Open Data: Good habits and best practices for effective research data management

Nora Wickelmaier

Leibniz-Institut für Wissensmedien

February 13, 2023

1 Introduction

Why are we here today?

Interaction with slido

Or go to `https://www.slido.com/` and enter `#TOSI2023`

Habits

> *"In the beginning, creating a new habit is more critical than actually achieving a goal."*

Six ideas for building the habits you want

1. Start your habit change process by building awareness
2. All change begins with making choices
3. Attach a new habit or behavior to something you already do regularly
4. Gain clarity about what you want to do and how you will do it
5. Start with a simple step
6. Remember the "why"

https://www.psychologytoday.com/us/blog/flourish-and-thrive/202002/6-powerful-ways-build-new-habits

The "why"

Reproducibility vs. replicability:

		Data: Same	Data: Different
Analysis	**Same**	**Reproducible**	Replicable
	Different	Robust	Generalizable

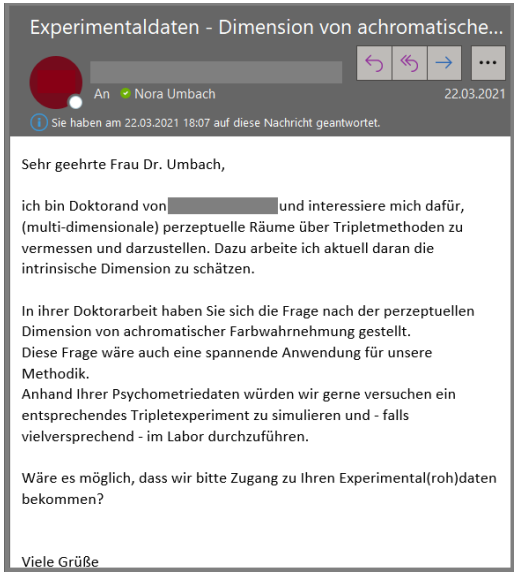
Ethical research standards:

https://the-turing-way.netlify.app/reproducible-research

Barriers

- Skills for doing reproducible research are not taught in a systematic way
- Supervisors are often not doing it
- Incentive system does not encourage to spend time on making research reproducible (yet!)
- Takes time
- Takes time
- Requires additional skills
- Learning these skills is often full of frustrating experiences

2 Personal example



Experimentaldaten - Dimension von achromatische...

An Nora Umbach 22.03.2021

Sie haben am 22.03.2021 18:07 auf diese Nachricht geantwortet.

Sehr geehrte Frau Dr. Umbach,

ich bin Doktorand von [redacted] und interessiere mich dafür, (multi-dimensionale) perzeptuelle Räume über Tripletmethoden zu vermessen und darzustellen. Dazu arbeite ich aktuell daran die intrinsische Dimension zu schätzen.

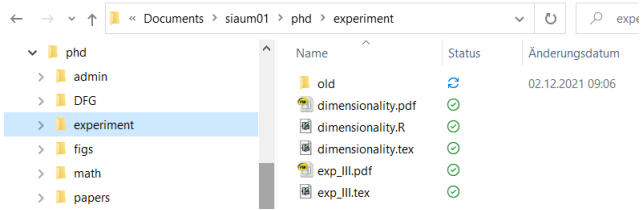
In ihrer Doktorarbeit haben Sie sich die Frage nach der perzeptuellen Dimension von achromatischer Farbwahrnehmung gestellt.
Diese Frage wäre auch eine spannende Anwendung für unsere Methodik.
Anhand Ihrer Psychometriedaten würden wir gerne versuchen ein entsprechendes Tripletexperiment zu simulieren und - falls vielversprechend - im Labor durchzuführen.

Wäre es möglich, dass wir bitte Zugang zu Ihren Experimental(roh)daten bekommen?

Viele Grüße

The situation

- I actually published the data of my first experiment in an R package
- BUT: He probably wants the data of my second experiment. . . 🙈
- First try:

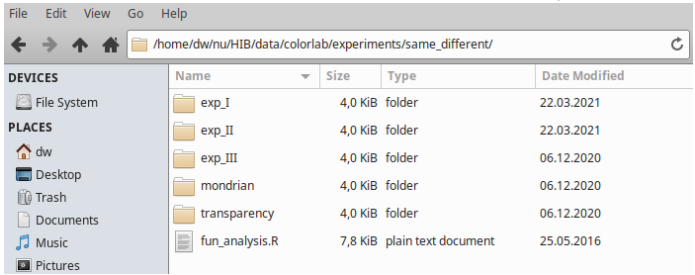


This does not look good. . . There's not even the actual folder with the final data. WTF?

- Remembering that I probably moved that to my "postdoc folder" – but this is not on my work computer, since it is so big. . . OK, I will check that at home tonight. . .

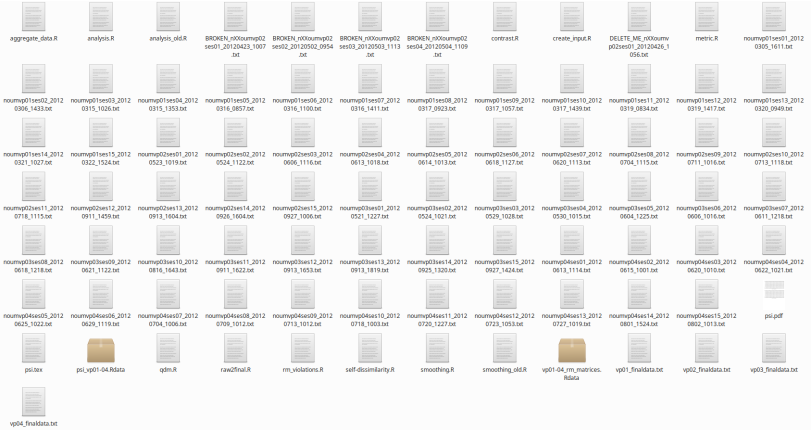
Getting closer I

- Puh, all the data are on my home computer (only copy, though. . .)



Getting closer II

- But not really organized...

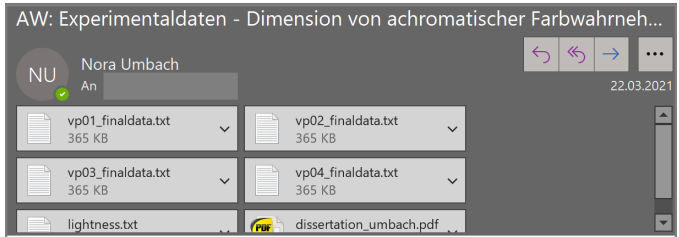


Getting closer III

- I check my analysis files and am pretty sure that the files `vp0[1-4]_finaldata.txt` are the ones I used in the analysis in my diss
- Checking the files again, I discover that the luminances of my stimuli are not in these files – only the stimulus names I used in the experiment, something like `stim_1_7` (pretty informative, huh?)
- The stimulus files are also in the folder, the one for `stim_1_7` looks like this

(Btw, I investigated the perception of black and white colors in my diss)

I finally answered this



Exercise

- Go to `https://nextcloud.iwm-tuebingen.de/s/8KoefDc6tZSSMwy` and download the data and additional material
- Are you able to understand what needs to be done in order to use this data based on the information provided?
- Write down the steps that need to be taken in order to make these data reusable
- What kind of skills do we need in order to perform these steps?

What I wish I could have answered

```
Hello,

All the data and analyses for my dissertation can be
found here:
https://www.mathpsy.uni-tuebingen.de/colorlab/

Let me know if you need anything else.

Best wishes,
Nora
```

What barriers stopped me from doing this?

- Back then, I did not even consider to publish my data
- (I only published the data for the first experiment so I had some data in my R package)
- Back then, I only wrapped up stuff before switching research topics for my postdoc phase
- However, I took some time to clean up the files, which allowed me to answer the request within one day
- I had most of the skills I needed, but nobody who emphasized how important it might be to make the data and analysis scripts available and reproducible

3 Workflow

What is a workflow?

> *A workflow consists of an orchestrated and repeatable pattern of activity, enabled by the systematic organization of resources into processes that transform materials, provide services, or process information.*

`https://en.wikipedia.org/wiki/Workflow`

Important aspects:

- Repeatable pattern
- Systematic organization
- Transformation processes

In short:

- A workflow answers the question: **What's the most efficient way to get this work done?**

Why do I need a workflow?

- It boosts productivity
- It reduces mental load
- A truly optimized workflow will:
  - Identify and remove unnecessary steps and processes that lead to slowdowns
  - Provide a sequential (chronological) order for accomplishing tasks
  - Automate some decisions and processes (freeing up time)
  - Reduce communication burdens (fewer e-mails, meetings, etc.)
  - Encourage collaboration
  - Track progress and assess performance
  - Keep records of previous processes and make future processes repeatable
  - Eliminate decision fatigue

Where to get started?

- Read Lowndes et al. (2017) – it's eye opening (and kinda funny)
- Consider your current research data management and think about what your current workflow is:
  - What is going well?
  - What could be improved?
  - What could be the benefits of an improved workflow in this area?

Project workflow

- Project workflow refers to how you organize projects and move through the various stages of the research cycle
- Kathawalla et al. (2021) say that a project workflow includes:
  - File folder structure
  - Document naming conventions
  - Version control
  - Cloud storage
  - Choice of who has access to a project and when (Collaborators? Public?)
- Developing a clear project workflow is much easier for PhD students than later career scholars who have many more projects to organize

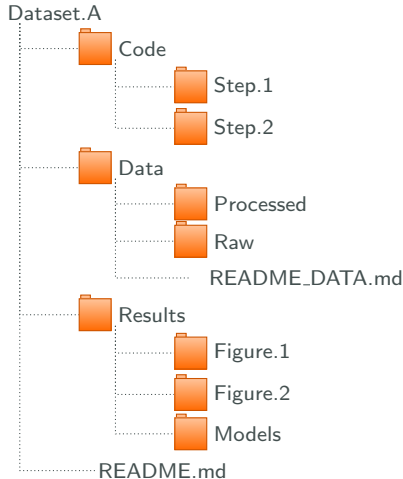
4 Folder structure

The basics

- One top level folder for each project
- Is anybody else working with you on this project?
- Will someone have to understand your system later? (Always imagine that someone has to!)
- Capture metadata about contents of folders and files
- Create README files for different levels
- Do not nest too deep
- Try to find a structure that works for more than one project

Directory structures

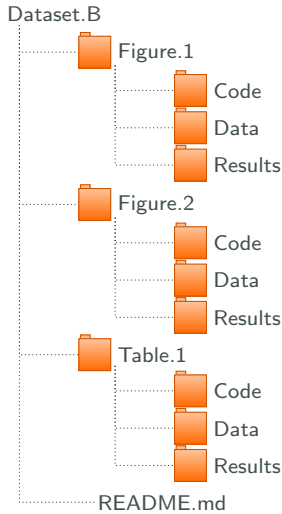
Organized by file type



https://datamanagement.hms.harvard.edu/collect/directory-structure

Directory structures

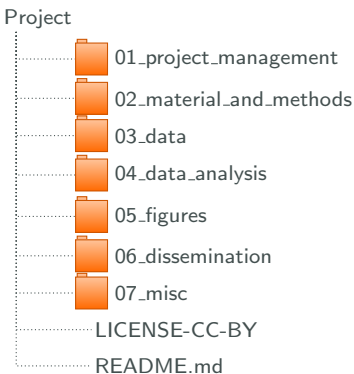
Organized by analysis



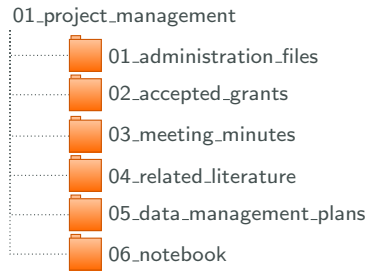
https://datamanagement.hms.harvard.edu/collect/directory-structure

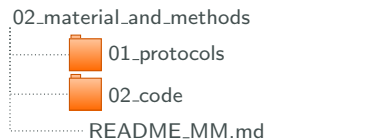
Research folder structure standard

Template for (Neuro)Science



Subfolders





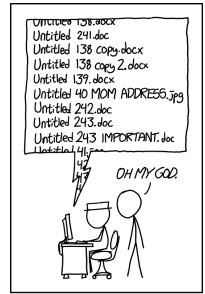
https://github.com/tonic-team/Tonic-Research-Project-Template

5 Naming conventions

Examples

- Files with no naming convention:

```
Test data 2016.xlsx
Meeting notes Jan 17.doc
Notes Eric.txt
Final FINAL last version.docx
```



- Files with naming convention:

```
20160104_ProjectA_Ex1Test1_SmithE_v1.xlsx
20160104_ProjectA_MeetingNotes_SmithE_v2.docx
ExperimentName_InstrumentName_CaptureTime_ImageID.tif
```

https://xkcd.com/1459/

https://datamanagement.hms.harvard.edu/collect/file-naming-conventions

The basics

- File names should contain only letters, numbers, underscores, and dashes
- A dash or underscore should be used instead of a space
- No special characters (& ' " ; : * ! # $, etc.)
- Maybe decide on a convention like
  - camel**C**ase
  - snake_case
  - **P**ascal**C**ase

Three principles for file names

1. Machine readable
2. Human readable
3. Plays well with default ordering

Steps to consider

1. Think about your files
2. Identify metadata
3. Abbreviate or encode metadata
4. Use versioning (incl. numbering, dates)
5. How will you search for your files?
6. Deliberately separate metadata elements
7. Write down your naming conventions

Naming conventions

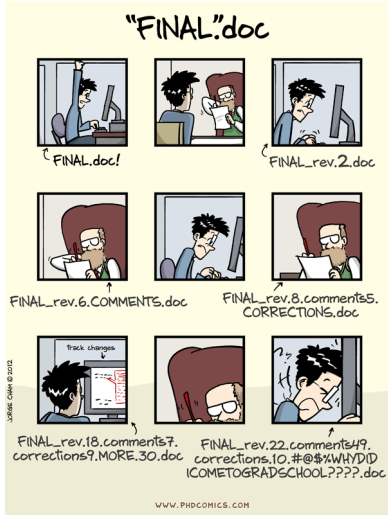
	Example	Documentation
Content-specific	`DATA_vp01_load_ses01.csv`	`DATA_[ID]_[cond]_[ses].csv`
Descriptive	`ANALYSIS_01_model-selection.R`	`ANALYSIS_[#]_[descrp].R`
Consistent	`ANALYSIS_02_plots.R`	`ANALYSIS_[#]_[descrp].R`
Leading date	`2022-09-29_exp1_vpall.txt`	`[yyyy-mm-dd]_[exp]_[type].txt`
Leading zero	`01_data-cleaning_study1.Rmd`	`[#]_[descrp]_[study].[R/Rmd]`

→ Think about your files
→ Identify metadata
→ Abbreviate or encode metadata
→ How will you search for your files?
→ Deliberately separate metadata elements
→ Write down your naming conventions

Exercise

- What would be a good and self-explanatory naming convention for my data files?
- Write down some metadata that could be useful here
- Come up with a naming convention standard for these files and write it down

Version control



https://phdcomics.com/comics/archive.php?comicid=1531

Version control

- Version control is a systematic approach to record changes made in a file, or set of files, over time
- File versioning can be as simple as using file naming conventions like suffixes `*_v1`, `*_v2`, `*_vn`

1. Create files – these may contain text, code or both
2. Work on these files, by changing, deleting or adding new content
3. Create a snapshot of the file status (also known as version) at this time
4. Document versions (e. g., in a README file)

https://the-turing-way.netlify.app/reproducible-research/vcs.html

6 Metadata

README files

- Provide a clear and concise description of all relevant details about data collection, processing, and analysis
- README files are created for a variety of reasons:
  - to document changes to files or file names within a folder
  - to explain file naming conventions, practices, etc. "in general" for future reference
  - to specifically accompany files/data being deposited in a repository
- Creating a README file at the beginning of your research process, and updating it consistently throughout your research, will help you to compile a final README file when your data is ready for deposit
- Find a template here: `https://cornell.app.box.com/v/ReadmeTemplate`

`https://datamanagement.hms.harvard.edu/collect/readme-files`

Metadata

Metadata

. . . is data about data.
. . . can be *descriptive*, *structural*, or *administrative*.

Contains information on origin and background of data like

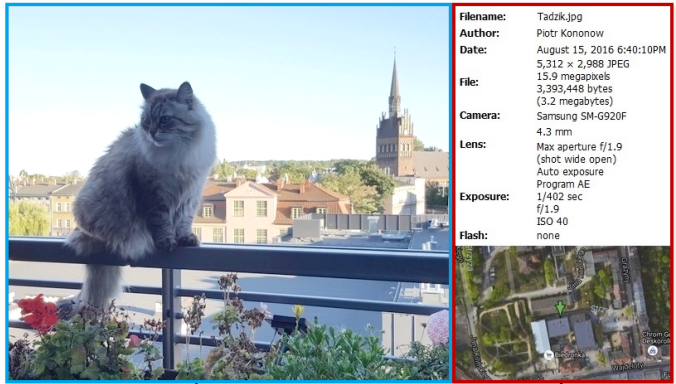
- Who, when, why, how, . . .
- Used resources
- Used abbreviations, units, names
- Licenses
- . . .

Data can be anything like

- Book content
- Pictures or audio files
- Website content or a blog post
- Journal paper
- Research data
- . . .

Metadata examples

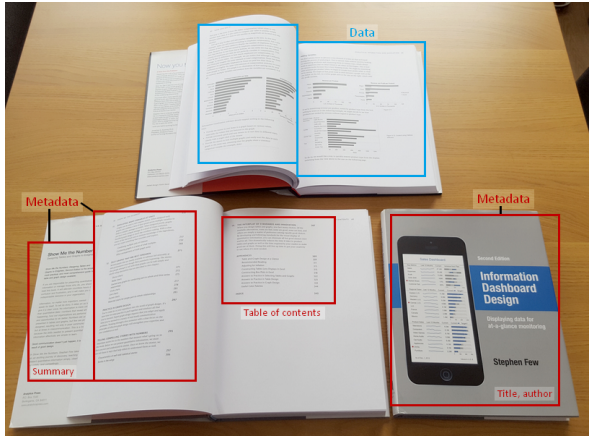
Photo



https://dataedo.com/kb/data-glossary/what-is-metadata

Metadata examples

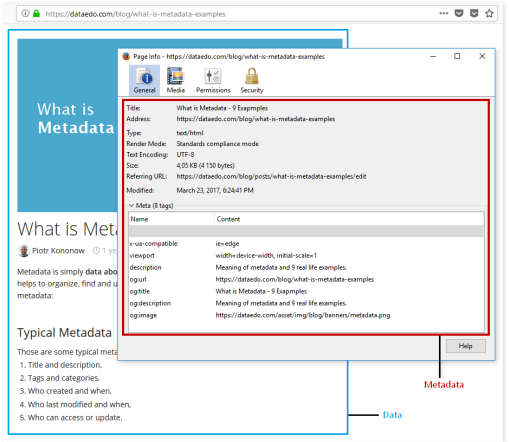
Book



https://dataedo.com/kb/data-glossary/what-is-metadata

Metadata examples

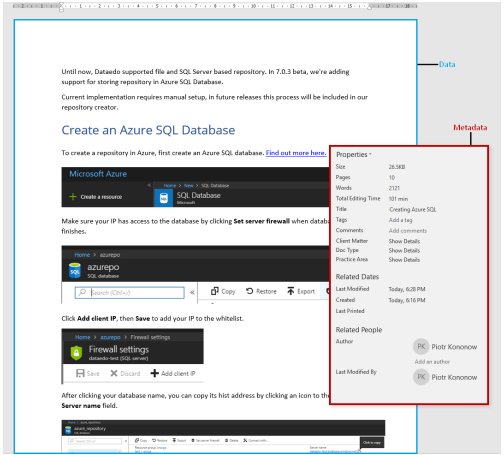
Webpage



https://dataedo.com/kb/data-glossary/what-is-metadata

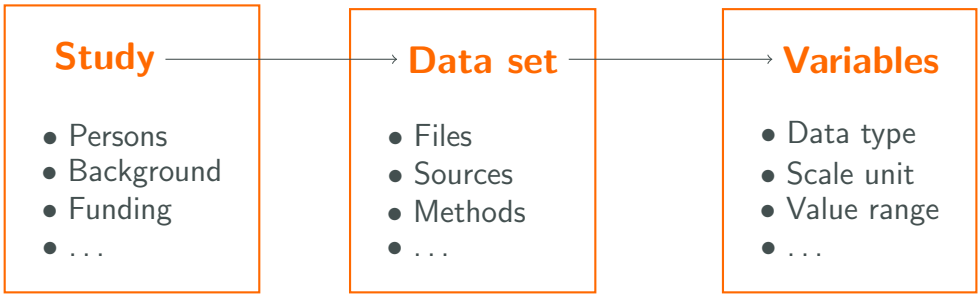
Metadata examples

WORD document



https://dataedo.com/kb/data-glossary/what-is-metadata

Metadata for research data



https://datamanagement.hms.harvard.edu/collect/readme-files

Metadata answers questions

- **Who** created the data?
- **Why** was the data created?
- **When** was the data created?
- **Where** is the data?
- **How** was the data created?
- **What** is the content of the data?

https://doi.org/10.5281/zenodo.7573695

7 Take away

Start small

- Start your 30 Days of Data Management Habits: https://nextcloud.iwm-tuebingen.de/s/A5HbJZmZ7W5sQjP
- Are you ten finger typing, yet? (If not, this is definitely something that will improve all of your workflows)
- Clean out the folders in your current project; rename the files
- Organize your literature folder
- Think about smart usage of cloud storage (there are many different options, there should be one that suits you – are you using the one that suits you best?)
- Next time you want to e-mail a document, think about a better way to share it
- Use R Markdown to write your next preregistration
- Read a book on R and data analysis
- Use Git for your next data analysis

What we didn't cover today

- Tidy data
- Codebooks (but see Jürgen's workshop!)
- Interactive reports
- Data loss prevention
- Cloud storage
- Version control
- Repositories
- Data management plans
- ...

Additional resources

- Resources from Kathawalla et al. (2021): `https://osf.io/w5mbp`
- Blogpost *A Guide to Open Science for People Who Are Already Too Busy*: `https://medium.com/@mullarkey.mike/a-guide-to-open-science-for-people-who-are-already-too-busy-e42f6ac3a1c7`
- Open Science Training Handbook: `https://open-science-training-handbook.gitbook.io/book/`
- The Turing Way: `https://the-turing-way.netlify.app/reproducible-research`
- R Markdown templates for preregistration: `https://github.com/crsh/prereg`
- The GIN-Tonic tool: `https://genr.eu/wp/towards-a-standardized-research-folder-structure/`

PS:

This week is International Love Data Week!

Remember: Documenting is like writing a love letter to your data ...

Love your data!

https://forschungsdaten.info/fdm-im-deutschsprachigen-raum/love-data-week/

References

Kathawalla, U.-K., Silverstein, P., & Syed, M. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology*, *7*(1).

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 1–7.

Wilbrandt, J. (2023). *Research Data Management Intro Series: Coffee Lectures & Espresso Shots.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7573695`