?**

Separate from Day 1

- Best practices for separating shareable and non-shareable data start on **Day 1**
- J-PAL guide for de-identification of data

Verify on Day T-1

- Did you properly separate shareable and non-shareable data?
- PII-checker
 - see Brodeur, Abel, and David Valenta. 2026. *On the Prevalence of Personally Identifiable Information (PII) in the Social Sciences.* I4R Discussion Paper No. 320.
- How can you decide what data to publish *without publishing the data*? Private LLMs...

A recent message from an author

After having been provided with the LLM-produced PII check:

We have removed the PII and resubmitted the deposit for your review. Here's a brief overview of the changes we made for reference:

- Removed PII: We removed the “name” field, subject IDs, and all location data (latitude/longitude).
- Deleted unused files: We removed files that were not used in the analysis and would have required PII to merge with the rest of the data.

Discussion

What could they have done earlier?

**Making non-
shareable data
shareable**

Methods

- Data **de-identification**
- True data **anonymization**
 - k-anonymity
 - differential privacy
- **Aggregation** and summarization
- **Synthetic** data generation

k-anonymity

Try it yourself:

 Open in Colab

Open In Colab

Differential privacy

-  Video
- Garfinkel, Simson, J. M. Abowd, and Christian Martindale. 2018. "Understanding Database Reconstruction Attacks on Public Data." *ACM Queue* 16 (5).
- Webb, Kaitlyn R., Soumya Mukherjee, Aratrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. 2026. *Assessing Utility of Differential Privacy for RCTs*. arXiv:2309.14581v2.

**Publishing data
securely, with
controls**

Methods to publish data with controls

- Data repositories have mechanisms
 - e.g., ICPSR, Dataverse
- Journal policies require “5 years retention” – how do you ensure that?

Flip it

You are the licensor!

See also

When data are not shareable

- Preparing code when data are not shareable:
Reproducibility when data are confidential

The end!

Source

- This document's source:
<https://github.com/labordynamicsinstitute/reproducibility-ethics>
- Licensed under 