Storing Data and Code

Robert Oostenveld
Cyril Pernet
Melanie Ganz

Life cycle

Planning

Practicalities: Work-in-progress and regular backups

Archiving and sharing/publishing

Data Management Plan (DMP)

Data Management Paragraph != DMP

The "paragraph" is a short description in a project (funding) application about data management. It describes for example that the data will be secure, sensitive, FAIR, shared, etc.

The "paragraph" is only written once.

The "plan" is the detailed version, which you update as you go along in the project.

Data Management Plan (DMP)

Compares to a packing list and describes:

- What data will be collected/handled
- Whether it is sensitive
- Who will handle it
- Where it will be stored
- Which formats
- Etc.

Sometimes a DMP is a *formal* requirement of the institute or funding agency.

An *informal* DMP is a good idea anyway.

Storing Data and Code

- Storage systems
- Backup
- Folders and organizing
- File naming
- File formats
- Archiving
- Tools

Not included here: FAIR, sharing, licenses, metadata, provenance, versioning, ...

Storage systems

Exercise - *where* do you store your files

That is, on (or in) which systems

— 5 minutes —

xxxxx

Teacher notes: use some platform for students to share (preferably allow anonymous input for those not feeling comfortable with that - also nothing is mandatory here)

Exercise - *where* do you store your files

That is, on (or in) which systems

— 5 minutes —

Storage systems - file organization

Different requirements

- Small files can stay where they are, e.g., email attachments
- Collection of many small files require an organization
- Large files don’t fit on a laptop SSD
- Files might contain sensitive (personal) information*
- Requirements for password-protected storage systems
- Requirements for sharing (with or without access control)
- Requirements for large computations

Storage systems - file organization

Multiple layers of complexity

- What if you have a 2nd office, apartment or house
- What if you have a 2nd (and 3rd, …) storage system that you must use

Example from the Donders Centre

- Shared lab computers -> unsafe
- Personal laptop (and the occasional USB attached SSD) -> my whole life
- Network drive for work-in progress, linked to compute cluster -> parts of my life
- Donders Repository for long term storage -> well-defined chunks

Backups …

Teacher notes: use some platform for students to share (preferably allow anonymous input for those not feeling comfortable with that - also nothing is mandatory here)

How do you make backups of your valuable data, documents and files?

xxxxx

What is the difference between a **backed-up** and an **archived** copy of a file?

Backups …

How do you make backups of your valuable data, documents and files?

What is the difference between a **backed-up** and an **archived** copy of a file?

Folders and organizing

How do we store and organize our photos

… assuming you are old enough to know what analog photos are

How do we store and organize our photos

- Holiday photos (per year, per destination)
- Renovation of our first house
- Wedding photos
- Our child’s first birthday

Or do you rely on GPS coordinates and dates (automatic metadata) and BigTech (Apple or Google) to make them findable?

Organize yourself

You are doing research right now, you have already started with different projects and folders.

Most frequently used principle is **one project = one folder**.

There are of course other ways, but this is simple and allows to find things quickly.

Organize yourself

@organizedbyaly on Instagram

Organize by category
- Clothes, Food, Books

Organize within category
- On type, color, size

For files, organize on type
- data files go together
- documents go together
- presentations go together

Temporal sequence and flow in your project

Temporal sequence in your project

- Study planning
- Experimental design
- Acquisition of raw data
- Analysis scripts and results
- Posters and a paper

Think of the optimal granularity of "projects".

One top-level folder for one study or paper,
not for your whole PhD project.

Organize yourself

How many (sub)folders do you use and for what purpose?

What additional files are required (beside the code to analyse data) and for what?

xxxxx

Work in teams, you have 5 minutes

Teacher notes: use some platform for students to share (preferably allow anonymous input for those not feeling comfortable with that - also nothing is mandatory here)

Organize yourself

How many (sub)folders do you use and for what purpose?

What additional files are required (beside the code to analyse data) and for what?

Work in teams, you have 5 minutes

How I (wish I) organize myself

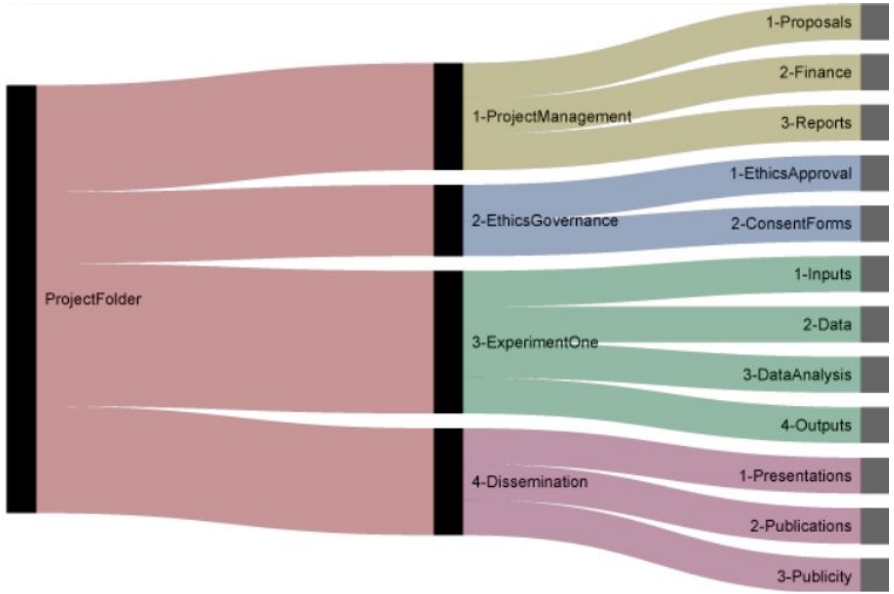
One project = one folder using a well-defined project ID + meaningful name

At the folder root: README.md file with full project title, PI, collaborators, funder, and details about what to expect in the folder.

- Project admin, ethics and governance folder
- Experimental design + material
- Raw neuroimaging data in BIDS folder (80% of your work)
- Code + requirements.txt
- Results
- Figures folder (for your talks & articles) –> share and get a DOI
- Dissemination or manuscript folder (outputs from that project)

Folders and organizing

Ask Enrico: Project management == Data management



https://eglerean.wordpress.com/2017/05/24/project-management-data-management/

Folders at DCCN

```
DCCN_NETWORK_DRIVE
├── Home_directories
│   ├── PI_Group
│   │   ├── Researcher1
│   │   └── Researcher2
│   ...
├── Project_directories
│   ├── 3011231.02
│   ├── 3055060.01
│   ├── xxxxyyy.zz
│   ...
```

```
xxxx    = PI group
xxxxyyy = budget number
zz      = sequential study number
```

Folders at BRC

```
Study_name
├── Analysis
│   ├── Excel files
│   ├── MATLAB files
│   ├── SPSS files
│   ├── Video coding files
│   └── other
├── DESCRIPTION_Experiment_name.docx
├── Data
│   ├── Raw data
│   │   ├── Behavioural
│   │   └── Neural
│   └── Video recordings
├── Experiment Info
│   ├── Hypothesis & Study Design
│   ├── Participant Info
│   ├── Program
│   ├── Project Proposal
│   ├── Stimuli
│   └── Testing Protocol
├── Final Results
│   ├── Final Presentation
│   └── Thesis
└── Literature
```

Folders at the level of RU university

```
RU_NETWORK_DRIVE
├── Department_name
│   ├── Common_files
│   ├── Researcher1_name
│   ├── Researcher2_name
│   │   ├── Study1_name
│   │   └── Study2_name
```

Different levels of organization

- Organization of my documents, travel itineraries, etc. -> per year
- Organization of my reusable source code -> per language
- Organization of other stuff -> domain specific standards
- ...
- Organization of the neuroimaging data in a project -> BIDS

Brain Imaging Data Structure (BIDS)

https://bids-standard.org/

Before: One directory with 1000s of DICOM files with cryptical names

After: top-level data folder with
- Study specification (funders, goal, what is done)
- Subjects details (age, sex, handedness, whatever)
- Individual folders per subjects (with subfolders and metadata)
- Derivatives (reproduce individual folders with secondary data if any)
- Code (all the stuff you do on the data documented here)

BIDS

Uses consistent and intuitive naming, including some metadata in the file name

```
subj-01/anat/subj-01-T1w.nii.gz
subj-01/anat/subj-01-T1w.json
subj-01/func/subj-01_task-reading_BOLD.nii.gz
subj-01/func/subj-01_task-reading_BOLD.json
subj-01/func/subj-01_task-reading_events.tsv
```

or

```
subj-01/technique/subj-01_<key>-<value>_..._datatype.png (data)
subj-01/technique/subj-01_<key>-<value>_..._datatype.txt (metadata)
```

File naming

File naming - principles

Principles for file naming

- human readable
- machine readable
- plays well with default sorting and ordering

File naming - exercise

How can we improve file naming?

Think about the file order in MATLAB/Python/R vs. Windows/Linux/macOS -- From the GitHub Repository https://github.com/CPernet/ReproducibleQuantitativeDataScience/tree/main/naming_files rename files

Work in teams, you have 10 minutes

Teacher notes: organize students in groups and share on white board allowing to compare side by side the different approaches

File naming - exercise

How can we improve file naming?

Think about the file order in MATLAB/Python/R vs. Windows/Linux/macOS
-- From the GitHub Repository
https://github.com/melanieganz/ReproducibleQuantitativeDataScience-2024/tree/main/naming_files rename files

Work in teams, you have 10 minutes

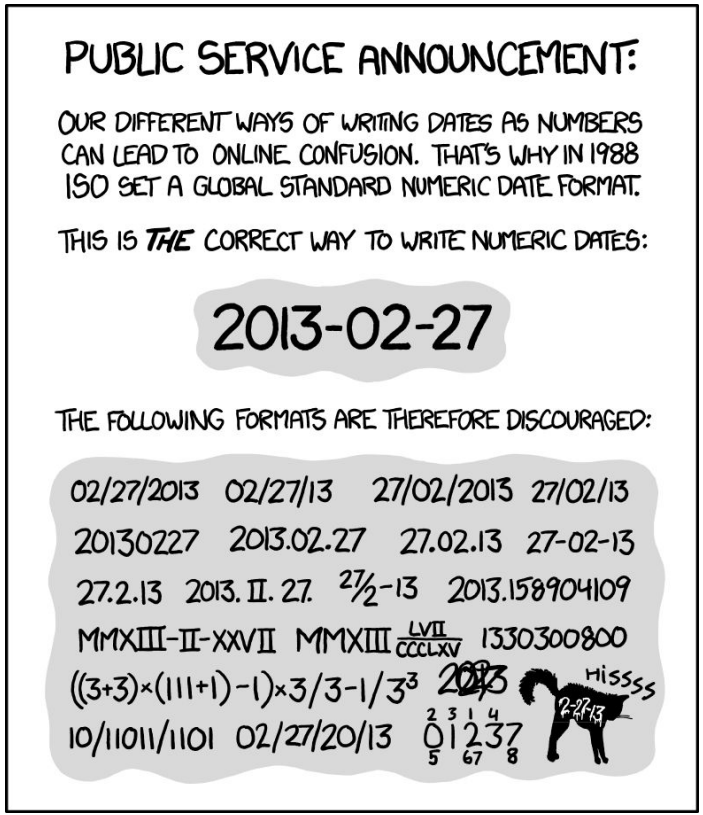
File naming - principles

“human readable” → name contains info on content, also logical order?

“machine readable” name can be constructed and/or parsed in a script

“plays well with default ordering” → make numbers part of the file name and left pad them with zeros (01, 02, … 99)

chronological order → use the ISO 8601 standard for dates



https://xkcd.com/1179/

File naming - principles

Machine readable facilitates use of RegExp (regular expressions) and globbing (wildcards).

Use of delimiters `“_”` underscore (used to delimit units of metadata) and `“-”` hyphen (used to delimit words), like `sub-01_task-attention_rawdata.ext`

- easy to search for files and filter
- easy to extract info by splitting file name in pieces

Be kind to yourself and avoid
- spaces and punctuation `. : ; , “ ‘`
- accented and special characters `$ @ & é æ ...`
- inconsistent/different capitalizations of files names like “foo” and “Foo”

File naming - parsing and sorting

Different ways to use upper/lower case.
For example snake_case, camelCase, kebab-case, ...

From: https://allisonhorst.com/everything-else

Short discussion: Why use case in file names and why not?

http://www.malte-grosser.com/post/introducing-the-snakecase-package/

File formats

File formats

Long-term accessible

Accessible in different (current and future) software

Accessible on different operating systems

Not encumbered with patents or use-limitations (mp3, gif)

Open Formats, often based on underlying standard
- MATLAB *.mat = hdf5
- MS Office *.xlsx and *.docx = zip+xml
- ascii, unicode
- pdf, tiff, png

Archiving

Archiving

You might have multiple storage systems, each with a different organization.

Consolidate everything when you archive.

Don’t leave multiple copies around as sloppy backups, move (or copy-and-delete) them so that you (and others) know where the final version is.

Don’t leave a full (but outdated) copy on the old system, but leave a note that redirects to the new location of the files.

Tools

Tools for managing data and code

Windows Explorer, macOS Finder, ...
Unix command line: ls, grep, find, locate, tree, ...
Rsync, CyberDuck, WinFTP, WinSCP, WebDAV, ...
Md5sum manifest, ...
Pip, VirtualEnv, Conda, Mamba, ...
git + GitHub/GitLab/etc + GitHub Desktop/GitKraken, ...
Datalad, Git annex, ...
... which additional tools are you using?


c|net