Open Data - Utrecht

Jürgen Schneider

24 September 2024

Table of contents

Welcome

This is a workshop on open and FAIR 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.

Transparency

Affiliation

DIPF | Leibniz Institute for Research and Information in Education
Project in the *Cooperation Center ShaReD - Sharing and Reusing Data*

See personal webpage

Cooperation

Working closely together, but not at the research data center at DIPF

Normalizing disability I guess

It's not a brain-freeze, I am not confused, I stutter ;)

Open Science

a very quick introduction

View slides in browser: click here

Part I

Reasons

1 Exercise: Your reasons

What are your reasons?

Now that you are here, there seem to be some drivers for you to share data. What are these?

1. Go to this whiteboard
2. Take one or more of the blue rectangles from the right
3. Write your reason in the rectangle. One reason per shape.

Questions to be answered at the end?
Please put them here!

2 External Incentives I

2.1 Research funders

DFG

DFG Guidelines on the Handling of Research Data

> "Assuming that the publication of research data [...] does not conflict with the rights of third parties (in particular data protection or copyright), research data should be *made available* as soon as possible [...] that allows it to be usefully *reused* by third parties" (DFG, 2015, p. 1)

Guidelines for Safeguarding Good Research Practice. Code of Conduct

> "Where possible and reasonable, this includes making the research data [...] available" (DFG, 2019, p. 19)

ESRC

ESRC research data policy

> "Publicly-funded research data are a public good, produced in the public interest, which shall be made openly available and accessible with as few restrictions as possible" (ESRC, 2018, p. 2)

NSF

Proposal & Award Policies & Procedures Guide
You will need to hand in…

> "plans for archiving data, samples, and other research products, and for preservation of access to them" (NSF, 2023, pp. II–32)

ERC

Open Research Data and Data Management Plans Information for ERC grantees

> "Grantees are required to deposit their research data in a repository and provide open access at least to those data" (ERC, 2022, p. 4)

Questions to be answered at the end?
Please put them here!

3 External Incentives II

3.1 Policies scientific societies

Examples of societies that have established a policy on open and FAIR data.

(APA, 2017; DGfE et al., 2020; Gollwitzer et al., 2021)

"Homework": Look up the scientific society most relevant to you and check if they have a *policy/recommendation/guideline* on open science or sharing data. What does it say on openness, FAIRness and limitations?

3.2 Journal policies

Some journals encourage, some journals mandate a **data availability statement** in the manuscript.

	Not Implemented	Level I	Level II	Level III
Data Transparency	Journal encourages data sharing, or says nothing.	Article states whether data are available, and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.

E.g.,

- Psychological Bulletin: Data transparency Level 2
- Meta Psychology: Data transparency Level 3
- Psychological Science: Data transparency Level 1

See the TOP Factor website to search for the data transparency rating of your favorite journal.

Evidence that **"Data available upon reasonable request"** often does not keep its promise. Only **6.8%** / **17%** of data sets were acutally provided in an investigation (Gabelica et al., 2022; Hussey, 2023)

Questions to be answered at the end?
Please put them here!

4 Research I

4.1 No data = not reproducible

Computational reproducibility :=

> "a second investigator (including the original researcher in the future) can recreate the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions" (Kitzes et al., 2018, p. xxii)

➡ *Same data + same analysis = same results*

Why reproducibility?
Allows *independent researchers* to assess the analytic choices, assumptions, and implementations that led to a set of *scientific claims*.
➡ Check for validity and generalizability (Clyburne-Sherin et al., 2019; Obels et al., 2020)

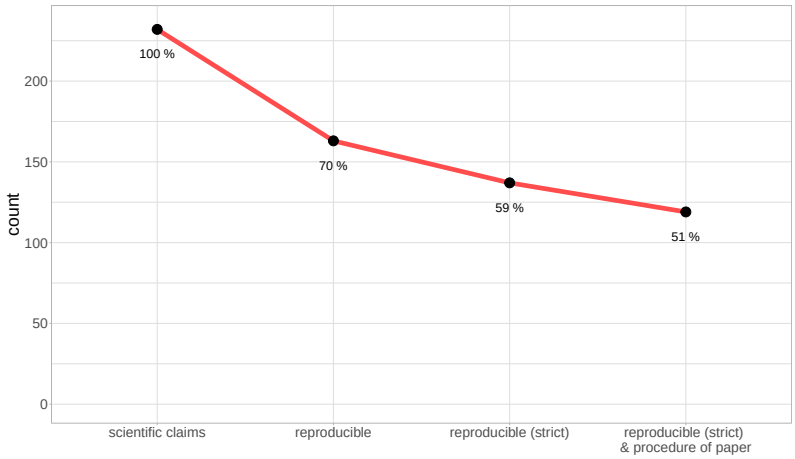
4.2 No FAIR data = reproducibility tedious

But computational reproducibility isn't as easy as it sounds (Artner et al., 2021)

4.2.1 Design

- checked 232 primary statistical claims
- from 3 journals
- after data was provided and accessible (33%, 25%, 26%)

4.2.2 Results



4.2.3 Conclusions

> **Vagueness Makes Assessing Reproducibility a Nightmare**
> most successful reproductions are predominantly the result of tedious and time-consuming work information about the provided raw data was often difficult to understand, and information about the relevant variables, data manipulations, and the used statistical model was often vague or inaccurate

(Artner et al., 2021, p. 12)

4.3 No data = barrier to replication

- Evidence e.g. from replication attempts in cancer biology (Errington et al., 2021)
- Due to various barriers, **50** of the **193** replication experiments could be **conducted** at all
- Missing data = major barrier to compute parameters to replicate

➜ data were open for **4** of **193** experiments

Questions to be answered at the end?
Please put them here!

5 Research II

5.1 Reuse

The reuse of research data can take many forms

Purpose	Advantage	Needs
Answer new research questions	Saves resources	analysis potential of data, good documentation (Logan et al., 2021; Steinhardt et al., 2021)
Teaching / student theses	Real-life-oriented education	good documentation
Meta-analyses	Easier estimation of parameters	Strictly reproducible code (e.g., Burgard et al., 2022)
Historical perspective	Data as historical artifacts	Potential of data varies

Questions to be answered at the end?
Please put them here!

6 Researcher

6.1 Get cited

- studies with available or link to data: **9%** or **25%** higher citation rates (Colavizza et al., 2020; Piwowar & Vision, 2013)
- But, **selection bias**: Willingness to share ↔ strength of evidence and quality of reporting (Wicherts et al., 2011)

- But, **higher trust** in authors openly sharing (Schneider et al., 2022)

6.2 Get funded

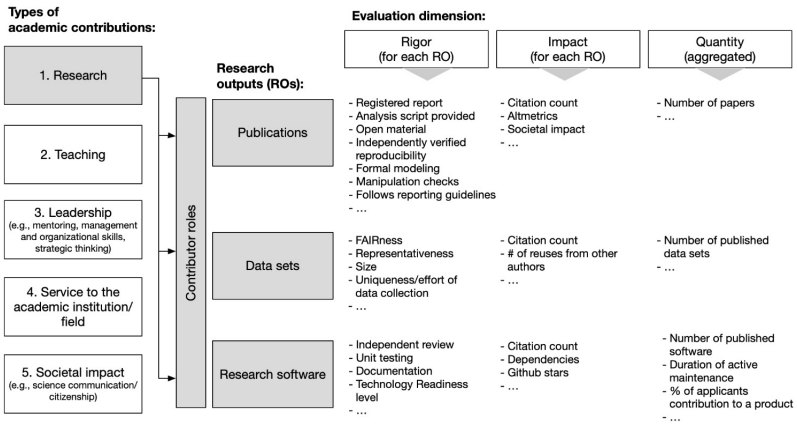
- Funders ERC (2022) & DFG (2015) require open data
- Federal agencies (IEA, 2022) and scientific societies (APA, 2017) endorse open sharing

6.3 Get published

- Some journals mandate *data availability statement*

6.4 Get hired

- new metrics for evaluation on the rise
- CoARA: "Value outputs associated with openness (FAIR data sets, […]" (CoARA, 2022, p. 21)→ *Signatories: ERC, League of European Research Universities, European University Association …*
- Example: DGPs recommendations on hiring and promotion (Gärtner et al., 2022; Schönbrodt et al., 2022)



(Schönbrodt et al., 2022, p. 4)

Questions to be answered at the end?
Please put them here!

Part II

Open and FAIR Data

7 Openness

Definition:

- **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 dogma and not a dichotomy
>
> "As open as possible as closed as necessary"
> (European Commission, 2023, p. 36)

Questions to be answered at the end?
Please put them here!

8 FAIRness

Purpose:

> "enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals" (Wilkinson et al., 2016, p. 1)

See also go-fair.org

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

8.0.1 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

  - search for a research data center that fits your needs: re3data.org
  - recommended repositories: Zenodo, 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?

8.0.2 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

8.0.3 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, …

8.0.4 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:

- Adhere to standards in folder organization
  - e.g., PSYCH-DS (see technical specification draft)
- 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

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

Questions to be answered at the end?
Please put them here!

9 Exercise

1. Go to this repository
2. Discuss for **which purposes** you consider this type of sharing to be **suitable** / **less suitable**
3. Discuss what you think makes this type of data sharing **FAIR** and what could be **improved**

Questions to be answered at the end?
Please put them here!

Part III

Limits

10 Limits

10.1 Not being open may be important

When does "as closed as necessary" apply?

The protection of individuals comes first and is more important than the potential reuse of data

10.1.1 Marginalized/vulnerable groups

- e.g., individuals traumatized by war or who experienced sexual abuse
- At the same time: Can sharing data help to protect these groups from being over-researched (possibly re-traumatization)?
  - Trauma patients support data sharing "to help others", when purpose and process of sharing is explained (Campbell et al., 2023; Lin et al., 2019)

10.1.2 "Closed doors"

- cases in which field access is obstructed or denied by the data provision (Prosser et al., 2023)
- cases in which sharing reduces the reduces the willingness to participate

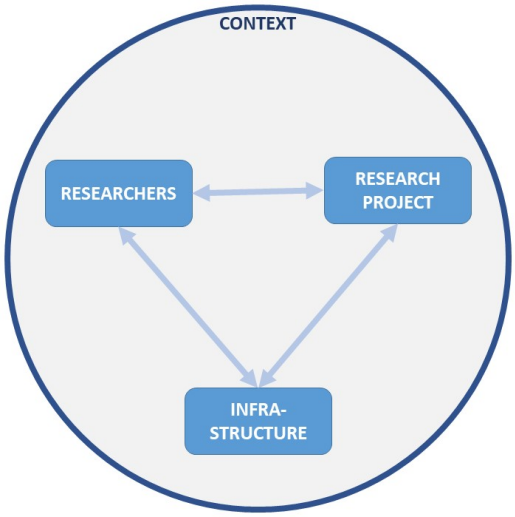
10.1.3 Costs > benefit

- e.g., low reuse potential -> publish for reproducibility
- e.g., when obtaining permission from the school authorities is extremely complicated

- e.g., epistemic problem: context of data collection is highly relevant and other researchers "haven't been there" (Mauthner et al., 1998) -> publish for intersubjective comprehensibility
- e.g., can't anonymize data -> synthpop, create input-output-documents via RMarkdown/Quarto

10.2 It's not all your responsibility

- Responsibility of opening research is a **collective responsibility** in the **"research ecosystem"** (European Commission, 2018; RfII, 2019)
- Researchers are just one part of this



Infrastructure

- Does suitable infrastructure exist?
- Is it "easy to use" and cheap?
- Is it tailored to my needs and type of data?
- Does it allow the implementation of FAIR data?
- Are there resources to support data sharing?

Context

- Do scientific societies, journals, or research funders encourage sharing?
- Is it common practice ("culture") in my field of research to share data? (Bishop, 2006)
- Are there standards established for data sharing?
- Do ethics committees request detailed reasoning for the intent to collect own data as opposed to reusing?

Research project

See reasons above on

- Marginalized/vulnerable groups
- "Closed doors"
- Costs > benefit

Researchers

- Willingness and concerns toward sharing data (Mozersky et al., 2021)
- Knowledge, experiences and skills with relevant processes

Questions to be answered at the end?
Please put them here!

Part IV

Workflow

11 Overview

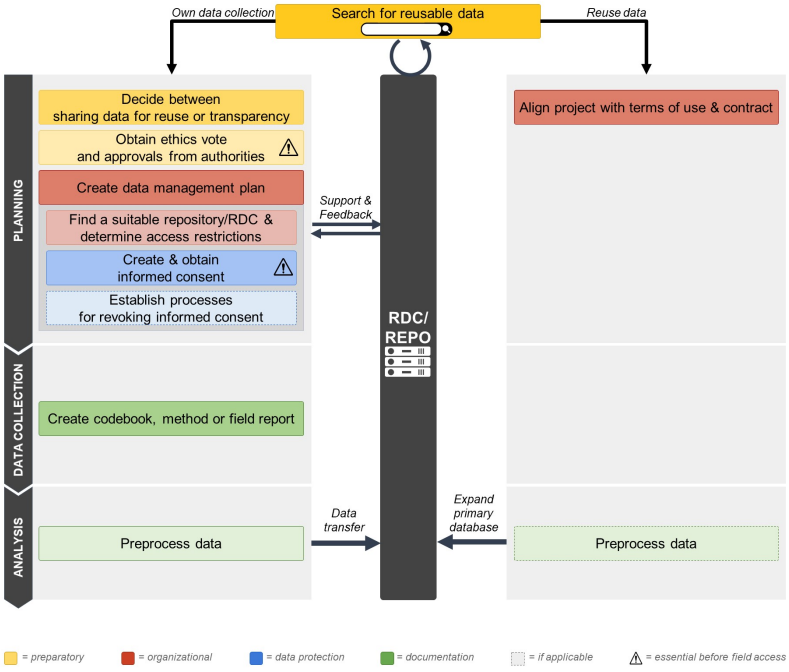

Own data collection
Search for reusable data
Reuse data
PLANNING
Decide between sharing data for reuse or transparency
Obtain ethics vote and approvals from authorities
Create data management plan
Find a suitable repository/RDC & determine access restrictions
Create & obtain informed consent
Establish processes for revoking informed consent
Support & Feedback
RDC/ REPO
Align project with terms of use & contract
DATA COLLECTION
Create codebook, method or field report
ANALYSIS
Data transfer
Expand primary database
Preprocess data
Preprocess data
= preparatory
= organizational
= data protection
= documentation
= if applicable
= essential before field access

Table: Steps & resources

What	Resources
Search for reusable data	Research data centers have searchable databases re3data.org, Verbund FDB, Open ICPSR (USA)UK Data Archive (UK)
Sharing for reuse or transparency	Costs: How much effort is required? Consent for reuse available?Benefits: Analysis potential, Quality of data
Data management plan	Templates/ToolsDMP template (ERC)Online tool (Open Aire)Online tool “DataWiz” (ZPID)StandardD-Psy-FAIR (ZPID)

Decide for a repository or RDC	Research data center:Search: re3data.orgVerbund FDB (education, Germany)Open ICPSR (USA)UK Data Archive (UK)RepositoriesZenodoosf.io

Informed consent	Recommendations of ICPSR Information, samples and model forms of UK Data Service Information, samples and model forms of VerbundFDB Good overview with Dos and Don'ts, University of UtrechtOverview + links (ZPID), EnglishExplanations + definitions (Michigan Tech), USChecklist, GermanTemplate, German standard languageTemplate, German plain languageTemplate (qualitative data), GermanExplanations + template (DGPs), German

Create codebook	R Package `codebook`codebook from DataWiz Institute of Education Science's "Guidelines for a Codebook"Information by ICPSR on "What is a Codebook?"

Questions to be answered at the end?
Please put them here!

12 Workflow 1

12.1 Search for reusable data

Why?
For resource intensive data collections this could save you a lot of time and money

Resources
Research data centers have searchable databases

- re3data.org (Database to search for databases)
- Verbund FDB (Germany)
- Open ICPSR (USA)
- UK Data Archive (UK)

12.2 Decide between sharing for reuse or reproducibility

Why?

- Providing data at a reseach data center costs time and money for you and the data center
- Typically,
  - sharing for **reuse** purposes is suited for **research data centers**
  - sharing for **reproducibility** purposes is suited for **repositories**

Resources
I am not aware of any standards to make this decision. Here are a couple of guidelines to decide, if your data is fit for reuse

- Costs:
  - How much **effort** is required for well-documented data sharing (e.g., Does a codebook exist? What steps are necessary for data cleaning/editing?)
  - Is there **consent for reuse** available or would it have to be obtained retrospectively?
- Benefits:
  - Analysis **potential** (e.g., not fully analyzed, type of data, connected with other data sources)
  - **Quality** of data (e.g., representative, size, special features of sample)

Questions to be answered at the end?
Please put them here!

13 Workflow 2

13.1 Write a data management plan

Why?

- supports researchers in the process of generating FAIR research data
- ensures good scientific practice

Resources
Templates and online tools for specific applications

- For rapid documentation: DMP template (European Research Council) with four open questions
- Online tool for machine-readable DMPs (Open Aire)
- Online tool supporting the creation of a DMP (ZPID)

Standards

- D-Psy-FAIR (Blask et al., 2022)
  - Manual
  - Online Tutorial
  - Slides

13.2 Decide for a repository (or research data center)

Why?
Deciding on a specific repository or research data center early on helps to

- meet its requirements
- identify whether there are templates available
- identify whether the employees can support the sharing process (in the case or RDC)

Resources

- search for a research data center that fits your needs: re3data.org
- recommended research data centers: Verbund FDB (education, Germany), Open ICPSR (USA), UK Data Archive (UK)
- recommended repositories: Zenodo, osf.io, ...

Questions to be answered at the end?
Please put them here!

14 Workflow 3

14.1 Informed consent

Why?

- *Personal data* is subject to General Data Protection Regulation (GDPR)
- Informed consent must therefore fulfill a number of requirements. E.g.,
  - purpose of data collection (includes *sharing the data* and *future use*)➡ therefore often "broad consent"
  - participation is voluntary and without disadvantages
  - revocation is possible at any time (until anonymized)

Resources

- Recommendations of ICPSR
- Information, samples and model forms of UK Data Service
- Information, samples and model forms of VerbundFDB
- Good overview with Dos and Don'ts, University of Utrecht
- Overview + links (ZPID), English
- Explanations + definitions (Michigan Tech), US
- Checklist, German
- Template, German standard language
- Template, German plain language
- Template (qualitative data), German
- Explanations + template (DGPs), German

14.2 Decide for access restrictions

Why?

- Some data *cannot or should not be anonymized* (e.g., losing their reuse potential)
  - Therefore access needs to be restricted to certain groups (as defined in consent form)
- Some researchers fear being *scooped* (Laine, 2017)

With repositories...

restriction levels are usually limited to

- public (everybody sees everything)
- private (only you and your collaborators see everything)

With research data centers...

there are different restriction levels possible for different files (*in the same project*). Restriction levels depend on what the research data center offers.

Level	Prerequesite	For what
Public Use-file		anonymized data, codebooks, transcription rules
Student Use-file	Short application states use purpose	non-anonymized data with right to use for teaching

Scientific Use-file	Longer application states use purpose, handling of data, and data analyses; identification via PostIdent	non-anonymized data with right to use for research
Remote Access	… + access only via virtual machine	non-anonymized sensible data with right to use for research
Safe room	… + access only in person at research institute	non-anonymized very sensible data with right to use for research

An example: Project DESI, where

- codebooks are publicly accessible (files on the right side)
- video data are restricted for scientific use (files on the bottom of page)

Alternatives to restricting

- *Embargo period*
  - Specify a time period, before data go public
  - Possible with research data centers and some repositories
- *Exclude certain research questions* from reuse
  - Specify these research questions in the terms of use
  - Usually only possible with research data centers, except you are writing a very good license yourself
- Create *synthetic data* (e.g., with R package `synthpop`)
  - Mimics the properties of your data
  - Then possible to share this synthetic data set

Questions to be answered at the end?
Please put them here!

15 Workflow 4

15.1 Create codebook

Why?
Remember?

> Vagueness Makes Assessing Reproducibility a Nightmare
> most successful reproductions are predominantly the result of tedious and time-consuming work information about the provided raw data was often difficult to understand, and information about the relevant variables, data manipulations, and the used statistical model was often vague or inaccurate (Artner et al., 2021, p. 12)

Resources

- R Package `codebook`
  - semi-automated creation of a codebook (depending on how well prepared/labelled your data set is)
  - in combination with the formr survey framework, this package saves you a ton of time
  - still has some minor bugs, be prepared to mingle with it
- Codebook as a result of using DataWiz
- Institute of Education Science's "Guidelines for a Codebook"
- Information by ICPSR on "What is a Codebook?"

Questions to be answered at the end?
Please put them here!

Part V

Thematic focus

16 Focus: Can I share?

When we ask ourselves the question 'Can I share the research data from my project?', we find ourselves navigating a tension between

- reproducible, efficient and verifiable results, which represent a public good, and
- the protection of participants' privacy.

Both are legitimate concerns.

16.1 Does my data count as personal data?

16.1.0.1 Personal data is...

Your data set counts as personal data, if individuals can be identified from someone else looking at the data (see EU data protection rules).

This is the case, if

- it includes at least **one direct** identifier, like the name, email-address, postal address, ip-address, social security number, ...

- it includes a **combination of indirect** identifiers from which the identity of an individual can be reconstructed. Such as simultaneous measures on gender, age, the combination of subjects studied, study location, and the place of birth.

> **Indirect identifier**
>
> In general, the existence of **one** indirect identifier in the data set does not constitute personal data. You have to judge for yourself at what point information/variables can be combined to identify a person. But in the end it comes down to: Better be save than sorry.

16.1.0.2 Data anonymization

If you collected personal data, the GDPR applies, and you'll need the **consent** of the subjects to share the data (see next section). However, the GDPR does not apply to **anonymous data** and it can generally be shared without the need to obtain explicit consent for this purpose.

Absolute anonymity:

Data are modified by coarsening or removing variables (direct or indirect identifiers) to such an extent that identifying individual participants becomes impossible.

Factual anonymity:

Deanonymisation cannot be ruled out completely but the allocation of data to the respective statistical unit is only possible with an unreasonable effort in terms of time, cost and human resources.

An example

Anonymizing interview transcripts is feasible. However, it may be possible to identify individuals based on their conversational tone and style using advanced text mining techniques and highly trained algorithms.

16.1.0.3 Task

Task

1. Check your data: Are individuals directly or indirectly identifiable?
2. Does it make sense for you to anonymize the data?
    - Is absolute anonymity possible?
    - What specific steps would you have to take to achieve anonymisation?

16.2 Does my informed consent allow data sharing?

16.2.0.1 Requirements

The GDPR, only applies to personal data. This might mean that only with personal data you are obliged to obtain consent from your participants to share the data.

Besides the usual requirements (voluntary participation, possibility of revocation, …), the informed consent **requires the following aspects** in order to be able to share data:

- a statement that the data **is being shared**
- a statement of **whom** the data will be shared with
- a description of the **purposes** for which the data are used and reused

> **Possibility of revocation**
>
> The possibility of revocation is **only required as long as the data is not anonymous** (i.e. personal). After anonymization, it is no longer possible to trace which data belongs to a particular person and it can therefore not be deleted.

16.2.0.2 Broad consent

Unfortunately, we are not all-knowing. This includes the fact that we **don't know what others will do with the data** we share. So what should we write as descriptions for the purposes?

Example: Consent tailored for project

"The data is processed and analysed for the purpose of investigating the correlations between learning strategies and learning success during studies. After completion of the project, the data will be made available to other researchers via an accredited research data center."

Examples: Broad consent

"The data is processed and analysed for research purposes. After completion of the project, the data will be made available to other researchers via an accredited research data center."

"I give permission for the data that I provide to be deposited in an accredited research data center, so it can be used for future research and learning."

Good resources

Recommendations of ICPSR
Information, samples and model forms of UK Data Service
Information, samples and model forms of VerbundFDB
Very good overview with Dos and Don'ts of the University of Utrecht
Maintaining privacy with open data (Video of presentation by Felix Schönbrodt)

16.2.0.3 But wait...

Important as an assessment criterion is the respondent's **capacity for understanding** what s/he is consenting to. If the participant doesn't know future uses of the data, can we really speak of an **informed** consent? Maybe the direction of future analyses go against the participant's will.

Whom and how do we ask?

- Some studies suggest that participants **rarely read** the information or consent form carefully (Parfenova et al., 2024; Pedersen et al., 2011)
- Some studies suggest that *very* **easy language** may be better for understanding (and thus being informed) than "standard" language (Geier et al., 2021)
- The **minimum age** differs between EU countries, can go down to 13 years (Germany: 16, Austria: 15)
- With studies involving **pupils in schools**, additional regulations may come into effect that require parental consent.

These are challenging questions that need to be **evaluated for every research project separately** and can lead to some researchers deciding **not** to share the data.

16.2.0.4 Task

Task

1. Check your informed consent: Did you mention that data is being shared? Does the purpose mentioned allow other researchers to reuse the data (e.g. for meta-analyses, reproducibility-checks)?
2. Did you include any sentences similar to these, that hamper data sharing?
    - "The data will only be used by *members of the project group*."
    - "The data will be *deleted* after termination of the project / after 5 years."

16.3 Sharing (restriction) levels

16.3.1 Features of data

Clearly, there is **not only the dichotomy** of " public access" vs. " no access". Even within a study, it may be necessary to restrict access to the data to varying degrees.

E.g. if you are working with a (factually) anonymized quantitative data set and interview scripts that you have been asked to share only with researchers.

Restrictions (or selective sharing) may be relevant due to…

- data type: **videos** of people always count as personal data and might be shared only for research purposes
- data type: **interviews** are usually tough to anonymize and anonymization process might decrease analysis potential
- **Sensitive** data or data of **vulnerable groups** should be particularly protected against unlawful dual use
- **Student** data and data of **minors** is generally considered to be in need of protection

16.3.2 Examples of sharing levels

With repositories...

sharing levels are usually limited to

- public (everybody sees everything)
- private (only you and your collaborators see everything)

With research data centers...

there are different sharing levels possible for different files (*in the same project*). Sharing levels depend on what the research data center offers.

Level	Prerequesite	For what
Public Use-file		anonymized data, codebooks, transcription rules

Student Use-file	Short application states use purpose	non-anonymized data with right to use for teaching
Scientific Use-file	Longer application states use purpose, handling of data, and data analyses; identification via PostIdent	non-anonymized data with right to use for research
Remote Access	… + access only via virtual machine	non-anonymized sensible data with right to use for research
Safe room	… + access only in person at research institute	non-anonymized very sensible data with right to use for research

An example: Project DESI, where

- codebooks are publicly accessible (files on the right side)
- video data are restricted for scientific use (files on the bottom of page)

16.3.3 Alternatives to restricting

- *Embargo period*
  - Specify a time period, before data go public
  - Possible with research data centers and some repositories
- *Exclude certain research questions from reuse*
  - Specify these research questions in the terms of use
  - Usually only possible with research data centers, except you are writing a very good license yourself

- *Create synthetic data (e.g., with R package synthpop)*
  - Mimics the properties of your data
  - Then possible to share this synthetic data set

16.3.4 Task

Task

1. Imagine a study results in anonymized quantitative data and an interview script. You have not received consent to share the interview data, but would like to store it securely.
   - Check osf.io to see if you can create a project where you can make your **quantitative data publicly available** and **restrict access to your interview transscript** so that only your research team can access it.
   - Check zenodo.org if you can create a repository to share your **quantitative data publicly**, create another repository to **store your interview data with restricted access** and the **link the two repositories** defining their relationship.
2. Is there a research data center, where you would share data that needs restriction? Such as when you obtained consent to share the data, but only for research purposes.
   - Check the website of the research data center and try to find out **what their sharing/restriction levels are**

17 Focus: How to share?

17.1 Describe

Create a codebook!

17.1.1 Why?

When others, or your future self, engage with the dataset, they need to clearly comprehend the contents and structure of the data. This relates to the **reuse** part of FAIR data.

"Vagueness Makes Assessing Reproducibility a Nightmare
most successful reproductions are predominantly the result of tedious and time-consuming work information about the provided raw data was often difficult to understand, and information about the relevant variables, data manipulations, and the used statistical model was often vague or inaccurate" (Artner et al., 2021, p. 12)

Questions that might arise are:

- Which item label relates to which variable in the data set?
- What are the answer options (value labels) for this variable in the data set?
- Which items make a scale?
- Which items are reversed?
- …

17.1.2 Resources

Number of item	*Name of the variable in the data frame*	*Type of the variable in the data frame*	*Wording of the item in the survey*	*English translation of the item, not validated*	*Questionnaire/ Source of the item*	*Dimension the item belongs to (for scales)*	*Response format*	*Response l*
number	variable	**type**	**item**	**english_translation**	**questionnaire**	**dimension**	**response_format**	**labe**
1	BFI_extra_1	numeric	Ich gehe aus mir heraus, bin gesellig.	I get out of myself, I'm sociable.	BFI-2	extraversion	5-point rating scale	response_list
2	CFT-total	numeric	–	–	CFT-20-R	–	–	–
3	gender	factor	Bitte geben Sie ihr Geschlecht an.	Please indicate your gender.	generated ad-hoc	–	single choice	response_list
4	time	date	–	–	–	–	–	–

Example from (Horstmann et al., 2020)

Resources quantitative data

- How and why to write a codebook: Link to paper which includes an example
- From a research data center: Straight to the point answer on FAQ
- Institute of Education Science's "Guidelines for a Codebook"
- Information and overview by ICPSR on "What is a Codebook?"
- Option to semi-automatically create a codebook, particularly useful if you have labelled data: R codebook package
  - Here is my introduction to the package including troubleshoot tips (use your browser to automatically translate it)
  - There is also a codebook web app, in case you don't want to use R

Resources qualitative data

- Comprehensive guide on "Opening up and Sharing Data from Qualitative Research", see p. 13 on contextualization
- "A guide to field notes"
- Further reading: "Engaging the challenges of sharing qualitative research data"

17.1.3 Task

Task

1. Check out this codebook example created by the codebook R package
2. What do you like about the codebook? What do you think could be improved to meet the goal of understanding the data set?

17.2 Share

Transfer your data to a repository or research data center!

17.2.1 Where?

Search for research data centers:

re3data.org
(Use filters to narrow down the search results)

Repositories:

zenodo.org
osf.io

17.2.2 Task

Task

Either:

1. Search re3data.org:
    - Is there a research data center that **fits your**

needs?
    - E.g. What sharing levels does it offer?
    - E.g. Will they process requests for data access or do you have to approve each time?
2. To **explore the repositories**, upload an empty Excel-File to zenodo.org as well as osf.io.
    - How easy is the process of filling out metadata?
    - Which repository gives the option for more detailed metadata?

17.3 Connect

17.3.1 Why?

Enable researchers to **find** the data from the paper and the paper from the data, regardless of **entry point**.

For example:

- Researchers trying to **replicate** your study will easily find the data
- Researchers trying to use your study in a **meta-analysis** will easily find the data
- Researchers who know about your data will always be able to establish a **clear link**.
- Researchers stumbling over your data (e.g. on OSF) will easily be able to **understand the context** by accessing the paper
- ...

17.3.2 How to connect?

Ideally, we use **persistent identifiers**. In our field, this usually means using a **DOI**.

> **Advantages of DOIs**
>
> **Permanence**: URLs can change or become broken over time if a webpage is moved or deleted. A DOI is designed to stay the same forever, even if the content is moved to a different location.
> **Reliability**: Since DOIs are managed by official organizations (like Crossref, DataCite), they guarantee that the content will be accessible for a long time.
> **Easy Tracking**: DOIs make it easy for others to track, cite, and reference your work consistently.

- osf.io offers one DOI per "component", but only if you switched it to "public" (→ click "Create DOI" under the title of your component)
- zenodo.org offers one DOI per repository. You can also get a DOI for repositories that are restricted in their visibility. (→ Switch "Do you already have a DOI for this upload?" to "No")

Ideally

- You'll put the DOI in a **section** of your paper that is **not paywalled** (that can be "data availability", "open practices" or "supplemental material")
- You'll use **standard fields** in the repository to put the DOI that links to your paper

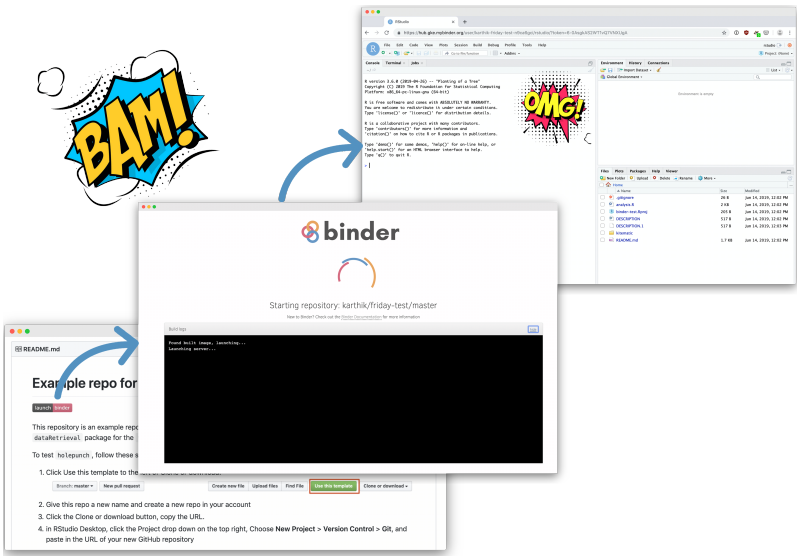
17.3.3 Task

Task

Imagine you'll need to upload a new version of your data set. E.g. after expanding your sample to include data on other populations.→ Simulate this by replacing your Excel file from the last task with a new excel file.

- Does the DOI change from before to after the upload?
- Is there an option to reference a specific version of your data set?
- Is there an option to automatically reference the latest version of your data set?

18 Focus: Tools and Resources



Holepunch

- **Share a copy of your R Studio project** via button click
- Accessible over **all platforms**: Project opens in browser
- **No installation** needed, only internet access and a browser
- *Best used in combination with renv or groundhog to secure package version stability*

Example

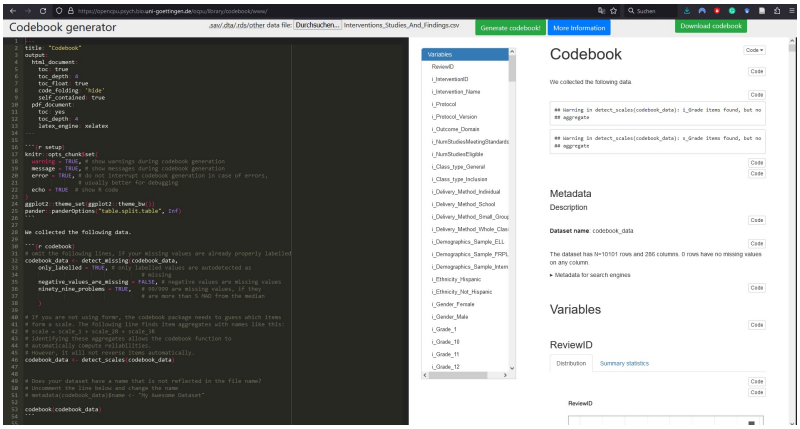
Go to this github repo and click on “launch binder”.

Enriched input-output-documents

- Enables data analysis and results to be reviewed **side-by-side**
- Possible to **add** formatted text, pictures, videos, …
- Several **layout options** (e.g. using tabs, table of content)
- with package xfun::embed_file() we can embed data *into* the HTML

Example

See this validation study of an instrument.



R codebook-package

- **semi-automatic** option to create a codebook
- labelled data → codebook with **minimal effort**
- See paper and introduction
- Not keen on using R? Use this web app

Examples

See

- the codebook of this small scale study or
- the codebook of the first sample from this larger study.

rix package

- The future of reproducibility?
- Creates a **self-contained** time-stamped capsule in which everything necessary is provided (like docker + renv at the same time)
- Uses the package manager **Nix**

> Introduction
>
> See this video of a workshop

Part VI

Reflection

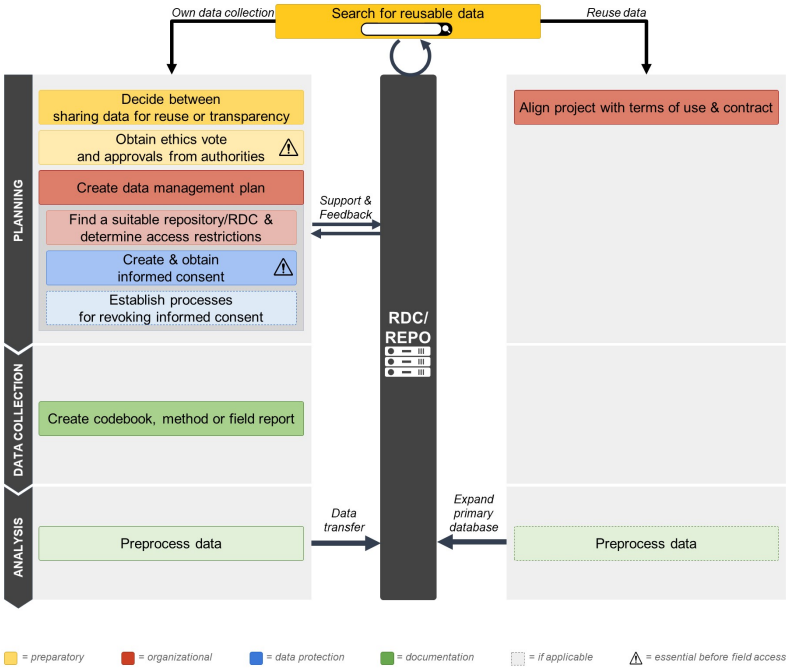
19 Exercise: Barriers

Let's assume: Sharing your data is possible. In one way or the other.

Why isn't everybody sharing all data "as open as possible as closed as necessary"?

1. Check out the flow chart again (see below)
2. Individually [3min]: Reflect on
   - What is a barrier/challenge to you?
   - What might be a powerful barrier/challenge for others?
3. In the breakout rooms [12min]: Discuss
   - What are the biggest barriers for you/others?
   - What would be different in an "ideal world" that would lead to you/others overcoming these barriers?
   - Document the barriers & your needs in this sheet

Questions to be answered at the end?
Please put them here!


Own data collection
Search for reusable data
Reuse data
PLANNING
Decide between sharing data for reuse or transparency
Obtain ethics vote and approvals from authorities
Create data management plan
Find a suitable repository/RDC & determine access restrictions
Create & obtain informed consent
Establish processes for revoking informed consent
Support & Feedback
RDC/ REPO
Align project with terms of use & contract
DATA COLLECTION
Create codebook, method or field report
ANALYSIS
Preprocess data
Data transfer
Expand primary database
Preprocess data
= preparatory
= organizational
= data protection
= documentation
= if applicable
= essential before field access

References

APA. (2017). *Ethical Principles of Psychologists and Code of Conduct*.

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Bishop, L. (2006). A Proposal for Archiving Context for Secondary Analysis. *Methodological Innovation Online*, *1*(2), 10–20. https://doi.org/10.4256/mio.2006.0008

Blask, K., Latz, M., Müller, M.-L., & Gellert, S. (2022). *D-PsyFAIR: Four simple steps to sustainable data documentation in psychology*. PsychArchives. https://doi.org/10.23668/PSYCHARCHIVES.12180

Burgard, T., Bosnjak, M., & Studtrucker, R. (2022). PsychOpen CAMA: Publication of community-augmented meta-analyses in psychology. *Research Synthesis Methods*, *13*(1), 134–143. https://doi.org/10.1002/jrsm.1536

Campbell, R., Goodman-Williams, R., Javorka, M., Engleton, J., & Gregory, K. (2023). Understanding Sexual Assault Survivors' Perspectives on Archiving Qualitative Data: Implications for Feminist Approaches to Open Science. *Psychology of Women Quarterly*, *47*(1), 51–64. https://doi.org/10.1177/03616843221131546

Clyburne-Sherin, A., Fei, X., & Green, S. A. (2019). Computational Reproducibility via Containers in Psychology. *Meta-Psychology*, *3*. https://doi.org/10.15626/MP.2018.892

CoARA. (2022). *Agreement on Reforming Research Assessment*.

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLOS ONE*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

DFG. (2015). *DFG Guidelines on the Handling of Research Data*.

DFG. (2019). *Guidelines for Safeguarding Good Research Practice. Code of Conduct*.

DGfE, GEBF, & GFD. (2020). *Gemeinsame Stellungnahme der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), der Gesellschaft für Empirische Bildungsforschung (GEBF) und der Gesellschaft für Fachdidaktik (GFD) zur Archivierung, Bereitstellung und Nachnutzung von Forschungsdaten in den Erziehungs- und Bildungswissenschaften und Fachdidaktiken*.

ERC. (2022). *Open Research Data and Data Management Plans. Information for ERC grantees*.

Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*, *10*, e67995. https://doi.org/10.7554/eLife.67995

ESRC. (2018). *ESRC Research Data Policy*.

European Commission. (2018). *OSPP-REC: Open Science Policy Platform Recommendations*. Publications Office.

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

Gabelica, M., Bojčić, R., & Puljak, L. (2022). Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, *150*, 33–41. https://doi.org/10.1016/j.jclinepi.2022.05.019

Gärtner, A., Leising, D., & Schönbrodt, F. D. (2022). *Responsible Research Assessment II: A specific proposal for hiring and promotion in psychology*. PsyArXiv. https://doi.org/10.31234/osf.io/5yexm

Geier, C., Adams, R. B., Mitchell, K. M., & Holtz, B. E. (2021). Informed Consent for Online Research—Is Anybody Reading?: Assessing Comprehension and Individual Differences in Readings of Digital Consent Forms. *Journal of Empirical Research on Human Research Ethics*, *16*(3), 154–164. https://doi.org/10.1177/15562646211020160

Gollwitzer, M., Abele-Brehm, A., Fiebach, C. J., Ramthun, R., Scheel, A., Schönbrodt, F., & Steinberg, U. (2021). Management und Bereitstellung von Forschungsdaten in der Psychologie: Überarbeitung der DGPs-Empfehlungen: DGPs-

Kommission „Open Science“ (beschlossen durch den Vorstand der DGPs am 26. 06. 2020). *Psychologische Rundschau*, *72*(2), 132–146. https://doi.org/10.1026/0033-3042/a000514

Horstmann, K. T., Arslan, R. C., & Greiff, S. (2020). Generating Codebooks to Ensure the Independent Use of Research Data: Some Guidelines. *European Journal of Psychological Assessment*, *36*(5), 721–729. https://doi.org/10.1027/1015-5759/a000620

Hussey, I. (2023). *Data is not available upon request* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jbu9r

IEA. (2022). *Sharing Study Data: A Guide for Education Researchers* (5f324947be2a4bf38e9ec89df7144f0e). Institute of Education Sciences.

Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). *The practice of reproducible research: Case studies and lessons from the data-intensive sciences*. University of California Press.

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

Laine, H. (2017). Afraid of Scooping – Case Study on Researcher Strategies against Fear of Scooping in the Context of Open Science. *Data Science Journal*, *16*, 29. https://doi.org/10.5334/dsj-2017-029

Lin, Y.-K., Liu, K.-T., Chen, C.-W., Lee, W.-C., Lin, C.-J., Shi, L., & Tien, Y.-C. (2019). How to effectively obtain informed consent in trauma patients: A systematic review. *BMC Medical Ethics*, *20*(1), 8. https://doi.org/10.1186/s12910-019-0347-0

Logan, J. A. R., Hart, S. A., & Schatschneider, C. (2021). Data Sharing in Education Science. *AERA Open*, *7*, 233285842110064. https://doi.org/10.1177/23328584211006475

Mauthner, N. S., Parry, O., & Backett-Milburn, K. (1998). The Data are Out there, or are They? Implications for Archiving and Revisiting Qualitative Data. *Sociology*, *32*(4), 733–745. https://doi.org/10.1177/0038038598032004006

Mozersky, J., McIntosh, T., Walsh, H. A., Parsons, M. V., Goodman, M., & DuBois, J. M. (2021). Barriers and facilitators to qualitative data sharing in the United States: A survey of qualitative researchers. *PLOS ONE*, *16*(12), e0261719. https://doi.org/10.1371/journal.pone.0261719

NSF. (2023). *Proposal & Award Policies & Procedures Guide (PAPPG) (NSF 23-1)*.

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229–237. https://doi.org/10.1177/2515245920918872

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

Parfenova, D., Niftulaeva, A., & Carr, C. T. (2024). Words, words, words: Participants do not read consent forms in communication research. *Communication Research Reports*, 1–11. https://doi.org/10.1080/08824096.2024.2379832

Pedersen, E. R., Neighbors, C., Tidwell, J., & Lostutter, T. W. (2011). Do Undergraduate Student Research Participants Read Psychological Research Consent Forms? Examining Memory Effects, Condition Effects, and Individual Differences. *Ethics & Behavior*, *21*(4), 332–350. https://doi.org/10.1080/10508422.2011.585601

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175

Prosser, A. M. B., Hamshaw, R. J. T., Meyer, J., Bagnall, R., Blackwood, L., Huysamen, M., Jordan, A., Vasileiou, K., & Walter, Z. (2023). When open data closes the door: A critical examination of the past, present and the potential future for open data guidelines in journals. *British Journal of Social Psychology*, *62*(4), 1635–1653. https://doi.org/10.1111/bjso.12576

RfII. (2019). *Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel*.

Schneider, J., Rosman, T., Kelava, A., & Merk, S. (2022). Do Open-Science Badges Increase Trust in Scientists Among Undergraduates, Scientists, and the Public? *Psychological Science*, *33*(9), 1588–1604. https://doi.org/10.1177/09567976221097499

Schönbrodt, F., Gärtner, A., Frank, M., Gollwitzer, M., Ihle, M., Mischkowski, D., Phan, L. V., Schmitt, M., Scheel, A. M., Schubert, A.-L., Steinberg, U., & Leising, D. (2022). *Responsible Research Assessment I: Implementing DORA*

for hiring and promotion in psychology. https://doi.org/10.23668/PSYCHARCHIVES.8162

Steinhardt, I., Fischer, C., Heimstädt, M., Hirsbrunner, S. D., İkiz-Akıncı, D., Kressin, L., Kretzer, S., Möllenkamp, A., Porzelt, M., Rahal, R.-M., Schimmler, S., Wilke, R., & Wünsche, H. (2021). *Opening up and Sharing Data from Qualitative Research: A Primer: Results of a workshop run by the research group „Digitalization and Science" at the Weizenbaum Institute in Berlin on January 17, 2020*. https://doi.org/10.34669/WI.WS/17

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828

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