I. Data documentation: Benefits and how to do

Katarina Blask & Erich Weichselgartner
Leibniz Institute for Psychology (ZPID)
Trier, Germany

You would like to archive and/or share your research data because

- Of funders' requirements (e.g., EU H2020)
- Of journals' policies (e.g., The Royal Society journals)
- Of recommendations by learned societies (e.g., German PS)
- Of "good research practice" (e.g. UKRI)
- You're an Open Science advocate (e.g., Community Networks)
    - The UNESCO Recommendation on Open Science has been adopted by Member States at the Science Commission plenary at its 41st General Conference (Item 8.1, 15 November 2021) (→International Science Council)

Leibniz Association

leibniz-psychology.org

Benefits of data sharing?

Leibniz Association

leibniz-psychology.org

Benefits of data sharing (why do you deserve a badge)✝

- Avoid unnecessary duplication of data collection
- Save time and money of respondents and of researchers
- Reanalysis: Verification (same problem, same data) [Reproducibility]

Confusing terminology: Reproducibility vs. Replicability.

"Everyone agrees that reproducibility and replicability are fundamental characteristics of scientific studies. But there are no formal definitions for these concepts, which leads to confusion since the same words are used for different concepts by different people in different fields."

(Patil, Peng, & Leek, 2016)

✝ Weichselgartner, E. (2008, Nov.) PsychData: An archive for primary research data in Psychology (invited talk). Keeping the Records of Science Accessible: Can we afford it? High-level strategic conference organized by the Alliance for Permanent Access, the European Science Foundation (ESF) and the Hungarian Scientific Research Fund (OTKA), Budapest, Hungary.

Reproducibility

"Reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use **the same raw data** to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results." (Bollen et al, 2015)

Reproducibility defined in this way mainly addresses issues of trust that data and analyses are as represented. "The idea in data sharing is to permit other researchers to reanalyze **the same data**." (Firebaugh, 2007)

Example

Of the 35 Registered Reports published in psychology that made **data** and code openly available, 57% were computationally reproducible✝ compared with 31% in a previous analysis of regular articles. “However, **suboptimal data curation**, unclear analysis specification and reporting errors can impede analytic reproducibility, undermining the utility of **data sharing** and the credibility of scientific findings.” (Hardwicke et al, 2018)

✝Computational reproducibility: Obtaining consistent results using **the same input data**, computational methods, and conditions of analysis (cf. NAS, 2019)

Replicability

Replicability refers to the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but **new data** are collected. New evidence is provided by new experimentation. (Bollen et al, 2015)

Replicability refers to the replication of scientific findings using independent investigators, methods, data, equipment, and protocols. (Peng, 2009)

"Whenever psychologists undertake large projects, like Many Labs 2, in which they replicate past experiments en masse, they typically succeed, on average, half of the time." (The Atlantic)

Crisis

Leibniz Association

leibniz-psychology.org

Example

2019

Many Labs 4: Failure to replicate *mortality salience effect* with and without original author involvement

(Klein, Dec 10th, 2019)

2021

"(Another) … 5 failed replications of ***mortality salience effects*** … it's dead …" (Lakens, Nov 16th, 2021)

Leibniz Association

leibniz-psychology.org

Reproducibility ⇒ Share your data

Replicability ⇒ Share your study-level documentation

Leibniz Association

leibniz-psychology.org

Benefits of data sharing (why do you deserve a badge)✝

- Avoid unnecessary duplication of data collection
- Save time and money of respondents and of researchers
- Reanalysis: Verification (same problem, same data)
- Secondary analysis (different problem, same data)
- Meta-analysis (same problem, several independent data sets)
- Refinement (alternative analyses)
- Testing the generality of research findings
- Create new enlarged databases
- Increase the amount of data available on any scientific question
- Applying new theories to existing data

Leibniz Association

leibniz-psychology.org

Benefits of data sharing (continued)

- Provision of resources for training
  - The reanalysis of previously collected data is one of the best ways of teaching research methodology
  - Secondary data are models for collecting one's own data✝
- Monitor historical changes
- Protection against faulty data
- Sharing research data is associated with increased citation rate
- Make data sets citable as scholarly publications (see DataCite)

✝ Sobal, J. (1982). The Role of Secondary Data Analysis in Teaching the Social Sciences. Library Trends, 30, 479-488.

Leibniz Association

leibniz-psychology.org

Benefits of data sharing (continued)

- Provision of resources for training
  - The reanalysis of previously collected data is one of the best ways of teaching research methodology

13 years later: “It is not uncommon to work with real datasets and code at the BSc/MSc level [for example, in high energy physics and some branches of economics and the life sciences]. However, across the board, training on reproducibility is not an established part of student curricula.”

(Chiarelli, Andrea, Loffreda, Lucia, & Johnson, Rob. (2021, Nov). The Art of Publishing Reproducible Research Outputs: Supporting emerging practices through cultural and technological innovation. Zenodo. https://doi.org/10.5281/zenodo.5521077)

What are research data?

- “Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.” (OMB CIRCULAR A–110)
- Research data in psychology are frequently represented as tabular data in flat rectangular files: Columns represent different variables, while rows represent different observations (e.g. subjects).

Leibniz Association

Research data in psychology are frequently represented as tabular data in flat rectangular files: Columns represent different variables, while rows represent different observations (e.g. subjects).

leibniz-psychology.org

pt1	pc	nitems	time	session	stor	age
452	0.142857	14	0.452	part 1	1	young
524	0.428571	14	0.524	part 1	1	young
596	0.538462	13	0.596	part 1	1	young
668	0.538462	13	0.668	part 1	1	young
776	0.636364	11	0.776	part 1	1	young
884	0.642857	14	0.884	part 1	1	young
1028	0.933333	15	1.028	part 1	1	young
1172	0.866667	15	1.172	part 1	1	young
1352	0.875	16	1.352	part 1	1	young
1568	0.954545	22	1.568	part 1	1	young
2432	0.904762	21	2.432	part 1	1	young

What do the values mean?

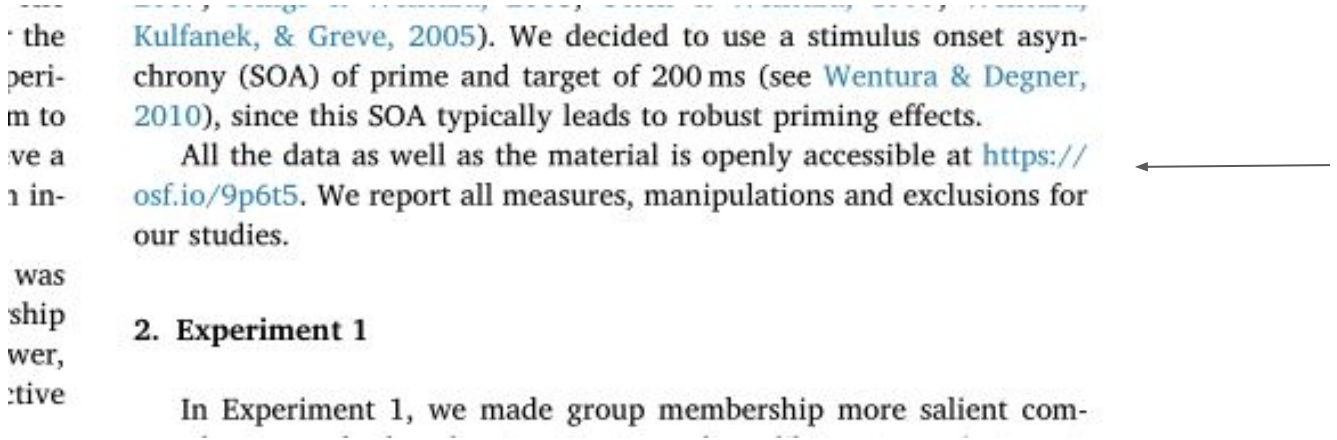
Oberauer, K., & Kliegl, R. (2004). Beyond resources - formal models for complexity effects and age differences in working memory. Primary data of the memory updating experiment. https://doi.org/10.5160/psychdata.orks96fo20

Why do I need to *document* research data?

- “Isn’t all the necessary info in the paper?” (see Video)
- “Just ask me anytime, I’ll remember” (see Video)

Data are published as **stand-alone objects** in repositories

[→re3data.org - Registry of Research Data Repositories]



the
peri-
m to
ve a
n in-

was
ship
wer,
tive

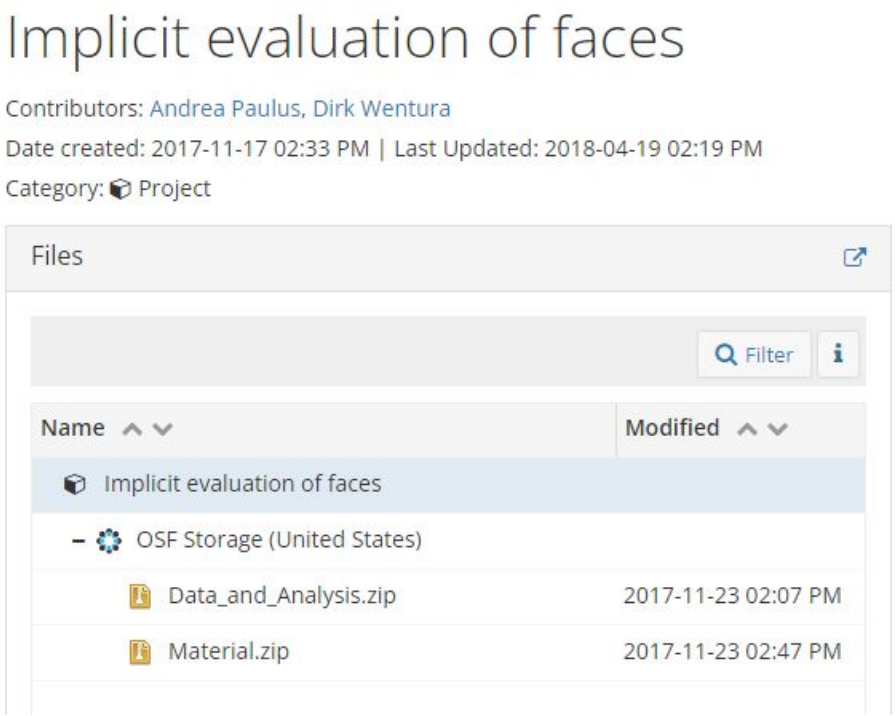
Kulfanek, & Greve, 2005). We decided to use a stimulus onset asynchrony (SOA) of prime and target of 200 ms (see Wentura & Degner, 2010), since this SOA typically leads to robust priming effects.

All the data as well as the material is openly accessible at https://osf.io/9p6t5. We report all measures, manipulations and exclusions for our studies.

2. Experiment 1

In Experiment 1, we made group membership more salient com-

"All the data as well as the material is openly accessible at https://osf.io/9p6t5"



"All the data as well as the material is openly accessible at https://osf.io/9p6t5"

Data_and_Analysis.zip

- Data_and_Analysis/**Analysis.sps**
- Data_and_Analysis/**Exp1.sav**
- Data_and_Analysis/**ImplicitEvaluation_Exp1.pl**
- Data_and_Analysis/**Raw_Exp1.txt**
- Data_and_Analysis/**rt_100_1000_Exp1.dat**
- Data_and_Analysis/**ReadMe.txt**

Please note:
In Experiment2 and Experiment3 some participant numbers are missing in the data file ...

Lack of documentation is common! ✝

✝ Günther, A., Kerwer, M. & Weichselgartner, E. (2017, Jul) Repositories for psychological research data: Overview and quality criteria. European Congress of Psychology, Amsterdam, The Netherlands..

What does “good” archiving/sharing of (digital) research data require?
Issues:

- File formats (.sav versus .csv)
- Naming conventions (e.g., xyz32nz.txt versus `experiment-1_data.csv`)
- Persistent storage (not your USB-drive, see Video)
- Ethics approval, participants’ consent
- Security (protect individuals' personal data)
- Metadata
- …

Use community-accepted, non-proprietary (open) norms and standards

Leibniz Association

leibniz-psychology.org

Additional requirements for sharing: Legal issues.

- Anonymization
- Clarification of rights (e.g., data ownership)
- Storage location (repository): European data protection law
- License for your data (e.g., Creative Commons)
- …

Leibniz Association

leibniz-psychology.org

Data documentation: Describing your data

- Why and how did you collect your data? When and where? Who was involved?
- Provide as much context and materials as possible
- You can do this by sharing supplemental information and by using metadata

From the DataWiz help text

Supplementary information is material that helps understanding how you carried out your research and evaluating your scientific claims. Materials can be publications (pre- or post-prints), reports, lab books, stimuli, computer code, data management plans, ethics approval forms, consent forms, pre-registration documents, multimedia files or anything that helps others† to understand your study.

† Or yourself a few years from now …

Metadata

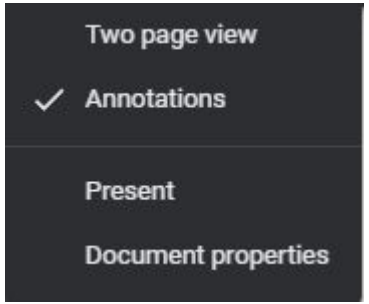
Metadata is data about data, i.e. metadata describes other data. Metadata uses standardised terms and is presented in a structured way. There are general and discipline specific metadata. Metadata is not just for description and discovery, but we will focus on these purposes here.

Leibniz Association

leibniz-psychology.org

Example metadata, PDF file

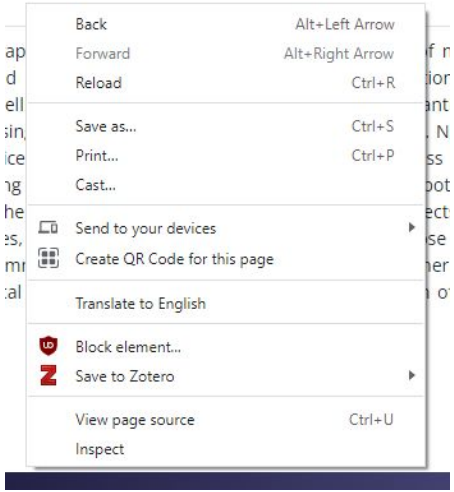
https://journals.sagepub.com/doi/pdf/10.1177/1940161220937239



Example metadata, HTML file

https://ejop.psychopen.eu/index.php/ejop/article/view/5391

Mouse right click



Leibniz Association

leibniz-psychology.org

Examples of Metadata standards relevant to psychology

General

- The Dublin Core (DC) Metadata Element Set: A set of fifteen "core" elements (properties) for describing digital as well as physical resources such as books, music scores or … research data

Disciplinary

- The Data Documentation Initiative (DDI) standard for describing study-level information in the social, behavioral, economic, and health sciences.
- The Brain Imaging Data Structure (BIDS) is a standard for organizing, annotating, and describing data collected during neuroimaging experiments.
- Psychology?

Leibniz Association

leibniz-psychology.org

Psychology? Project PsyCuraDat: Developing user-oriented curation criteria for psychological research data✝

What kind of information do researchers need in order to understand and interpret the data of other researchers, e.g. for reanalysis, secondary analysis or meta-analysis?

Existing documentation “standards” used in Psychology (e.g. DC, DDI) are derived from Computer Science, Library and Information Science, and the Social Sciences. But what do psychologists need? Find out empirically.

The DataWiz documentation schema is adopted from preliminary results of PsyCuraDat. Katarina will introduce the schema.

✝ Grant 16QK08 (2019-2022) by the German Federal Ministry of Education and Research awarded to Erich Weichselgartner and after his retirement transferred to Dr. Katarina Blask.

Leibniz Association

leibniz-psychology.org

Katarina

Leibniz Association

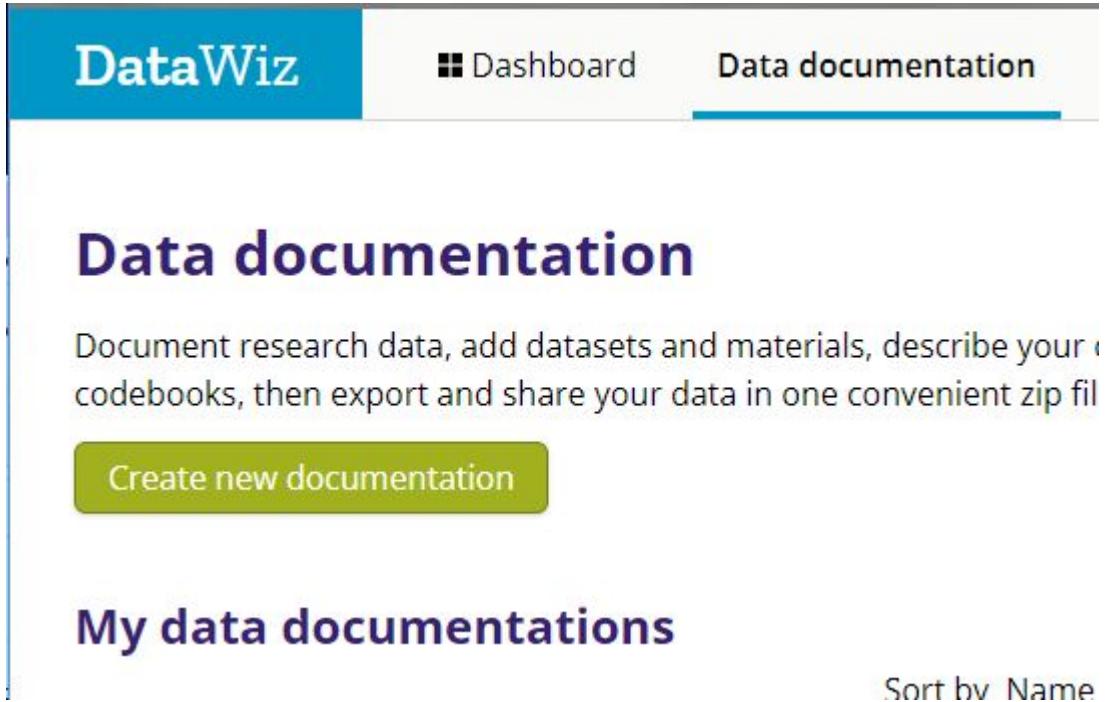
leibniz-psychology.org

Data sharing / data documentation is “extra” work and takes time and effort! How can we support researchers with these tasks?

There are two design goals behind DataWiz:

- Since researchers know their data best **they** should do the documentation
- A web-based structured tool with built in guidance and help can on the one hand assure that documentation is sufficient and on the other reduce the extra workload

https://datawiz2.dev.zpid.de/



End of Part I.
Thank You!

ZPID and DataWiz are funded by

RheinlandPfalz
MINISTERIUM FÜR WISSENSCHAFT, WEITERBILDUNG UND KULTUR

Bundesministerium für Gesundheit

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Leibniz Association

leibniz-psychology.org

References (see also “Suggested Reading”)

- K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, J. L. Olds (2015). Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science (National Science Foundation, Arlington, VA).
- Firebaugh, G. (2007). Replication data sets and favored-hypothesis bias: Comment on Jeremy Freese (2007) and Gary King (2007). Sociological methods & research, 36(2), 200-209.
- Günther, A., Kerwer, M. & Weichselgartner, E. (2017, Jul) Repositories for psychological research data:
- Overview and quality criteria. European Congress of Psychology, Amsterdam, The Netherlands..
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. Royal Society open science, 5(8), 180448.
- National Academies of Sciences: Reproducibility and Replicability in Science. Washington (DC): National Academies Press (US); 2019 May 7. 3, Understanding Reproducibility and Replicability. Available from: https://www.ncbi.nlm.nih.gov/books/NBK547546/
- Peng, R. D. (2009). Reproducible research and biostatistics. Biostatistics, 10(3), 405-408. https://doi.org/10.1093/biostatistics/kxp014
- Weichselgartner, E. (2008, Nov.) PsychData: An archive for primary research data in Psychology (invited talk). Keeping the Records of Science Accessible: Can we afford it? High-level strategic conference organized by the Alliance for Permanent Access, the European Science Foundation (ESF) and the Hungarian Scientific Research Fund (OTKA), Budapest, Hungary.