Workshop on Data Organisation

Justine Vandendorpe & Sophie Boße,
ZB MED - Information Centre for Life Sciences

Photo by ZB MED

What you will find in this slide deck

Slides with general aspects of data organisation

Slides with discipline-specific content

Engagement methods

Source: ZB MED 2024

Content

- Motivation
- File naming
- File versioning
- Folder structure
- Tidy data
- Further resources

Motivation

Ideas out loud

Why do we need to organise data in a structured way? (5 min).

Source: modified from Biernacka *et al.* 2020; photo by Volodymyr Hryshchenko on Unsplash

Structured approach

A structured approach will result in more **efficient work**.

- To ensure that what, how and why things were done remains **traceable**.
- To avoid **duplication** and **loss** of data.
- To make **naming conventions** known to everyone.
- To facilitate **collaboration**.
- To make it easier to **search** for and **find** data.
- To make it possible for **other researchers** to work with the data.
- To identify the **current state** without effort.
- To ensure **machine readability**.

Source: modified from Biernacka *et al.* 2020 by ZB MED 2024

5S methodology

1. **Sort:** delete unnecessary files
2. **Set in order:** develop and document naming conventions and folder structures
3. **Shine:**
   - Comply with conventions
   - Develop routines
4. **Standardize:**
   - Document rules and responsibilities
   - Develop best practices and Standard Operating Procedures (SOPs)
5. **Sustain:**
   - Regularly check whether rules are followed
   - Implement improvements if necessary

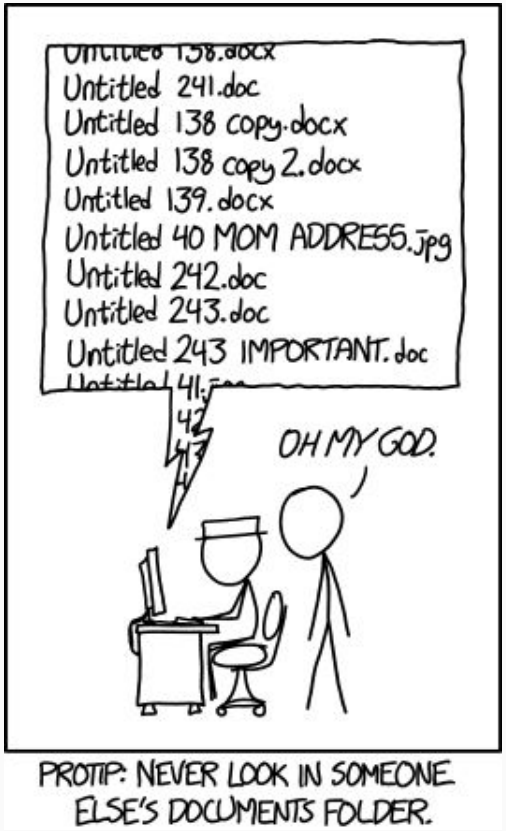
Source: modified from Assmann *et al.* 2022 by ZB MED 2023; icon from Kücklich 2020

File naming

Why do we need a naming convention?

"A file naming convention is a framework for naming your files in a way that describes what they contain and how they relate to other files." [Longwood Research Data Management]

- ➔ maximizes access to your records
- ➔ helps to stay organized
- ➔ helps to quickly identify your files
- ➔ helps others to navigate your files



Source: LMA RDMWG 2024a; comic from xkcd

Criteria for a good naming convention

Good filenames are

➔ human-readable

➔ machine-readable

➔ and they play well with default ordering

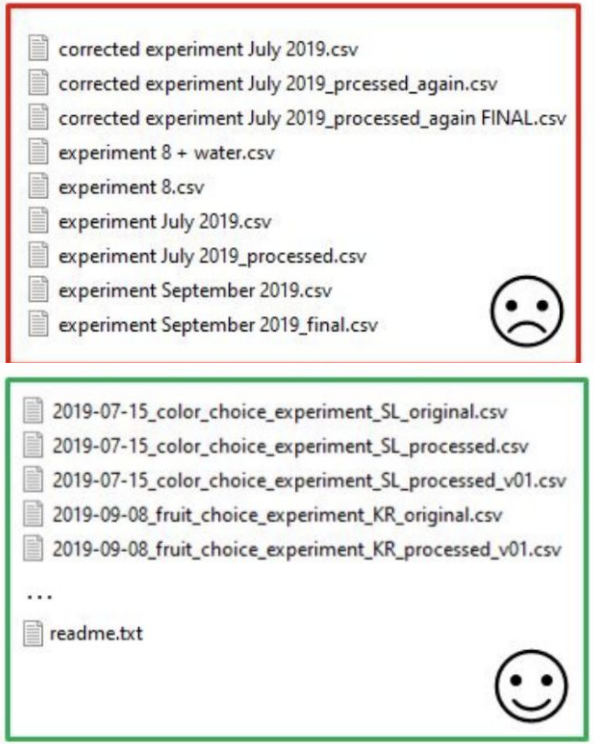
Source: Goldman 2020a, Bres et al. 2022; photo by Girl with red hat on Unsplash

Examples of file names (1/2)

Good structure: YYYY-MM-DD_JV_ProjectID_ExperimentID

Uniform naming

- 20160512_ClimateMeasurement1_original.jpg
- 20160522_ClimateMeasurement1_MHU_excerpt.jpg
- 20160523_ClimateMeasurement1_MHU_excerpt_edited_color.jpg



Source: modified from Biernacka *et al.* 2020 and Bobrov *et al.* 2021 by ZB MED 2024; Figure by Rehwald et al. 2022

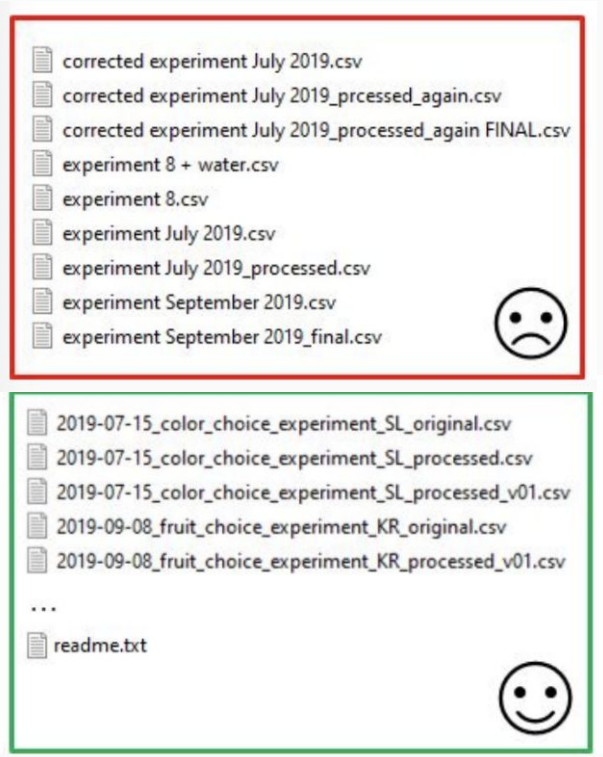
Examples of file names (2/2)

Good names

- 2016-01-04_ProjectA_Ex1Test1_SmithE_v1.0.xlsx
- 2000_USNM_379221_01.tiff
- USNM_379221_01.tiff

Bad names

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

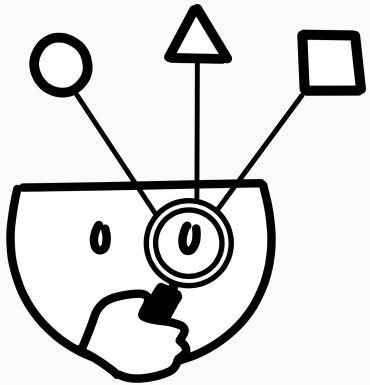


Source: modified from Bres *et al.* 2022 by ZB MED 2024; Figure by Rehwald et al. 2022

Exercise: naming conventions

Which of these examples follow a good naming convention? (5 min).

- 4hr3nofnf3w49389utz304.mp3
- Projekt001_Probe045_MassSpec_20200824.csv
- Workshop_RDM.pdf
- Probe1 3004 Britta+Olga new edited!corrected
- 2021-05-18_US_NoDirtyDishesDay

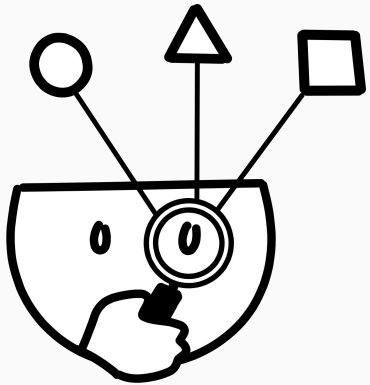


Source: modified from Bres *et al.* 2022 by ZB MED 2023; icon from Kücklich 2020

Exercise: naming conventions - Solution

Which of these examples follow a good naming convention?

- 4hr3nofnf3w49389utz304.mp3 → No
- Projekt001_Probe045_MassSpec_20200824.csv → Yes
- Workshop_RDM.pdf → No/Yes
- Probe1 3004 Britta+Olga new edited!corrected → No
- 2021-05-18_US_NoDirtyDishesDay → Yes



Source: modified from Bres *et al.* 2022 by ZB MED 2023; icon from Kücklich 2020

Creating a Naming convention in 7 Steps

1. Identify **what** group of **files** your naming convention will cover
    a. Check for established naming convention in your discipline or group!

Source: modified from Briney 2020, Goldman 2020a and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

Creating a Naming convention in 7 Steps

2. Identify **metadata** that are needed to **easily locate** a specific file
   a. pick 3 to 5 metadata to easily understand what is in each file
   b. Consider including a combination of the following information:
      i. Experiment conditions
      ii. Type of data
      iii. Researcher name/initials
      iv. Lab name/location
      v. Project or experiment name or acronym
      vi. Date or date range of experiment
      vii. Experiment number or sample ID
   c. Information included in Folder Names must not be repeated

Source: modified from Briney 2020, Goldman 2020a, Biernacka et al. 2022 and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

Creating a Naming convention in 7 Steps

3. **Abbreviate** or encode metadata if needed
    a. Limit file names to **≤ 32 characters**
        i. (32CharactersLooksExactlyLikeThis.txt)
    a. If any metadata has regular categories
        i. standardize them
        ii. replace them with 2 or 3-letter codes
    b. Don’t forget to document any codes!
    c. Do not use special characters e.g. {}[]<>()* % # ' ; " , : ? ! & @ $

Source: modified from Biernacka et al. 2022, Briney 2020, Goldman 2020a and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

Creating a Naming convention in 7 Steps

4. Select the order of the metadata so that **filenames play well with default sorting**
    a. Think about how you want to sort and search for files
    b. Put the most important metadata first
    c. Use default ordering: alphabetically, numerically, chronologically
    d. Use ISO 8601 formatted dates (YYYYMMDD or YYYY-MM-DD)
    e. Use leading zeros for sequential numbering systems
        i. 001,002, … 101,102 instead of 1, 2, … 101, 102

Source: modified from Briney 2020, Goldman 2020a and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

Creating a Naming convention in 7 Steps

5. Decide how to **separate** each metadata element
    a. **Avoid white spaces** for making the names machine readable
    b. Use dashes (-), underscores (_) and/or CapitalLetters to make filenames machine and human readable
    c. Use periods only before file extensions

Source: modified from Briney 2020, Goldman 2020a and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

Creating a Naming convention in 7 Steps

6. Use **versioning** if you need to track different versions of each file
    a. Use a Version number (e.g. “v01”)
    b. Or a Version Date
    c. Or a status of your workflow (e.g “raw”, “processed”, “composite”)
    d. Put the versioning at the end of the file name

Source: modified from Briney 2020, Goldman 2020a and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

Creating a Naming convention in 7 Steps

7. **Document** your Naming convention!
    a. Create an SOP or a README.txt
    b. Keep it with your files
    c. Make the documentation **available** to all research group members
    d. Stay **consistent**

Source: modified from Briney 2020, Goldman 2020a and LMA RDMWG 2024a by ZB MED 2024; photo by Andreas Brücker on Unsplash

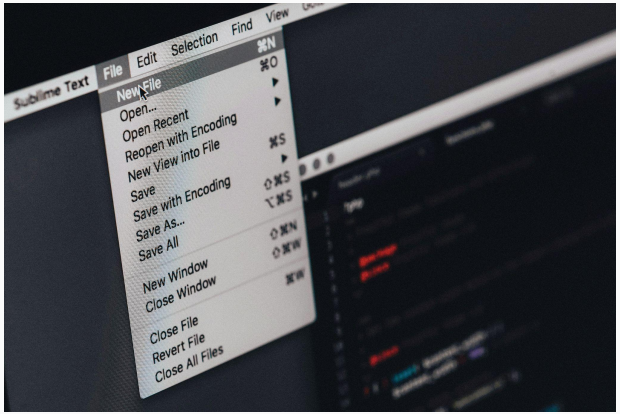
Creating a Naming convention in 7 Steps

```
File naming convention:
strain-treatment-specimenNum_age_version_imageNumber.fileFormat

Attributes:
       strain = indicates whether the mouse was C57BL/6 or BALB/c
    treatment = indicates whether the mouse was ovariectomized or non-surgical control
  specimenNum = a 2-digit number assigned to each mouse in the treatment groups (between 01-06)
          age = age of the mouse, in weeks, at the time of the microCT scan
      version = indicating whether image is raw data, reconstructed or processed data file
  imageNumber = 4-digit image slice number in the microCT z-stack (e.g. 00xx)
   fileFormat = 8-bit bitmap (.bmp) image files as generated by the microscope

Acronyms and Codes:
          strain = {C57BL/6: "BL6", BALB/c: "BaC"}
       treatment = {ovariectomized: "OVX", non-operated control: "CTL"}
     specimenNum = {"01", "02", "03", "04", "05" ,"06"}
     age (weeks) = {"w14", "w16", "w18", "w20", "w22", "w24"}
         version = {raw: "raw", reconstructed: "rec", processed: "proc"}

Examples:
The raw image file of the 3rd C57BL/6 mouse from the ovariectomized group, microCT scanned at 18-weeks-old
        - BL6-OVX-03_w16_raw_0001.bmp
The raw image file of the 1st BALB/c mouse from the non-surgical control group, microCT scanned at 22-weeks-old
        - BaC-CTL-03_w22_raw_0001.bmp
```

Source: [Adelaide microscopy]

Exercise: naming conventions

- Please design a naming convention for your files and give some examples [Biernacka et al. 2022].
- You can use a worksheet and/or a checklist that might help you



Source: modified from Biernacka et al. 2022 by ZB MED 2024; photo by Ilya Pavlov on Unsplash

Tools for simultaneous renaming of files

Multiple

- Adobe Bridge
- jExifToolGUI

Linux

- Gnome Commander
- GPRename

Mac

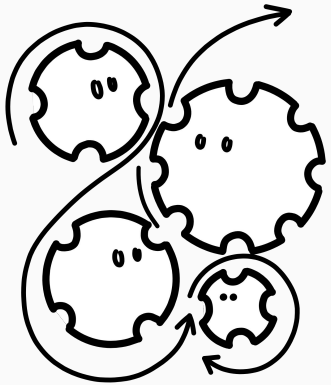
- ExifRenamer
- NameChanger
- Renamer 6

Unix

- mv command

Windows

- Advanced Renamer
- Altap Salamander
- Ant Renamer
- Bulk Rename Utility
- ExifToolGUI
- Rename-IT
- Total Commander
- WildRename



Source: modified from Biernacka *et al.* 2022 and Malinowski 2020 by ZB MED 2023; icon from Kücklich 2020

File versioning

To keep in mind

- **Decide** with project partners
- **Write down** how a version change is to be defined
- **Document** version changes

Source: modified from Bres *et al.* 2022 by ZB MED 2023; photo by Jason Goodman on Unsplash

Purposes

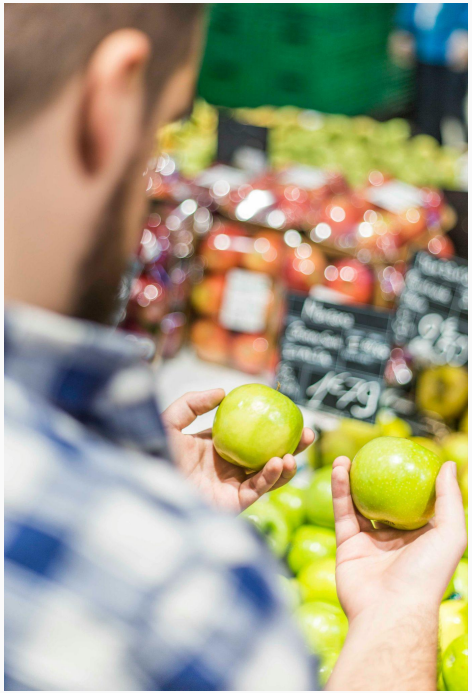
Versions and their history help:

- To **track** and **trace** your steps.
- To easily **go back** one step.
- To support **debugging**.
- To **create** new versions.

Source: modified from Biernacka *et al.* 2020 by ZB MED 2024; photo by eberhard 🖐 grossgasteiger on Unsplash

Options

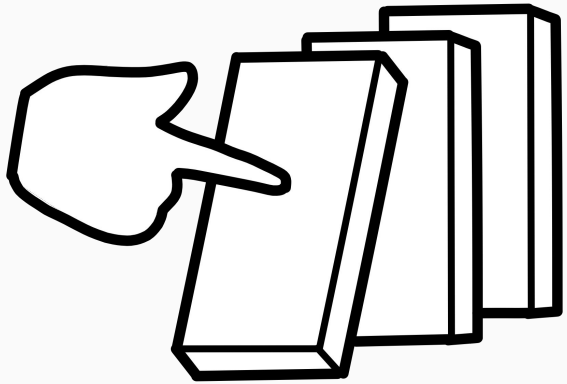
- In **file names**
- Within **data**
- In **text files**
- With a **Version Control System** (VCS)
- **Versioning** and **change tracking** are available for collaborative documents and storage locations.



Source: modified from Biernacka *et al.* 2020 and Bres *et al.* 2022 by ZB MED 2024; photo by Raquel Martínez on Unsplash

Manual file versioning

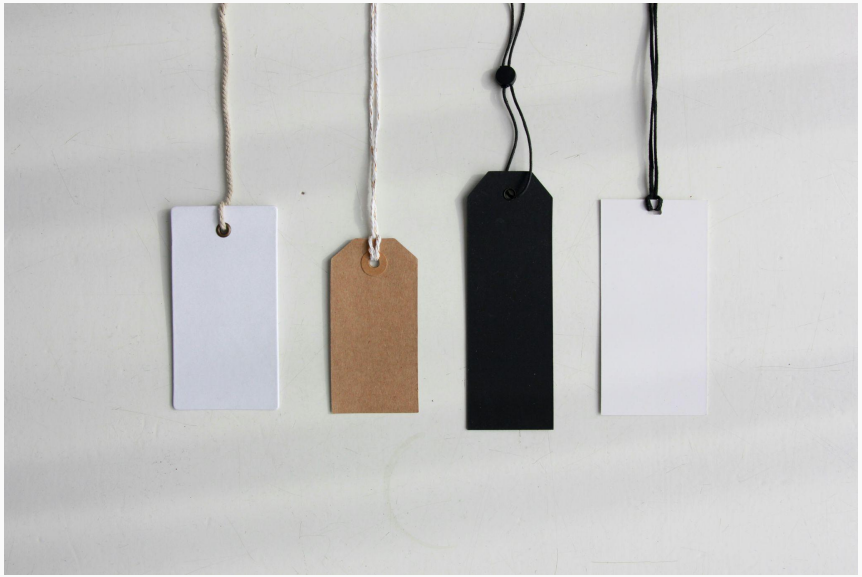
- Use a version control **table** (e.g. doc from the University of Sydney)
- Define **responsibilities** for completion of files
- Use semantic versioning: **MAJOR.MINOR.PATCH**
  - Ex1Test1_SmithE_v**1.0.0**.xlsx
  - Ex1Test1_SmithE_v**1.2.5**.xlsx
  - Ex1Test1_SmithE_v**2.1.1**.xlsx
- Save **milestone versions**
- Store **obsolete versions** separately after backup



Source: modified from Bobrov *et al.* 2021 and Bres *et al.* 2022 by ZB MED 2023; icon from Kücklich 2020

Examples of file labelling with version control

- [document name][version number]
- Doe_interview_July2010_V1
- Lipid_analysis_rate_V2
- 2017_01_28_MR_CS3_V6_03

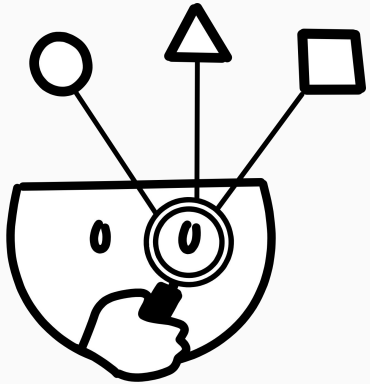


Source: modified from Biernacka *et al.* 2020 by ZB MED 2024; photo by Angèle Kamp on Unsplash

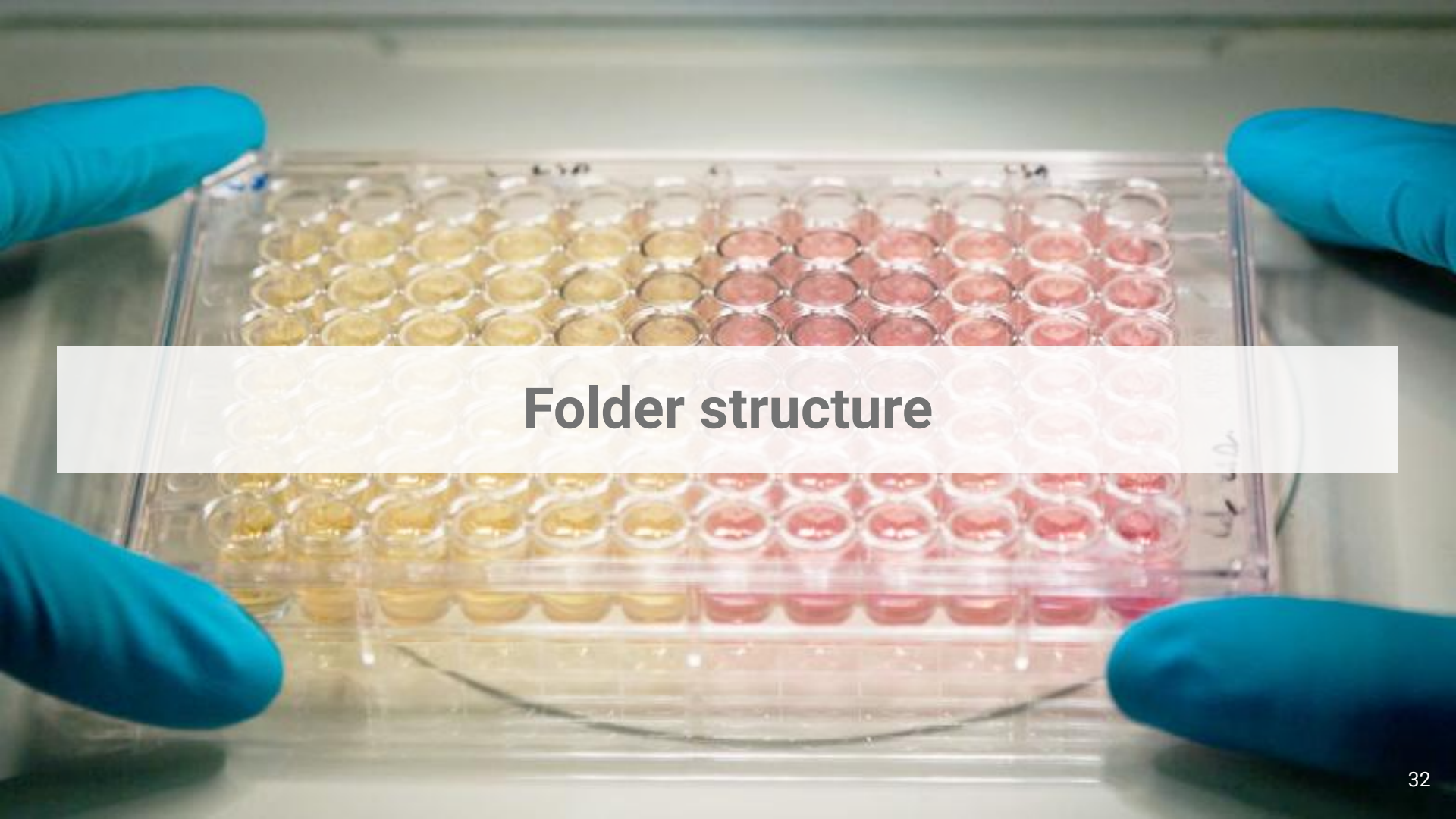
Exercise: file versioning

Can you improve this filename?

Final FINAL last version.docx



Source: modified from Bres *et al.* 2022 by ZB MED 2023; icon from Kücklich 2020



Folder structure

Why organizing files and folders systematically

- **Save time and nerves** when searching for files
- **Improve workflow** during the project
- **Collaboration** made easy

Source: modified from Goldberg 2020b by ZB MED 2024; photo by Vladalek on Unsplash

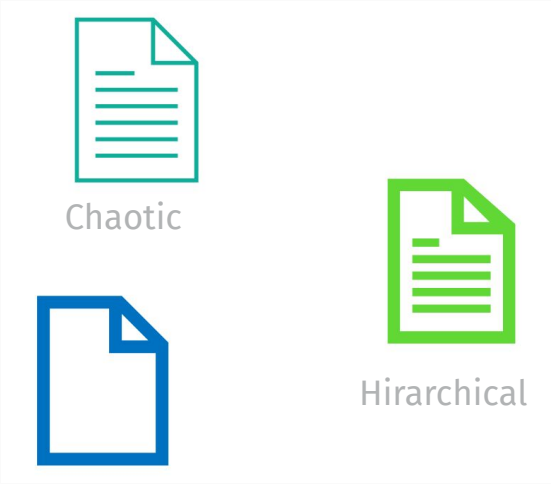
Why organizing files and folders systematically

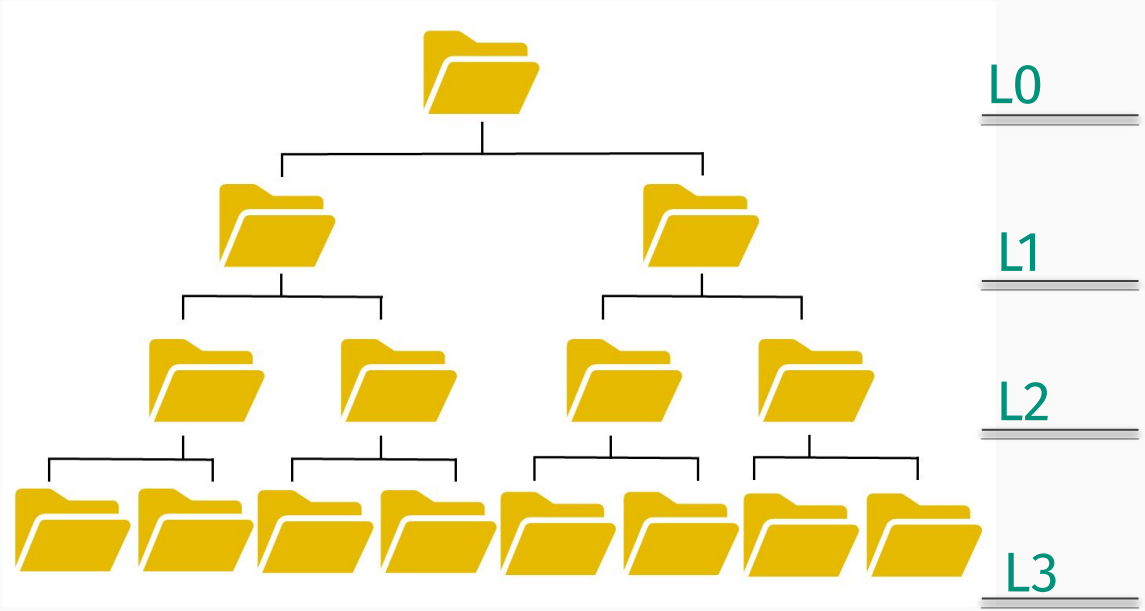
- **Save time and nerves** when searching for files
- **Improve workflow** during the project
- **Collaboration** made easy

Imagine someone (maybe your future self) looks at your files and immediately understands in detail what you did and why.

Source: modified from Goldberg 2020b by ZB MED 2024; photo by Jason Leung on Unsplash

Recommendations for folder structure

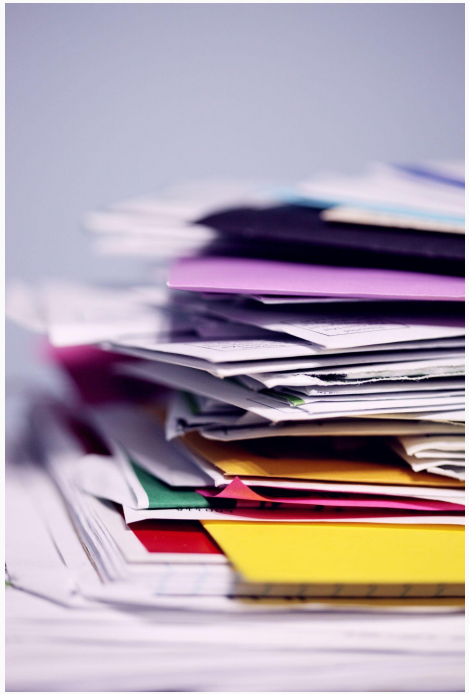




Source: modified from Boße et al. 2024 by ZB MED 2024

Recommendations for folder structure

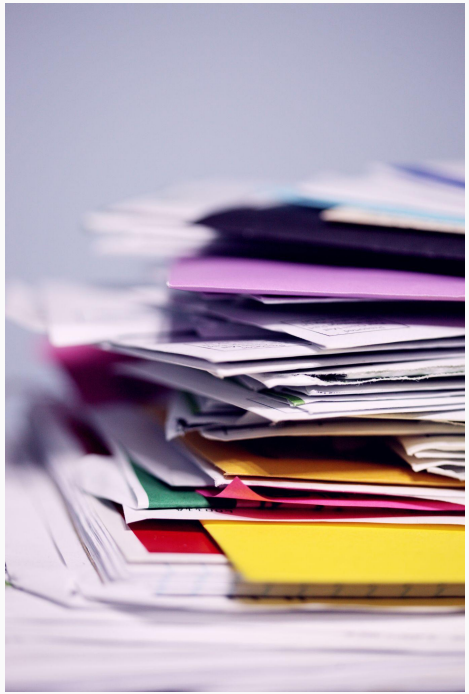
- Choose a folder structure that is:
  - **Clear**
  - **Comprehensive**
  - **Efficient**
  - **Hierarchical**
- Avoid **overlapping categories**
  - no copies of files in different folders!
- Apply naming conventions
- Maximum **4 levels**
- Maximum **10 elements** per folder



Source: modified from Bobrov *et al.* 2021 and Bres *et al.* 2022 by ZB MED 2023; photo by Alexander Grey on Unsplash

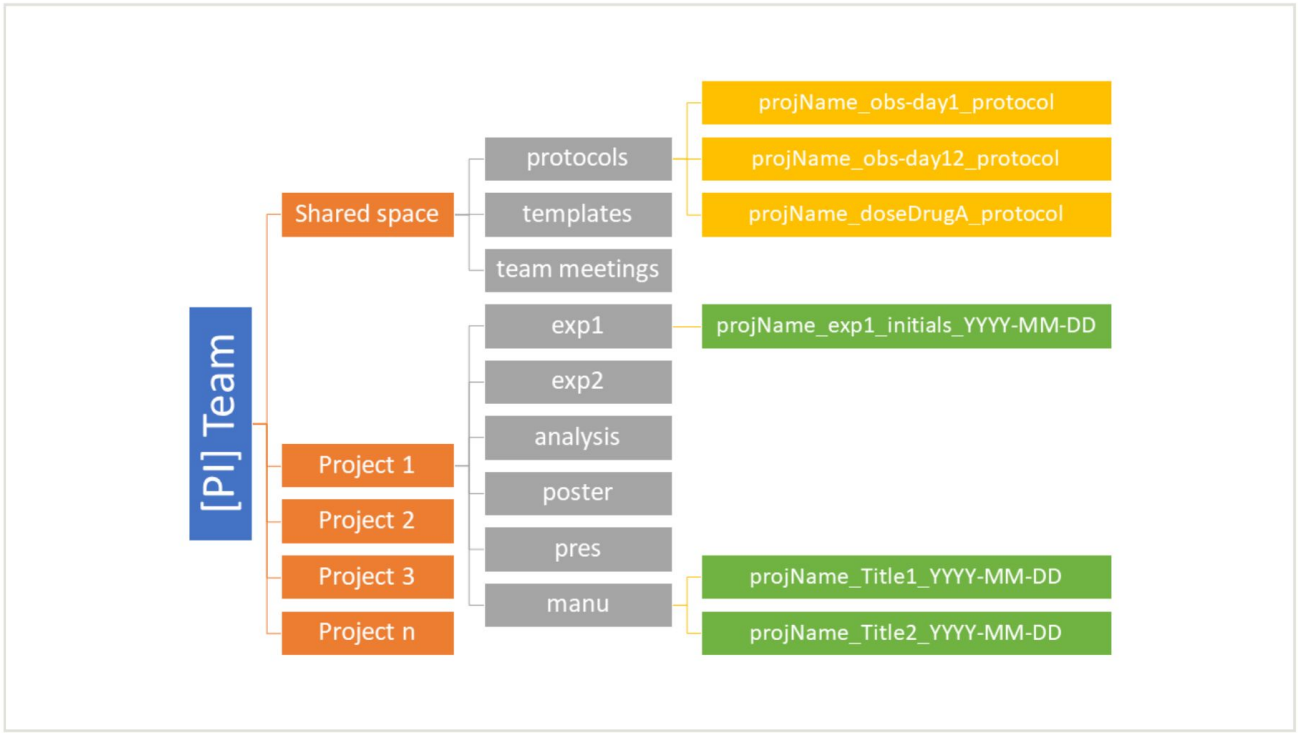
Recommendations for folder structure

- there is **no one fits all solution**
- Distinguish between….
  - **work** vs. **private** material
  - **own** vs. **others'** work (papers vs. literature)
  - **research** vs. **administrative** content
  - **raw** vs. **processed** vs. **final** data
  - **experiment** vs. **analysis**
- Avoid generic folders
- Avoid researcher-specific folders - folders are **about the content**, not the authors!



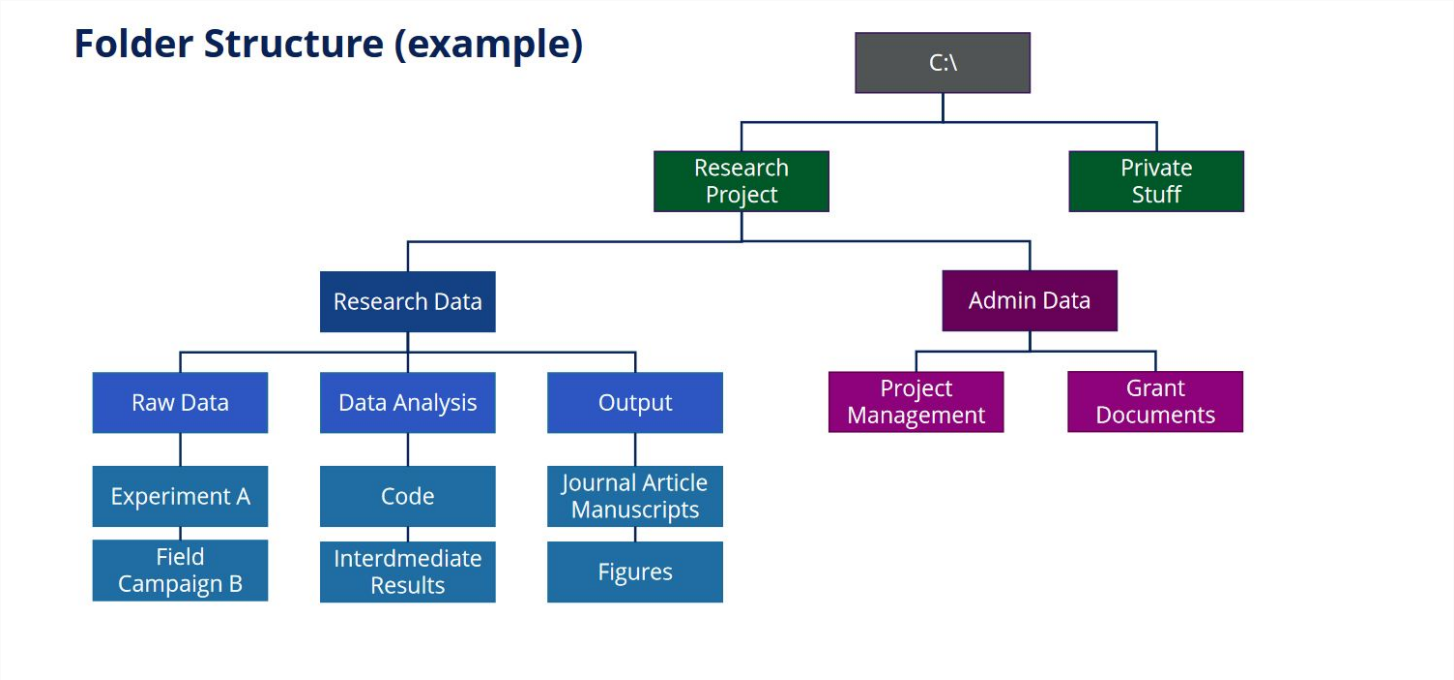
Source: modified from Bres et al.Schmid 2021 by ZB MED 2024; photo by Alexander Grey on Unsplash

Example of folder structure (1/4)



Source: Briney et al. 2020

Example of folder structure (2/4)



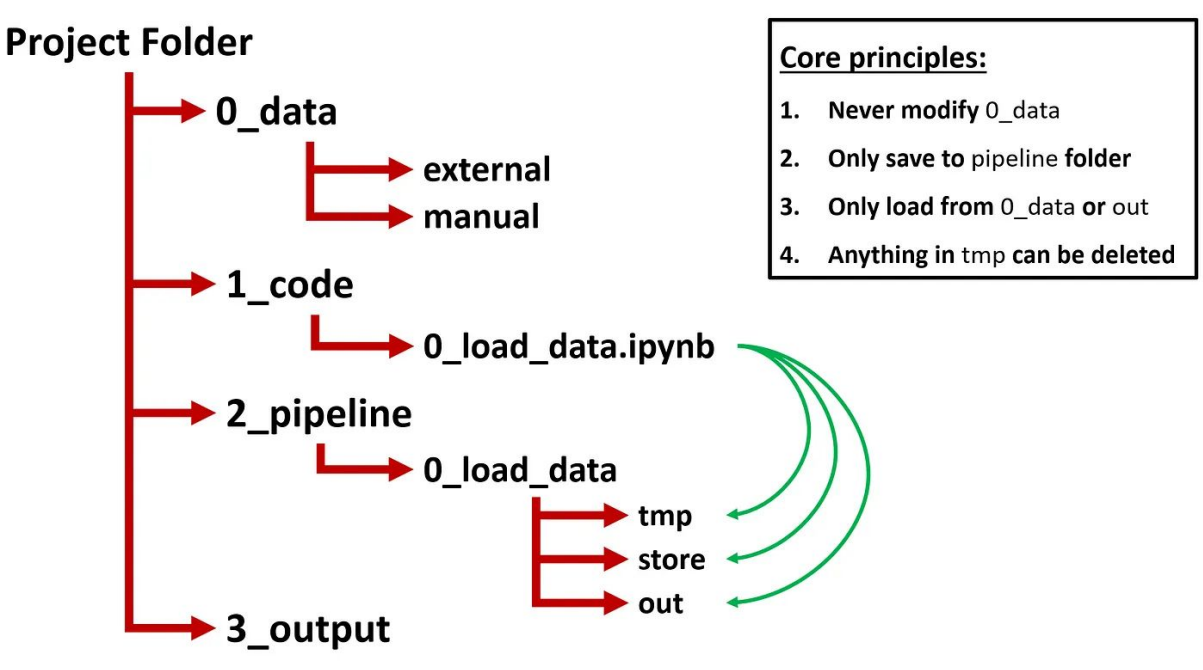
Source: Briney et al. 2020

Example of folder structure (3/4)

- Project
  - Data
    - Raw_data
    - Processed_data
    - Documentation
  - Code
    - Src
    - Output
      - Plots
    - Documentation
  - Protocols
- Manuscripts
- Conference_reports
- Administrative_information

Source: modified from Bobrov *et al.* 2021 by ZB MED 2023; photo by Agnivesh Jayadeep on Unsplash

Example of folder structure (4/4)



Source: de Kok 2018

Exercise

Think about one thing you could improve in your current folder structure!

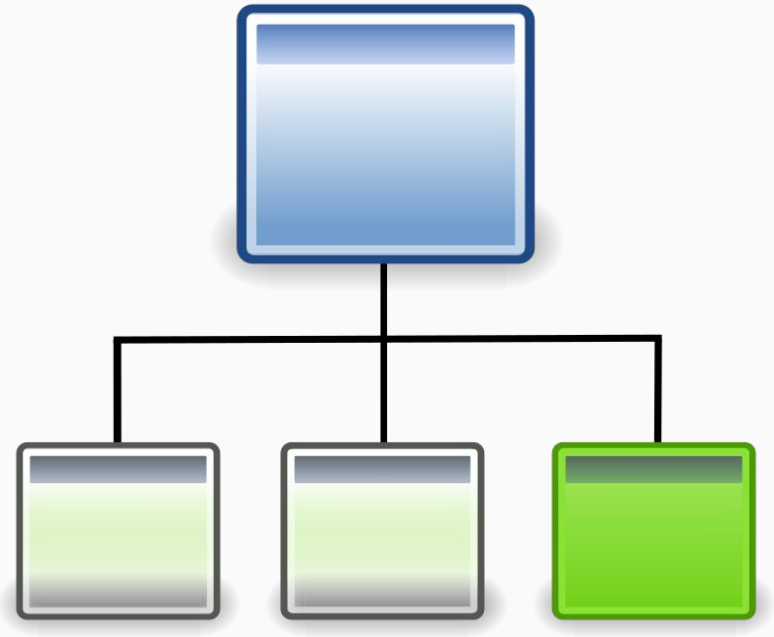
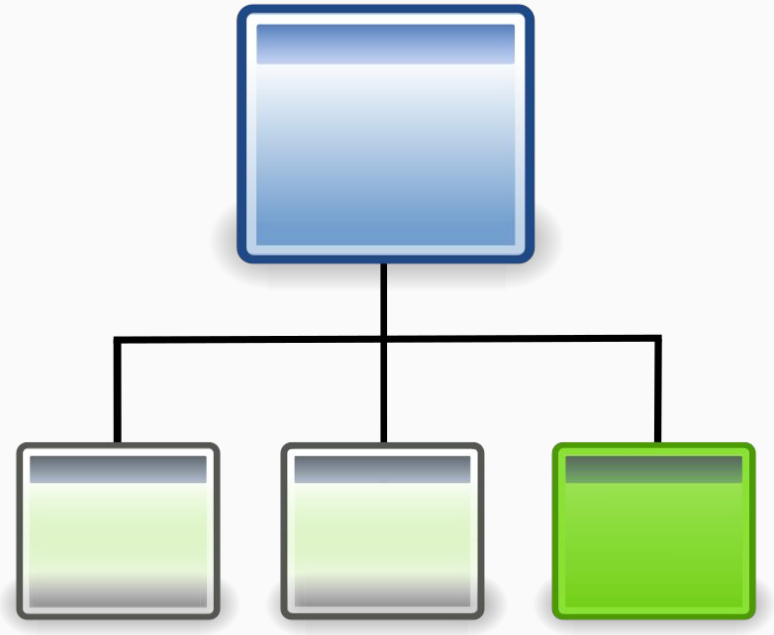


photo from RRZEicons 2009

Exercise

Create a structure for your storage in the form of a directory tree (10 min).



Source: modified from Biernacka *et al.* 2020 by ZB MED 2024; photo from RRZEicons 2009

Standardized folder structure

Benefits

- Facilitate **research collaborations** and **lab management**
- Promoting **data management**
- **Easy to use** once set up
- **Time saving** when browsing for files
- Good at representing the **nested structure** of information

Drawbacks

- Too detailed folder structure seems to be rather inconvenient for **saving** files
- **Full impact** reached only if the whole research group uses it
  - different “understanding” of data
- High **cost of transition** for ongoing project
- Hard to find a **good balance** between **breadth** and **depth**
- Categories are inherently discrete and can result in **forced separation**

Source: modified from Colomb *et al.* 2021, Kleppert 2024 and US Merced Library by ZB MED 2023

Tag-Based File Structure

- use tags to organize your files
- decide on a **set of tags** to use in the beginning of the project
- find a **balance** between
  - too many tags (makes it impossible to find your file)
  - too less tags (your file is not unique and cannot be found)
- be **consistent**
- **Programs** for tagging files:
  - Tagsistant (Unix, free)
  - Tabbles (Windows, free w/ limited features)
  - TMSU (Unix, Windows, free)

Source: modified US Merced Library and Shanshan Ma 2010 by ZB MED 2024, photo by Jon Tyron on Unsplash

Tag-Based File Structure

Benefits compared to a hierarchical Folder structure

- easier to set up
- easier to combine different filesets (for collaborative work)
- one file can be assigned to multiple categories

Drawbacks compared to a hierarchical Folder structure

- doesn't represent the structure of information
- risk of inconsistency -> loss of files
- not for all file formats/organisation tools available
    - risk of loss of information when moving to another system

Hybrid approach: You can also add tags to your hierarchical folder structure

Source: modified US Merced Library and Shanshan Ma 2010 by ZB MED 2024, photo by Jon Tyron on Unsplash

Recommendations for establishing a system

- Invest **time** planning out folder structure
- Establish the system **as a group**
- Provide a method for **easy adoption**
  - provide a **template**
  - **document** the system in a README file

```
project/
  code/                code needed to go from input files to final results
  data/                raw and primary data (never edit!)
     raw_external/
     raw_internal/
     meta/
  doc/                 documentation of the study
  intermediate/        output files from intermediate analysis steps
  logs/                logs from the different analysis steps
  notebooks/           notebooks that document your day-to-day work
  results/             output from workflows and analyses
     figures/
     reports/
     tables/
  scratch/             temporary files that can safely be deleted or lost
  README.txt           file and folder description
```

Source: modified from Bobrov *et al.* 2021 and Bres *et al.* 2022 by ZB MED 2023; Figure from Mičetić et al., elixir

Further Resources

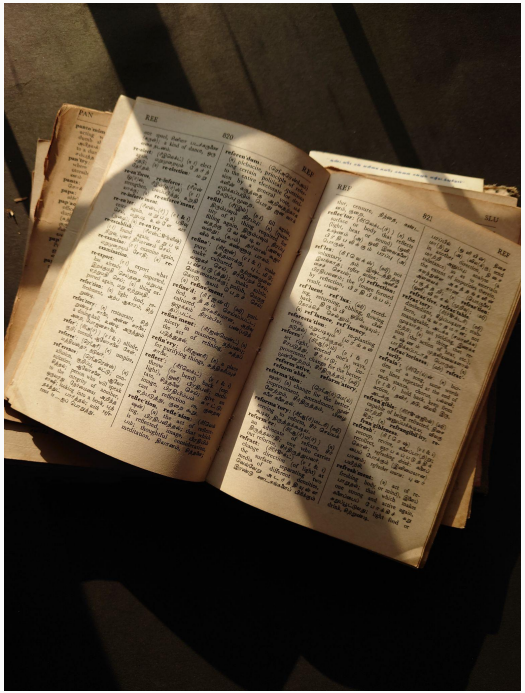
- Checklist organize your files
- downloadable folder structures to reuse:
  - Organised Folder Structure for Research Projects
  - Folder Structure Generator for Research Projects
  - GIN Tonic - Research Folder Structure Standard (not just, but also for working with Git Serves - GIN, GitHub, GitLab)
- Guide on tagging and finding files

Source: modified from Colomb *et al.* 2021, Kleppert 2024 and US Merced Library by ZB MED 2023

Tidy data

Tidy data

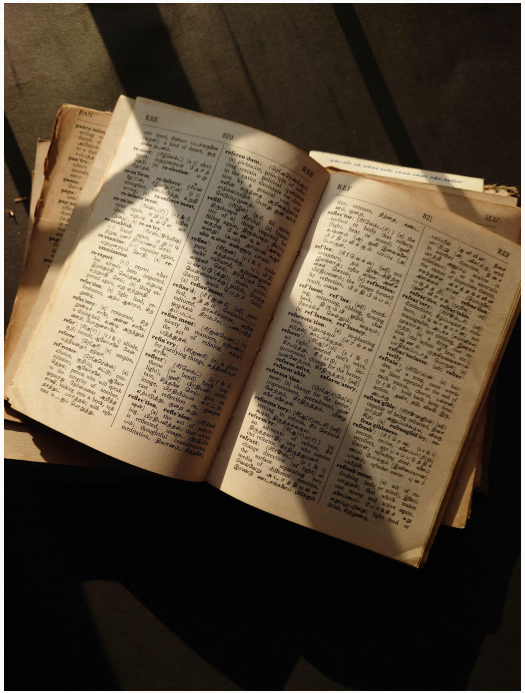
Tidy data = “data that is well designed for working with computers” [The Carpentries].



Source: ZB MED 2024; photo by Antony Hyson S on Unsplash

Spreadsheet programs

Spreadsheet programs = “very useful graphical interfaces for designing data tables and handling very basic data quality control functions” [The Carpentries].



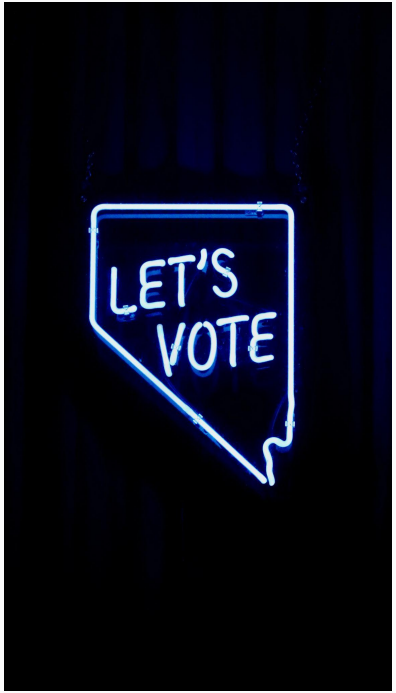
Using spreadsheet programs for data organisation

- Data entry
- Organising data
- Subsetting and sorting data
- Statistics
- Plotting

Source: modified from The Carpentries by ZB MED 2024; ZB MED 2023; photo by Antony Hyson S on Unsplash

Instant poll

- Have you ever used spreadsheets in your work?
- What kind of work do you do in spreadsheets?
- What do you think spreadsheets are good for?



Source: modified from The Carpentries by ZB MED 2024; photo by Manny Becerra on Unsplash

Ideas out loud

What are some things you've accidentally done in a spreadsheet, or been frustrated by not being able to do easily? (5 min).

Source: modified from The Carpentries by ZB MED 2024; photo by Volodymyr Hryshchenko on Unsplash

Problems with spreadsheets

- Create **tables** for reports.
- **Replicate** your or someone else's steps.
- Accidentally apply a slightly different **formula** to several adjacent cells.

Source: modified from The Carpentries by ZB MED 2024; photo by eberhard 🖐 grossgasteiger on Unsplash

Improving messy data

Download a messy version of some of the Portal Project data. This includes information about the site, date, species identification, weight and sampling area (within the site) for some small mammals.

Think about what could be improved about this data and write down answers to the following questions (15 min):

- Describe five things about this data that are not tidy and how you could fix each of these problems.
- Could this data be easily imported into a programming language or database in its current form?
- Do you think it's a good idea to enter the data as it is and clean it up later, or to have a good data structure for analysis at the time of data entry? Why?

Formatting data tables in spreadsheets

- Use one **column** for one variable.
- Use one **row** for one observation.
- Use one **cell** for one value.
- Never change your **raw data**. Always make a copy before making any changes.
- Keep all of the **steps** you take to clean your data in a plain text file.

RDM training					
Date	Length (hours)	PGR	PDRA	other	Delivered by
4 Feb	1.5				GQ
7/8 Feb					GQ
20 Feb					GQ & DF
03/03/17	2	15	3	0	DF
04/03/17	2	30	0	0	DF
08/04/17	2	30	0	1	DF
26/05/17	2	27	0	0	DF
2 June?	2	24	2	0	DF
3 June?	1.5	12	7	4	DF

Source: modified from The Carpentries by ZB MED 2024

Exercise: one cell one value

Structure this table better (5 min).

Mass
26g
0.2kg

Source: modified from The Carpentries by ZB MED 2024

Exercise: one cell one value - Solution

Structure this table better (5 min).

Mass	Unit
26	g
0.2	kg

Source: modified from The Carpentries by ZB MED 2024

Formatting problems

- Don't use **multiple tables** on the same sheet.
- Don't use **multiple tabs** on the same file.
- Fill in **zero** when you mean zero.
- Use an appropriate **null value** to record missing data.
- Don't use **formatting** to convey information or to make the spreadsheet look pretty.
- Don't put **units** or **comments** in cells.
- Don't put more than **one value** in a cell.
- Be careful with **column names** (i.e. spaces, numbers and special characters).
- Avoid including **special characters** (e.g. {}[]<>()* % # ' ; " , : ? ! & @ $ ~) in your data file.
- Put **metadata** in a separate file.

Source: modified from The Carpentries and Biernacka *et al.* 2020 by ZB MED 2024

Appropriate null values

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

Source: table from White *et al.* 2013

Further resources

Further resources

5S methodology

- The 5S Methodology in Research Data Management
- 5S Data: Setz dich auf deine 5 Buchstaben und organisiere deine Daten! (Coffee Lecture)

File naming

- Briney 2020 (convention worksheet)
- Briney *et al.* 2020 (Table 1. File naming examples)
- ELIXIR Belgium

File versioning

- Biernacka *et al.* 2020 (Checklist: Versioning p. 66)

Folder structure

- Colomb *et al.* 2020 (template for research repositories)
- de Plaa 2021 (simple Open Data template)
- MIT Libraries
- MIT Libraries Data Management Services 2018

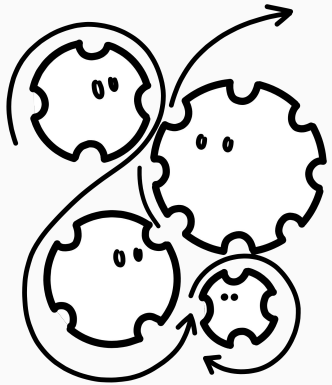
Data Organization in Spreadsheets

- Broman & Woo 2017
- Library Carpentry (tidy data)
- Perkel 2022

Source: ZB MED 2023

Discipline-specific tools

- Data Curation Tool (FAIR4Health)
- G-Node Infrastructure (GIN) = Modern Research Data Management for Neuroscience



Source: ZB MED 2023; icon from Kücklich 2020

Thank you!

For further information we are at your disposal

ZB MED – Information Centre for Life Sciences
Gleueler Straße 60
50931 Köln

forschungsdaten@zbmed.de
www.zbmed.de

This slide deck is based on the lesson plan on data organisation, available at DOI: https://doi.org/10.4126/FRL01-006484175