PTOS-open-data

Jürgen Schneider

Invalid Date

Table of contents

Welcome

This is a workshop on open data.

PUBLICATIONS AND DATA

If you like the workshop...

0.0.0.1 and want to keep it forever, make it yours

For that...

1. Fork the github repo this Quarto book is based on
2. Go to settings of your new repo and go to the "pages" section. Then set the "Branch" option to `gh-pages` (leave the dropdown to the right of this at `/root`)
3. Wait a minute to let the website get deployed. You can check on the status in the "Actions" tab of your repo.
4. Back on the main repo site, click on "About" (top right). In the URL of the website, change "j-5chneider" to your username "[your github username].github.io/PTOS-open-data/" (you might need to activate GitHub Pages for that, by creating a GitHub Pages repo)
5. open your new webpage by clicking on that link in the "About" section

0.0.0.2 give it a star in GitHub

So you get noticed if I update something on the github repo.

And I get that sweet sweet dopamine. Hmm dopamine.

Part I

Open and FAIR Data

1 Openness

- **anyone**
- can **readily access** the data
- at no more than a **reasonable reproduction cost** (i.e., internet connection)

(Open Knowledge Foundation, 2023)

> Openness is not a dichotomy
>
> "As open as possible as closed as necessary"
> (European Commission, 2023, p. 36)

2 FAIRness

https://www.go-fair.org/fair-principles/

(Wilkinson et al., 2016)

> **i** FAIRness vs. openness
>
> "does not necessarily mean that data has to be "open" [...] even highly protected data can be FAIR data"
> (Kraft, 2023)

(Kraft, 2023) (FAIR principles and the role of scientists)

2.0.1 Findable

The problem:
Just because we provide data online, doesn't mean that others will find it.

We could have the greatest data set to answer further research questions - if our colleagues don't know it exists or can't locate the data, openness will be of little value.

The solutions:

- Get a persistent identifier (e.g., DOI), where you provided your data
  - research data centers: Verbund FDB, RDC at ZPID, ...
  - repositories: Zenodo, psycharchives.org, osf.io, ...
- Mention DOI in publication that builds on this data (e.g., in the "data accessibility statement")
- Describe your data as richly as possible (metadata). *Research data centers* offer form fields tailored to the discipline or data type. With *repositories* use alternative possibilities, such as keyword fields.
  - e.g., which variables does the quantitative data set contain?
  - e.g., which topics does your data cover?
  - e.g., which population did you draw your sample from?

2.0.2 Accessible

The problem:
Just because others find our data doesn't mean the *access barriers* are as low as possible and doesn't mean they know *in which way* they are allowed to access it. Examples:

- Providing a link to the data in the text of a paywalled journal article
- Unclear licensing / use conditions when providing data (e.g., are non-researchers allowed to access the data or is it only open for qualified researchers?)

The solutions:

- Make sure access is free of charge (or as cheap as possible)
  - e.g., by providing link to data in publicly accessible sections of journal articles that are not open access
  - e.g., by using repositories or research data centers that allow access free of charge
- Make sure users know if they can access and under which conditions
  - e.g., *research data centers* ensure that terms of use are clear (who may access under what conditions) and offer different levels of access restriction
  - e.g., on *repositories* provide a readme-file and an open license (e.g., CC0, CC-BY, CC-BY-SA) with data sets for access cases

2.0.3 Interoperable

The problem:
Just because others downloaded our data doesn't mean they can open and manipulate it.

The solutions:

- Use file formats with open licenses
  - e.g., tabular data: CSV (with additional labelling script), RData
  - e.g., text data: PDF, HTML, ODT, RTF
- Make sure users know how different files are related to one another
  - e.g., define which file contains student data and which teacher data
  - e.g., define which file contains data from cohort 1 and which cohort 2, …

2.0.4 Reusable

The problem:
Just because others opened our data doesn't mean they understand the data and its use-conditions. Examples:

- Others can't understand what the column names of the tabular data set mean: Which columns in the data set relate to which variables in the journal article?
- Can someone from sociology use the data set from psychology they found on osf.io?
- Does someone reusing a data set have to cite the authors?

The solutions:

- Rich description/explanation of what user will find *in* the data set (meta descriptions about the data set as a whole, as for accessibility)
  - e.g., provide a codebook. How to semi-automatically create a codebook, see the R package `codebook`
- Provide a license for the use-cases
  - again, *research data centers* ensure that terms of use are clear (who may use under what conditions)
  - again, on *repositories* provide a readme-file and an open license (e.g., CC0, CC-BY, CC-BY-SA) with data sets for the use-cases

3 A good definition

Includes aspects of fairness. See

https://opendatahandbook.org/guide/en/what-is-open-data/

References

European Commission. (2023). *Horizon Europe (HORIZON). HE Programme Guide. Version 4.0*. Publications Office.

Kraft, A. (2023). *The FAIR Data Principles*. https://doi.org/10.23668/PSYCHARCHIVES.13577

Open Knowledge Foundation. (2023). What is Open Data? In *Open Data Handbook*. https://opendatahandbook.org/guide/en/what-is-open-data/.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18