gesis

Leibniz Institute
for the Social Sciences

Tools and Workflows for Reproducible Research in the Quantitative Social Sciences

Sharing, collaborating, publishing

Johannes Breuer, Bernd Weiss, & Arnim Bleier

2022-11-18

Leibniz Association

Sharing is caring

One prerequisite for research being reproducible (by others) is sharing research materials. There are many parts of their work that researchers can share to increase the reproducibility as well as the (potential) replicability of their work (see Klein et al. , 2018). Four main types of output are:

1. Data
2. Code & scripts (for data collection, processing, and analysis)
3. Other study materials (e.g., questionnaires or stimulus materials)
4. (Detailed) Information about the study procedure

Notably, all of these outputs should be well-documented (e.g., via README files, metadata or comments in code).

Sharing for reproducibility

While sharing study materials and information about the procedure are important for replicability, for reproducibility, the most important things to share are the data and code.

Sharing options

In the previous sessions, we have seen how we can use *Jupyter Notebooks* and *Binder* to share our code and results in a way that also enables interactivity. The underlying files were hosted on *GitHub*. However, while *GitHub* is very useful for developing and sharing code, it is not an ideal platform for sharing data. There are different reasons for that:

- data files are not necessarily plain-text formats
- privacy/data protection issues with public repositories
- questions regarding how to document materials

Sharing options

There are many different ways in which researchers can share their data and code/scripts (see Klein et al., 2018). Keeping only local copies of things and sharing upon personal request is not a very sustainable or scalable solution. A better option is sharing via institutional or public archives and repositories.

Fantastic repositories and where to find them 

The paper by Klein et al. (2018) provides an overview of public repositories that hold psychological data.[1] A good tool for finding suitable repositories is the *Registry of Research Data Repositories*.

[1] However, parts of this overview have inevitably become somewhat outdated since the paper was published.

How to choose a repository

In general, research data (and code) that are publicly archived should follow the so-called FAIR principles (Wilkinson et al., 2016), meaning that the shared materials should be...

- **F**indable: Persistent identifiers; metadata; indexed
- **A**ccessible: Retrievable by identifier; controlled access where necessary
- **I**nteroperable: Standardized metadata; open, lightweight, and interoperable file formats (e.g., CSV, TSV, JSON, ODS)
- **R**eusable: Documented; clear usage license

How to choose a repository

Some more specific key criteria for choosing a repository are that it should...

- use persistent and unique identifiers (such as DOIs)
- accommodate licensing
- feature access controls (e.g., allowing the restriction of access to a particular set of users)
- have persistence guarantees for long-term access
- store data in accordance with local legislation (e.g. the GDPR in Europe)

See Klein et al. (2018) for further details

General public archiving options

Two archiving options that are quite popular among researchers are the *Open Science Framework* (OSF) and *Zenodo*.

While these two archives are not curated, which somewhat reduces findability, they are quite flexible and easy to use. They can be used to share different types of content, including data and code, and also offer some degree of access control. A nice feature of the *OSF* and *Zenodo*, especially for sharing code, is that they offer integration with *GitHub*.

Curated disciplinary archives

A good way to maximize findability and reusability is to make use of curated disciplinary archives. Two of their key advantages are that they a) offer support for researchers who want to archive and share their data, and b) employ structured metadata.

Curated disciplinary archives

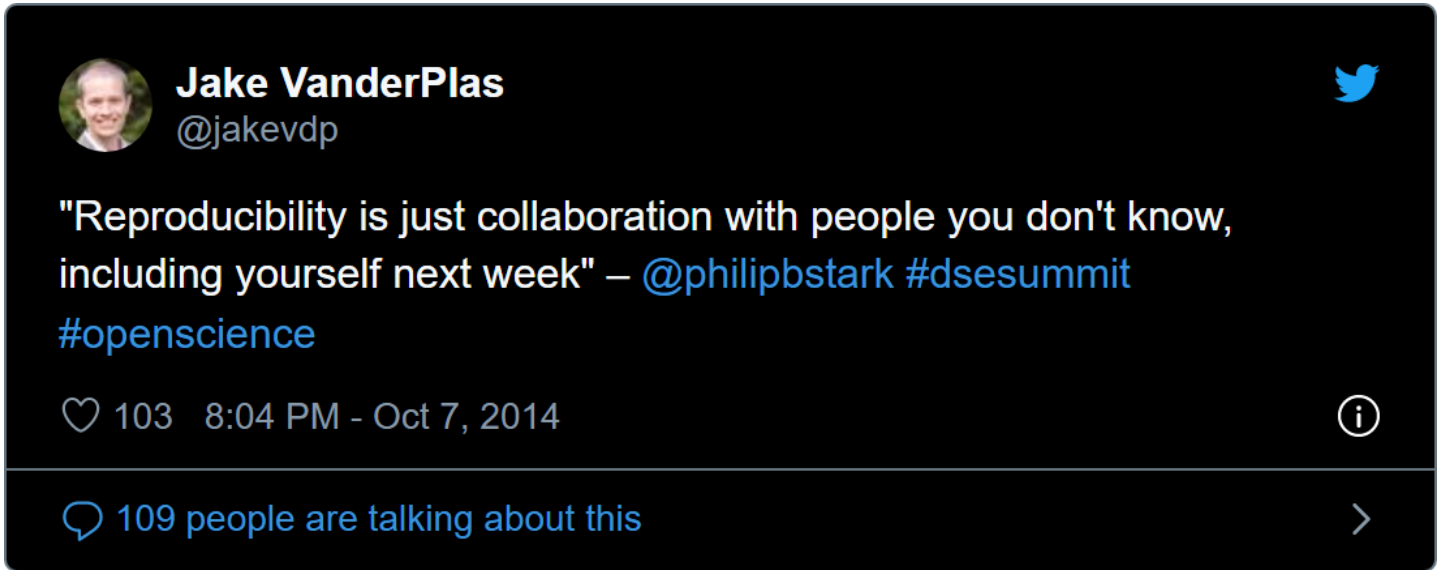
GESIS offers a data archive for quantitative social science data with a focus on survey data (and now also digital trace data).

There also are data archives with a social science focus in other countries. Many of the European ones are members of *CESSDA*, and in the U.S. there is the *ICPSR*.

Curated disciplinary archives

There are, of course, other curated repositories with a focus on specific disciplines and/or data types, such as *PsychArchives* by *ZPID* for psychology, or *Qualiservice* in Germany and the *The Qualitative Data Repository* in the U.S. for qualitative data.

It's all about collaboration 🤝



https://twitter.com/jakevdp/status/519563939177197571

Teaming up for reproducible research

Up until now, the scenario in this workshop was that we work alone on our code and data (as well as the reports we produced based on those). However, we typically have collaborators/coauthors with whom we do this. In such cases, we need to have solutions in place for collaboratively editing the materials and resulting output.

Lucky for us, *Git* & *GitHub* have options for collaborating with others (which we will discuss in the next session).