Data organisation practices

Introduction to Data Management Practices course

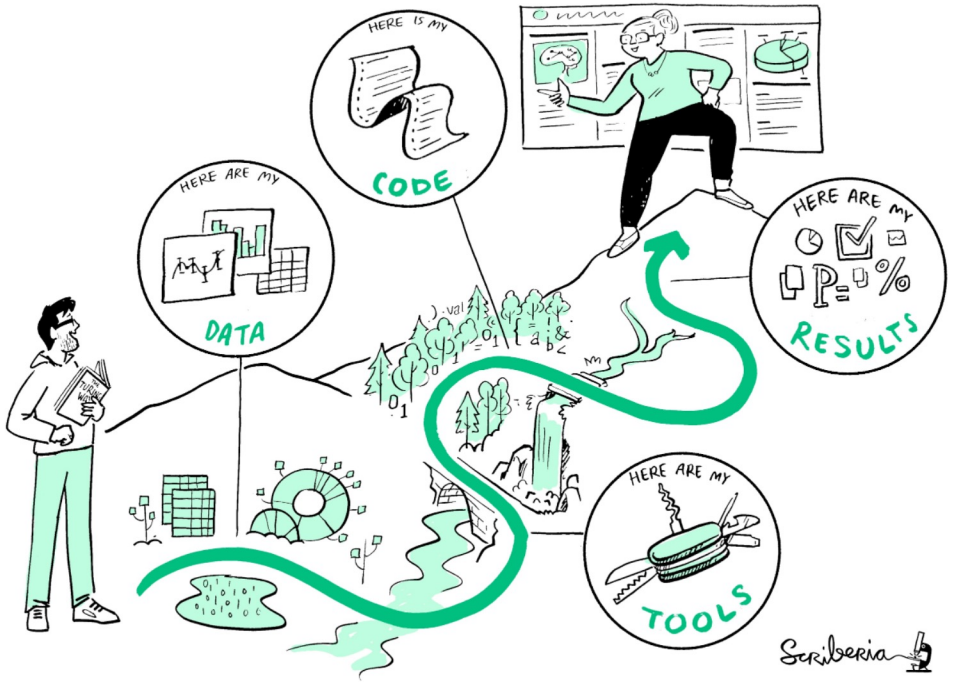
NBIS DM Team

data@nbis.se

https://nbisweden.github.io/module-organising-data-dm-practices/

Objectives

- What to consider for maintaining data organization strategies in a project
- What to consider when settling for a file structure
- Understanding good practices for data storage, processing and documentation (**FAIR-ification**)



Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Exercise 1

What measures do you take in order to avoid file chaos in your data organisation?

Managing Research Data

- Research **data is a core component** of any research project or publication.
- Good data management practices are **important in all phases** of research
  - Ethics and legislation
  - Information security
  - Research documentation
  - Project organisation
- Research data needs to stay authentic and be secured **beyond the project**'s time frame

Credit: Illustration from Digitalbevaring.dk / Jørgen Stamp (CC BY 2.5 Denmark license).

The Data Lifecycle

Data *content* is subject to changes at all phases of its life cycle

- Creation
- Error corrections
- New variables
- Changing file formats
- Etc.

Data *structure* may change at all phases of data life cycle

- Data splits
- Re-organization
- Storage change
- Etc.

Planning for Data Management

- Large impact in the planning phase
  - Intended folder structure
  - Clustering of files
  - File naming convention
  - Standards for dates/measures
  - Documentation procedures
- Prepare your project for receiving data by making a **Data Management Plan**

Data Management Recipients

- Colleagues
- Scientific community
- Society
- Yourself

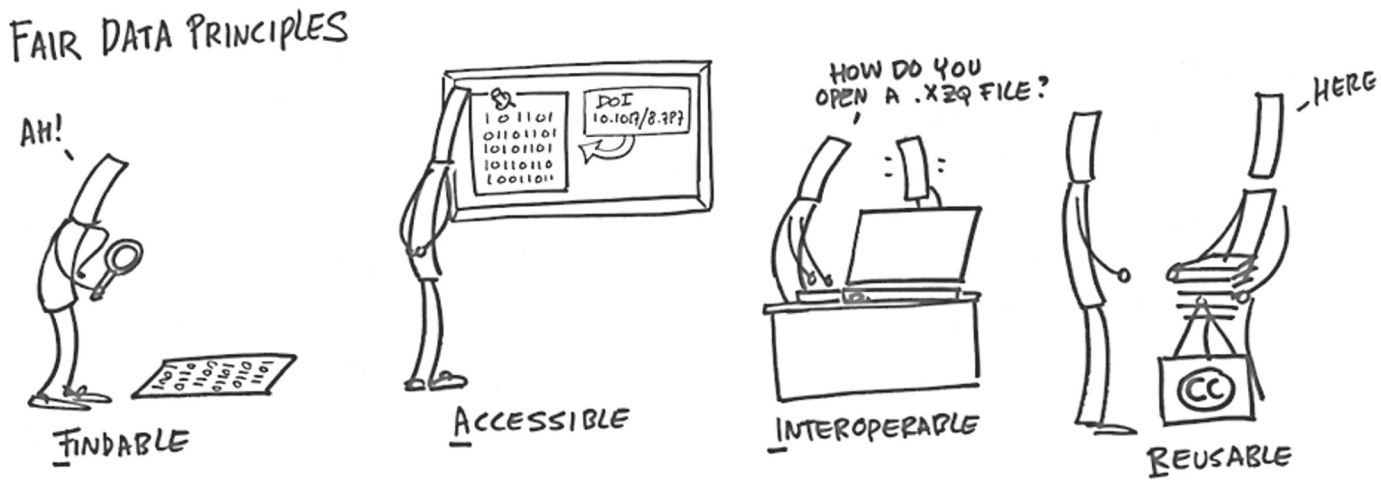
People I collaborate with must understand what I do with the data

Scientists wanting to reuse or review my data can find and understand the data

The society funding my research have a right to know what happens to the data

Your future You will not always remember what Your present You decided today

Reach for FAIR

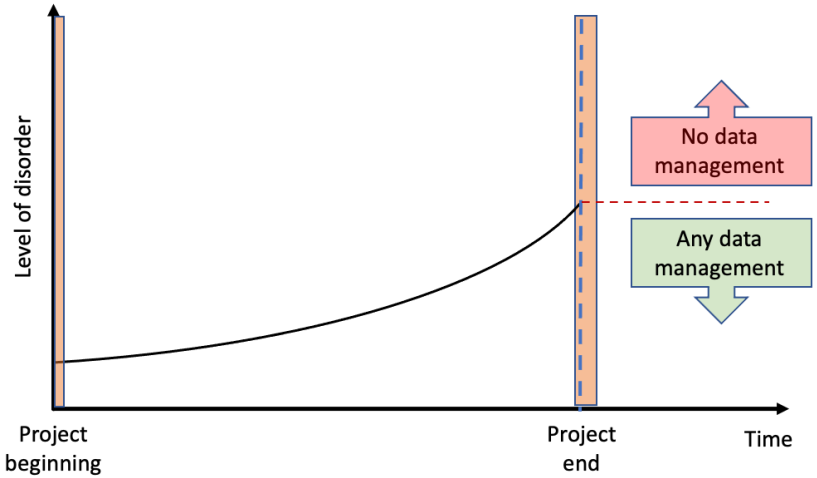


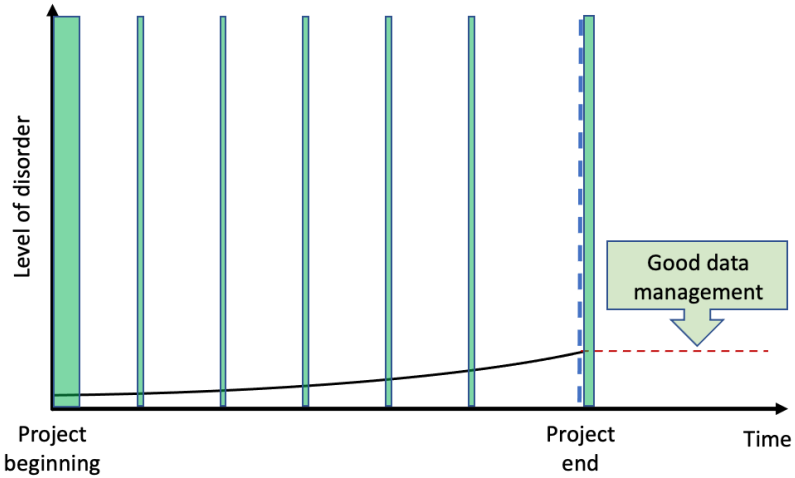
Adopting good practices for data organization, makes research data more **FAIR**

Image: https://book.fosteropenscience.eu/

Time vs. Effort

Planned data organization reduces disorder and optimizes time vs. effort





Exercise 2

Rank the following data organization steps from 1-5 (1 being the one you believe you think is most important, and 5 the least). Also mark with an "X" the steps you have implemented in your own research.

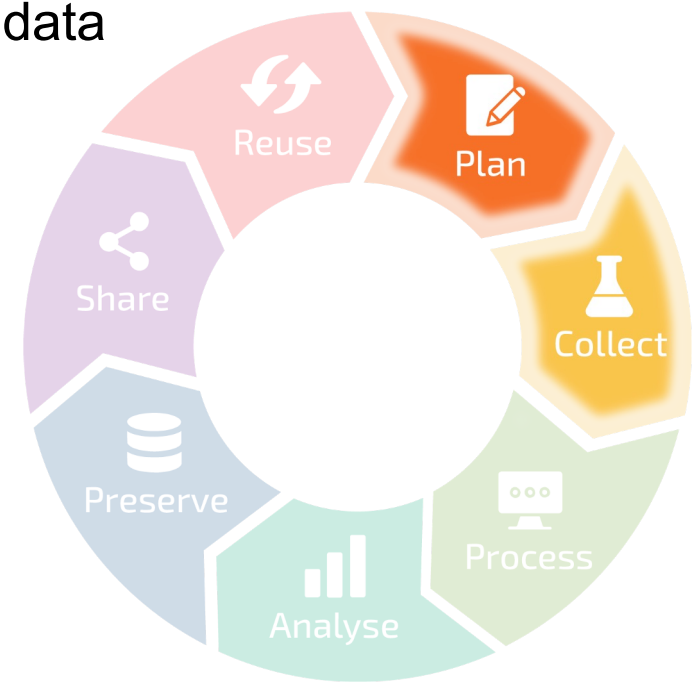
- File naming convention
- Folder naming convention
- File versioning system
- File organisation documentation (README.txt)
- File and folder maintenance (moving, deleting)

Raw data

- Raw data is the purest form of scientific data
- NOT for analysis
- Preservation and access restriction

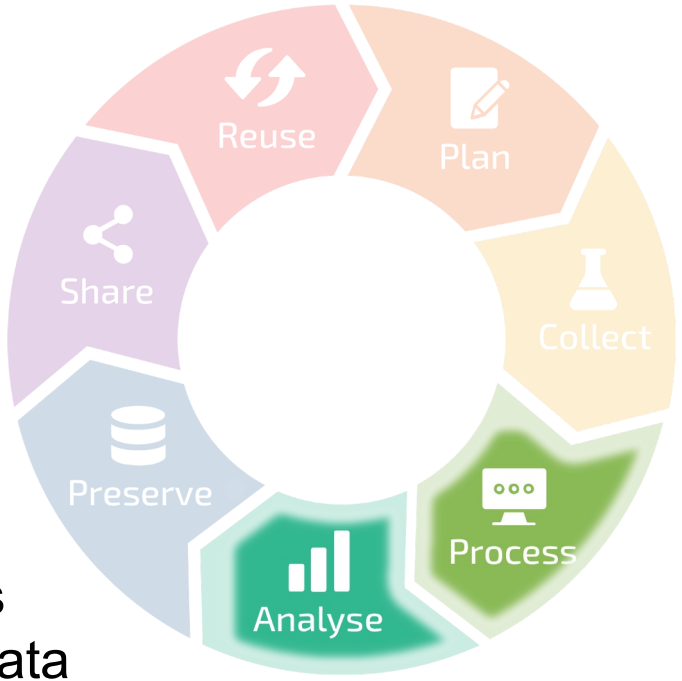
- Extense of raw data package(s)
- Versioning
- Documentation

- Status of sub-selections
- Associated data
- Long term / Publication package



Analyses and code

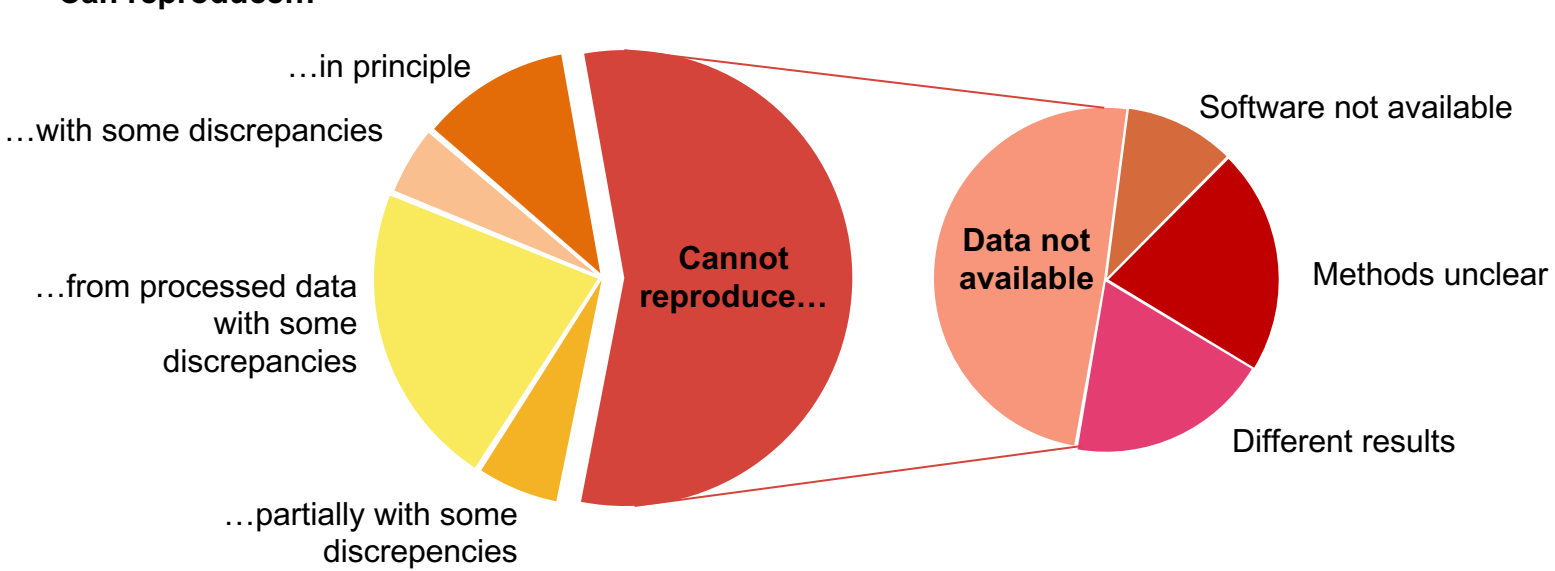
- Data (subset) + Analyses = Results
- Document for reproducibility
- Describe analyses, avoid ”default”
- Self-sufficient descriptions



- Rich metadata
- Store code close to data
- Document and publish code for analyses
- Cross-reference code and data in metadata

Analyses and code

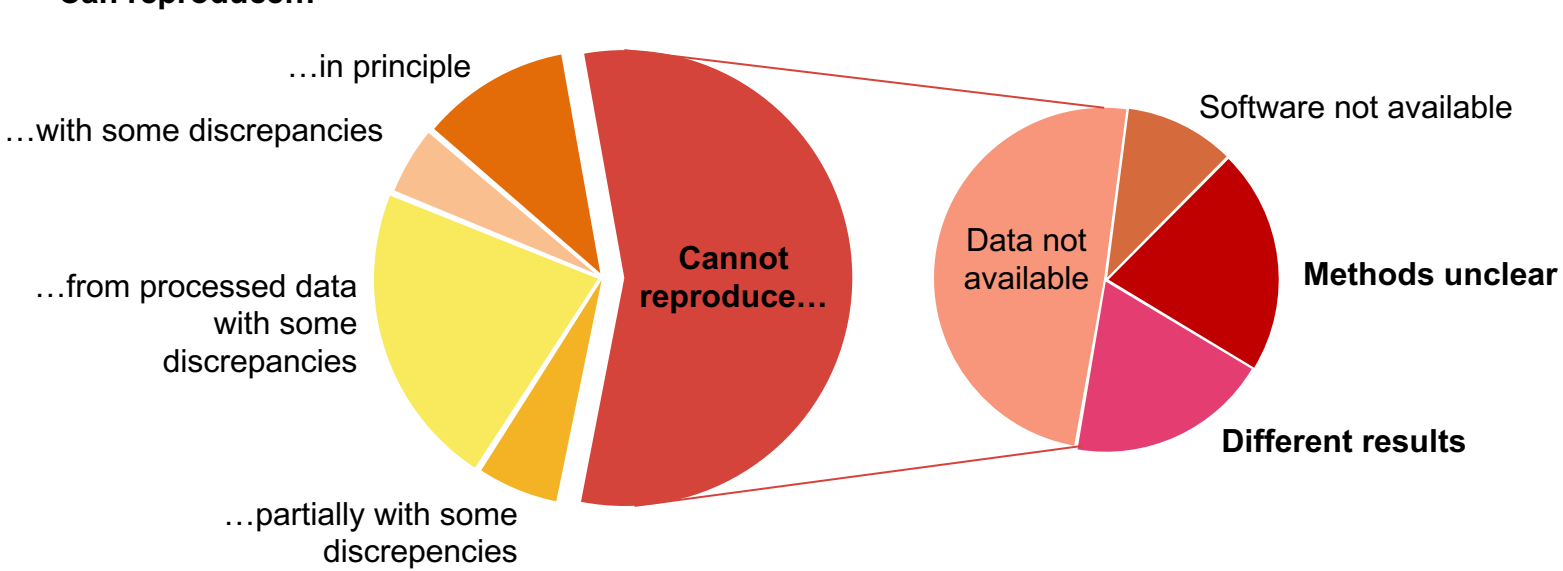
Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:



Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41** (2009) doi:10.1038/ng.295

Analyses and code

Reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:



Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses. *Nature Genetics* **41** (2009) doi:10.1038/ng.295

Tabular data

- Tabular data is not a data type, but a mode to organise data
- Do not mix with notes/code/analyses
- Data visualization ≠ data use
- Missing values

- Interoperable as .csv or .tsv

Tabular data

- Column = Variable
- Row = Observation
- Cell = Value

Open Access training					
Date	Length (hours)	Registered	Attended	Delivered by	Canceled
16/01/17	1	26	23	JM	N
05/02/17	1	38	26	JM	N
17/02/17	1	19	25	PG	N
07/03/17	1	27	17	JM	N
29/03/17	1	32	15	PG	N
02/04/17	1	41		PG	Y
24/04/17	2	44	44	JM	N
25/05/17	1	43	37	PG	N
16/06/17	1	15	15	JM	N

Tabular data

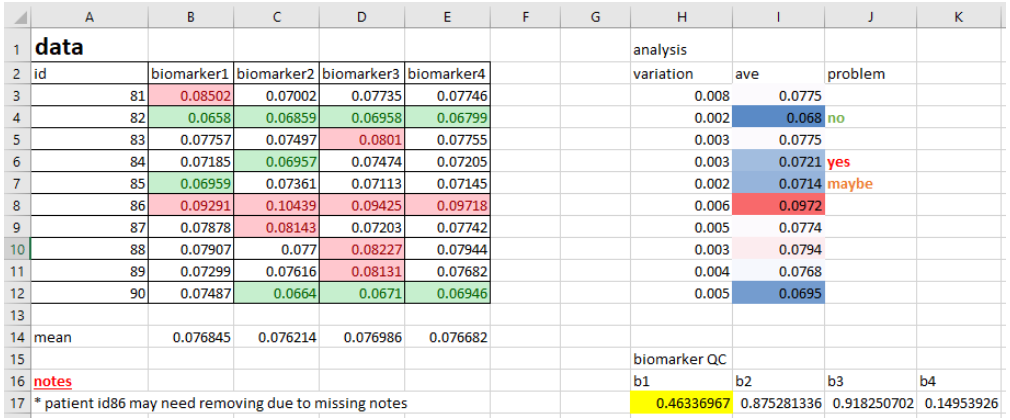
Do not:

- Spatially distribute data
- Combine values in cells
- Split compatible data in tables
- Use colors

	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2		RDM training					Open access						
3		Date	Length (hours)	PGR|PDRA|other	Delivered by		Date	Len	Attendees	Delivered by		
4		12 Jan	1.5	45|0|0	FG		8 Jan	1.5 hours	20	FG		
5		7 Feb	2	38|0|0	GH		13 Jan	1 hour	21	JM		
6		4 Mar	2	43|3|0	GH		22 Jan	1 hour	35	JM		
7		6 Mar	1	21|7|0	GH		2 Feb	1.5 hours	36	JM		cancelled
8		17 Mar	1.5	34|1|0	FG		3 Feb	1.5 hours	22	JM		
9		21 Mar	1	25|2|0	DQ		3 Feb	1 hours	30	JM		
10		23 Mar	2	32|10|0	FG		20 Feb	1.5 hours	36	FG		
11		19 Apr	1	34|0|0	GH		28 Feb	1.5 hours	28	JM		
12		30 Apr	1.5	37|0|0	FG		19 Mar	1.5 hours	33	FG		
13		4 Jun	1	45|0|0	GH		19 Mar	1 hour	39	JM		
14		12 Jun	2	36|0|0	DQ		4 Apr	1.5 hours	21	JM		
15		22 Jun	1.5	38|0|0	DQ		5 May	1.5 hours	25	JM		
16		25 Jun	1	35|4|0	GH		18 May	1 hour	22	JM		
17		30 Jun	1.5	44|3|0	FG		19 May	1.5 hours	20	FG		
18		1 Jul	1.5	40|0|4	FG		21 May	1.5 hours	21	JM		
19		6 Jul	1.5	21|0|0	GH		14 Jun	1.5 hours	37	JM		
20		7 Jul	1	37|4|1	DQ		18 Jun	1.5 hours	25	JM		
21		9 Jul	1	29|7|0	GH		4 Jul	1.5 hours	39	JM		
22		30 Jul	2	22|3|0	FG		6 Jul	1.5 hours	39	JM		
23		29 Aug	1.5	22|4|0	GH		10 Jul	1.5 hours	34	JM		
24		10 Sep	1	38|0|0	FG		13 Jul	1.5 hours	23	FG		
25		21 Sep	1	31|0|0	GH		17 Jul	1.5 hours	30	JM		
26		1 Oct	2	26|9|5	DQ		3 Aug	1.5 hours	28	JM		
27		25 Oct	1.5	20|4|0	DQ		20 Aug	1.5 hours	32	JM		
28		4 Nov	1.5	38|5|5	FG		26 Aug	1.5 hours	25	JM		
29		5 Nov	2	40|0|0	GH		28 Aug	1.5 hours	33	FG		
30		8 Nov	2	22|7|0	FG		1 Oct	1.5 hours	38	JM		
31		1 Dec	2	41|6|0	DQ		21 Oct	1.5 hours	34	JM		
32		19 Dec	2	39|9|1	GH		9 Nov	1.5 hours	32	JM		
33							15 Nov	1.5 hours	35	JM		
34							15 Nov	1.5 hours	27	JM		
35							2 Dec	1.5 hours	35	FG		
36							7 Dec	1.5 hours	23	JM		
37							11 Dec	1.5 hours	38	FG		
38							19 Dec	1.5 hours	20	FG		
39													

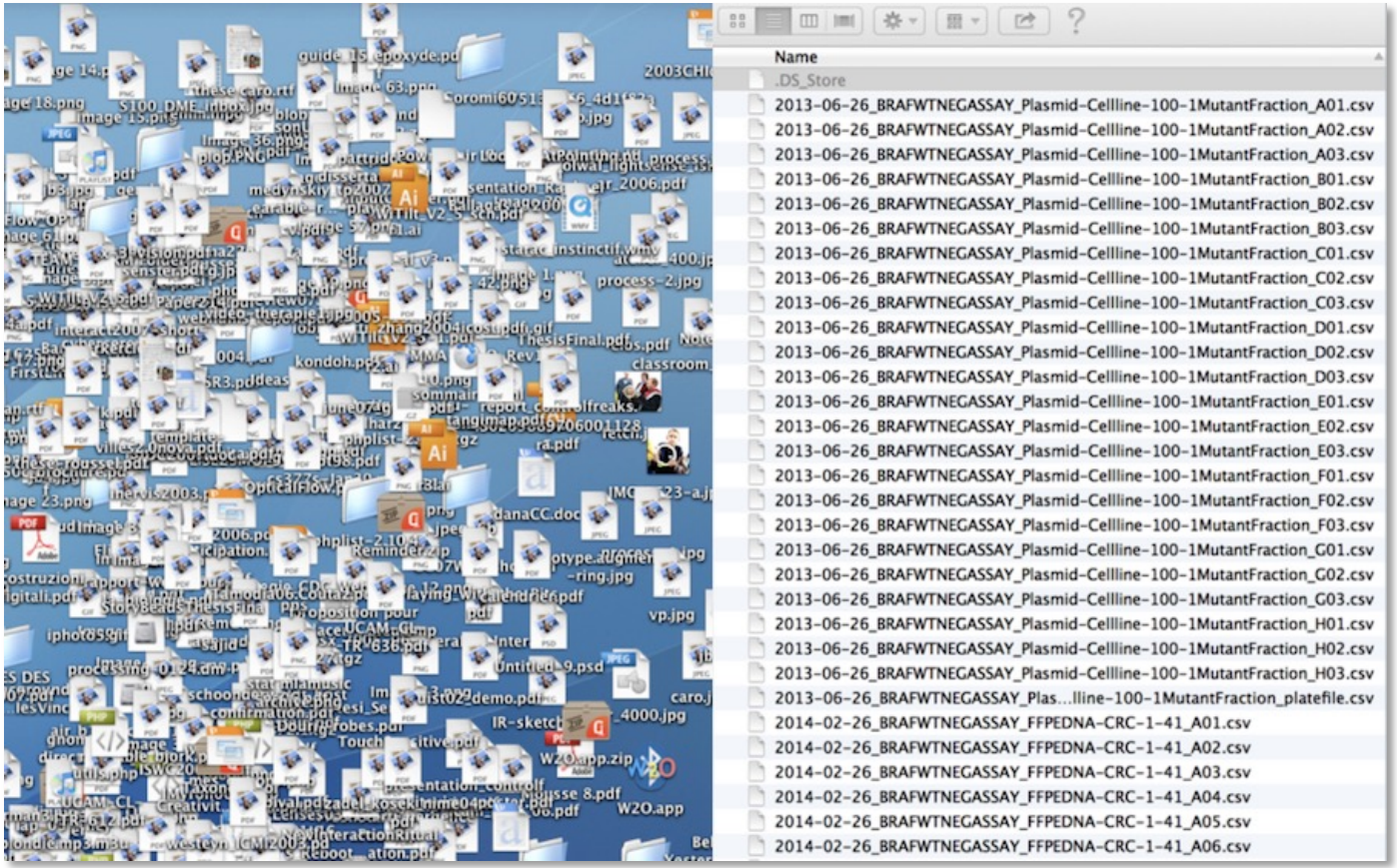
Tabular data

- ✓ Raw means raw!
- ✓ Tidy data tables
  - One cell–one value
  - One column–one variable
  - One row–one observation
- ✓ Beware of Excel “features”
  - Misguided ”auto-corrections” of dates, casing, numbers etc.
  - Misaligned formulas
  - Limited numerical precision
  - Limited number of rows/columns



	A	B	C	D	E	F	G	H	I	J	K
1	**data**							analysis			
2	id	biomarker1	biomarker2	biomarker3	biomarker4			variation	ave	problem	
3	81	0.08502	0.07002	0.07735	0.07746			0.008	0.0775		
4	82	0.0658	0.06859	0.06958	0.06799			0.002	0.068	no	
5	83	0.07757	0.07497	0.0801	0.07755			0.003	0.0775		
6	84	0.07185	0.06957	0.07474	0.07205			0.003	0.0721	yes	
7	85	0.06959	0.07361	0.07113	0.07145			0.002	0.0714	maybe	
8	86	0.09291	0.10439	0.09425	0.09718			0.006	0.0972		
9	87	0.07878	0.08143	0.07203	0.07742			0.005	0.0774		
10	88	0.07907	0.077	0.08227	0.07944			0.003	0.0794		
11	89	0.07299	0.07616	0.08131	0.07682			0.004	0.0768		
12	90	0.07487	0.0664	0.0671	0.06946			0.005	0.0695		
13											
14	mean	0.076845	0.076214	0.076986	0.076682						
15								biomarker QC			
16	notes							b1	b2	b3	b4
17	* patient id86 may need removing due to missing notes							0.46336967	0.875281336	0.918250702	0.14953926

Organising files and folders



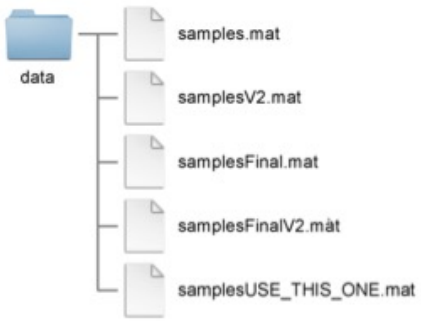
Practices

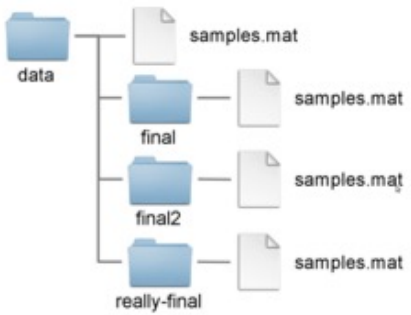
Your primary collaborator is yourself from 6 months ago, and that person is really difficult to communicate with!

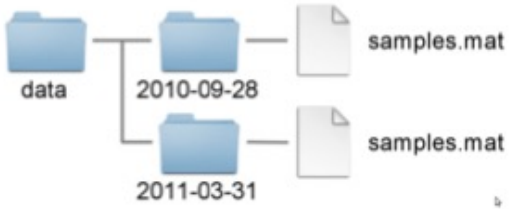
Good practices

- Organise files hierarchically
- Use folders to divide files into categories
- Choose a file naming strategy
- Create documentation files

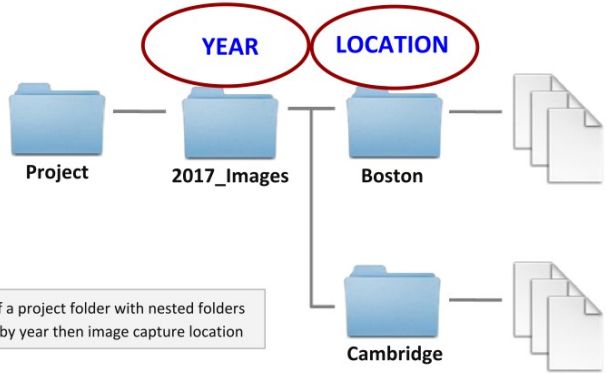
Divide files into categories





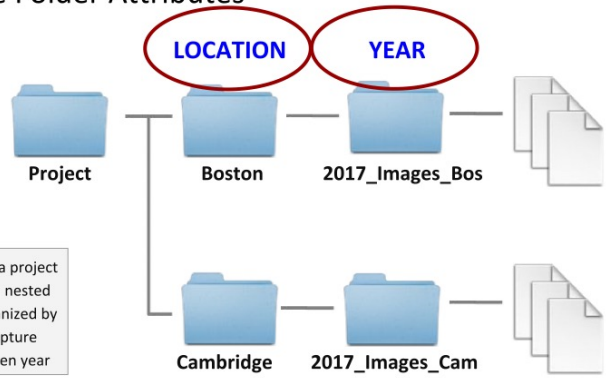


Create Folder Attributes



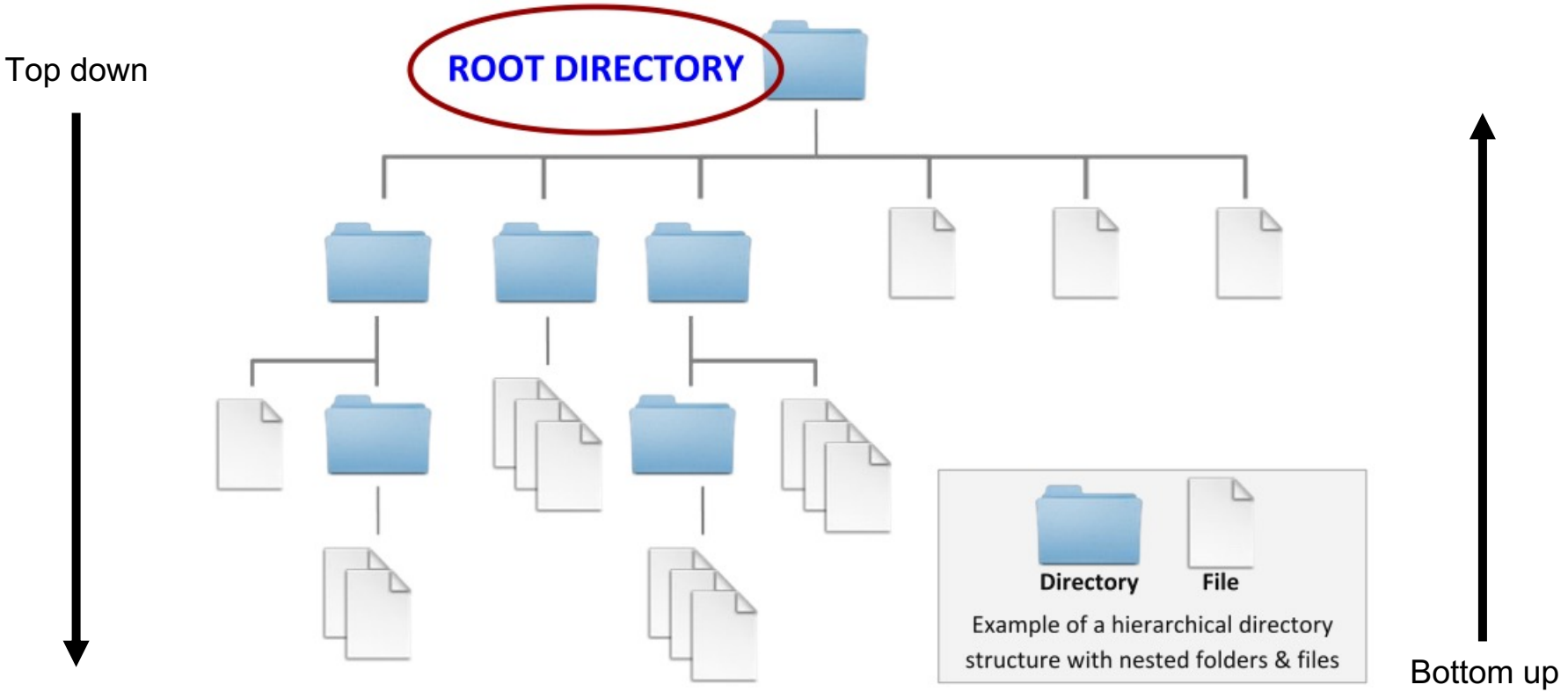
Example of a project folder with nested folders organized by year then image capture location

Create Folder Attributes

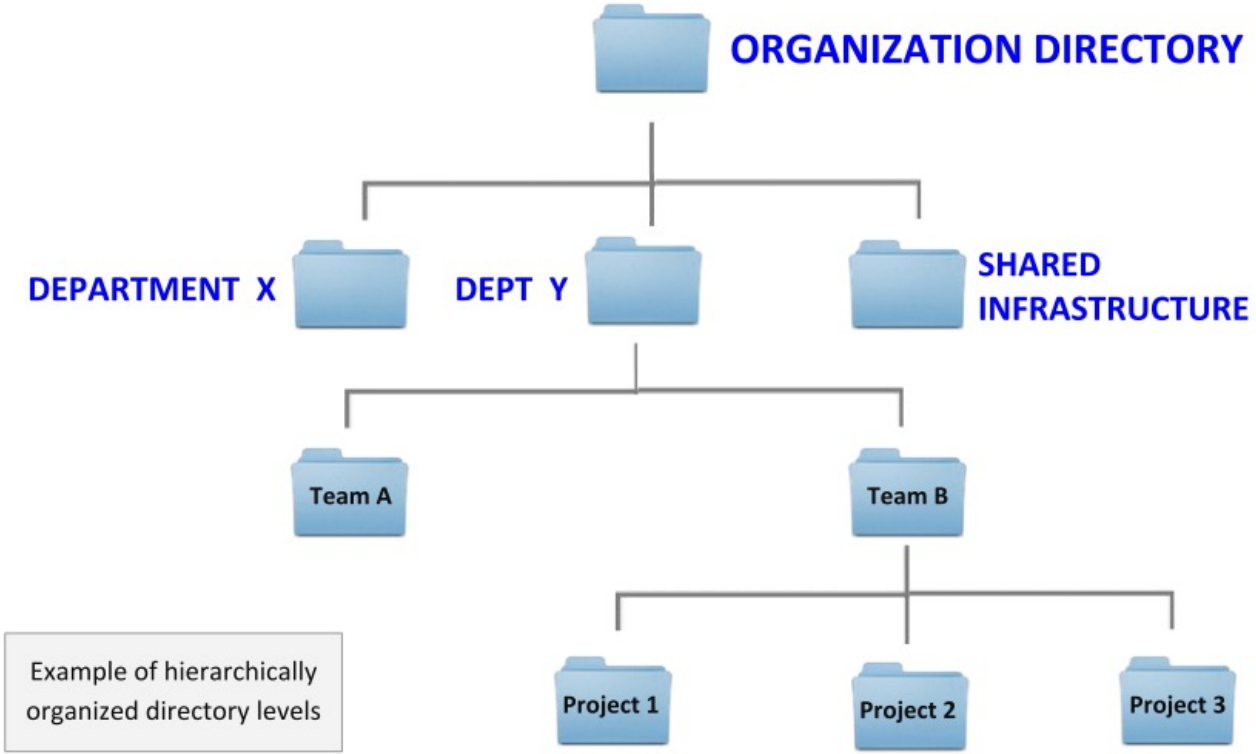


Example of a project folder with nested folders organized by image capture location then year

Organise files hierarchically

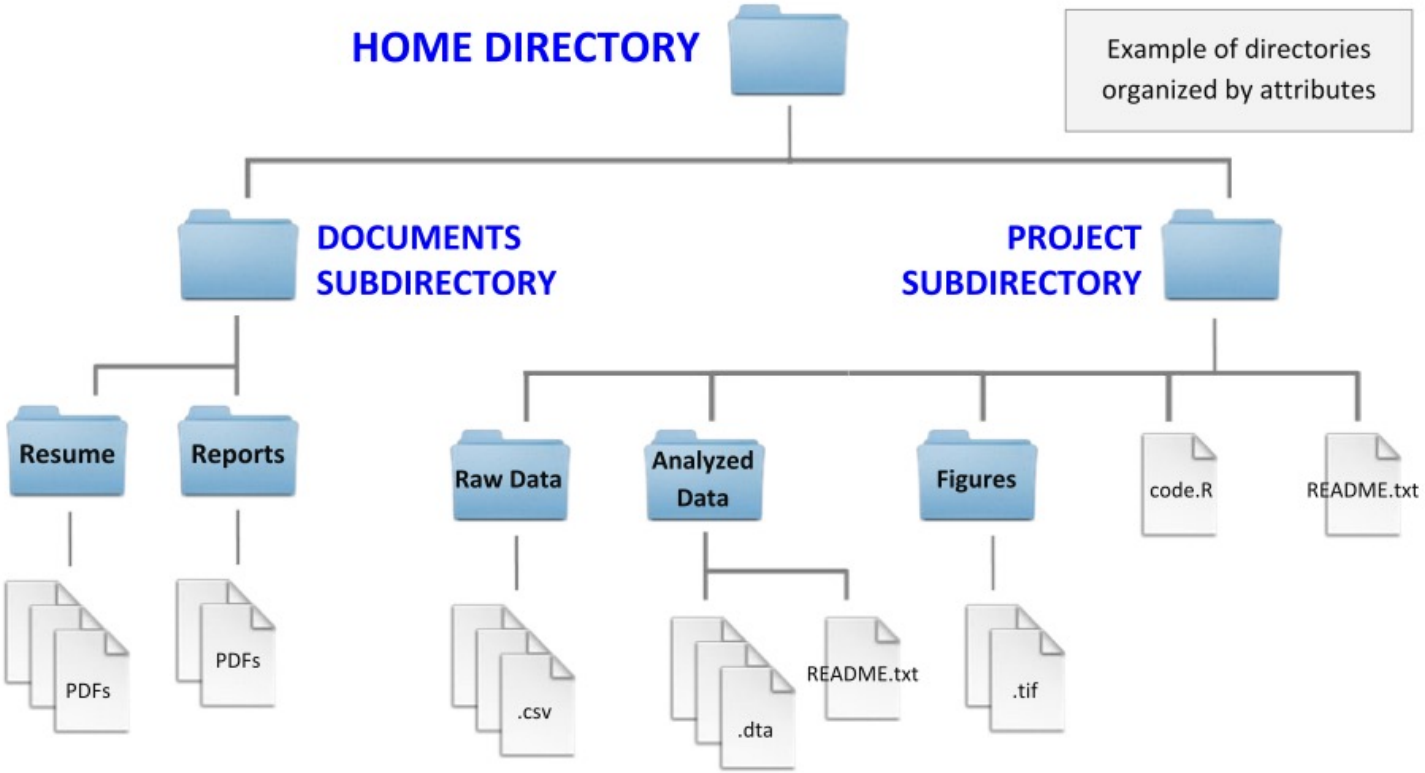


Hierarchy?

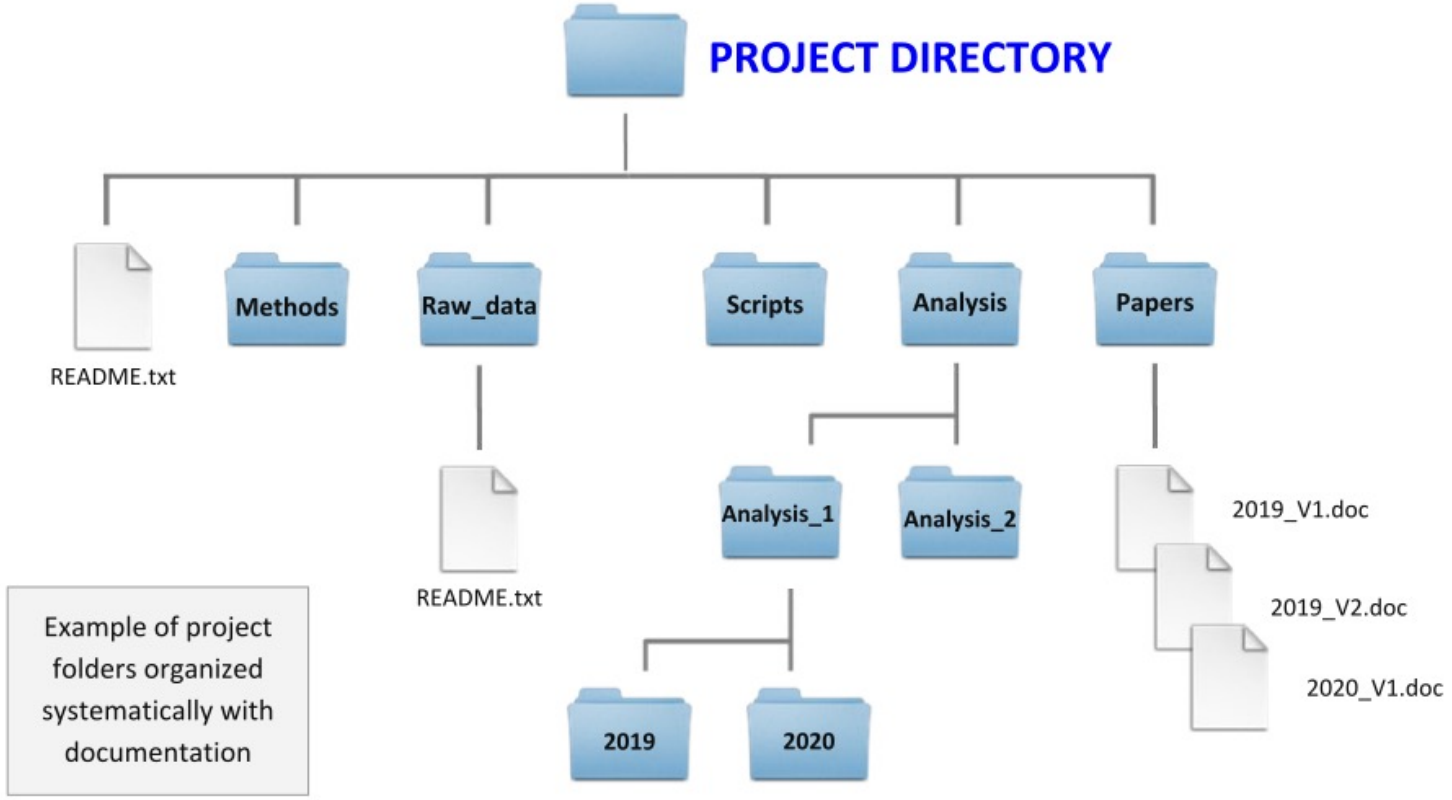


Example of hierarchically organized directory levels

Hierarchy?


HOME DIRECTORY
Example of directories organized by attributes
DOCUMENTS SUBDIRECTORY
PROJECT SUBDIRECTORY
Resume
Reports
Raw Data
Analyzed Data
Figures
code.R
README.txt
PDFs
PDFs
.csv
.dta
README.txt
.tif

README.txt


PROJECT DIRECTORY
README.txt
Methods
Raw_data
Scripts
Analysis
Papers
README.txt
Analysis_1
Analysis_2
2019_V1.doc
2019_V2.doc
2020_V1.doc
2019
2020
Example of project folders organized systematically with documentation

Exercise 3

Considering your own file structure, or the file structure used in your research group, write in the shared document:

1. If your current file structure is top-down or bottom-up, or a mix of the two.
2. Who the "inventor" of your file structure is (title, not name!), and to what extent you can influence how it is organized?
3. One thing you believe you could improve in your current file structure.

Digitalbevaring.dk

File names

- Year or other date
- Type of data, document or file
- Project stages
- Analysis version or revision
- Experiments
- Instruments
- Time periods
- Geographic location
- Storage requirements
- Team member, institution or project site

Helpful characteristics!

Digitalbevaring.dk

Choose a file naming strategy

- A file name is a principal identifier for the file
- Consistent in time and among different people
- Practically useful when accessing files, such as sorting and filtering

Chronologically
(ISO 8601 date standard)

```
20171028_001.tiff
20171028_002.tiff
20171028_003.tiff
20171029_001.tiff
20171029_002.tiff
```

Classification or code
(standardized)

```
USNM_379221_01.tiff
USNM_379221_02.tiff
USNM_379221_03.tiff
USNM_379222_01.tiff
USNM_379222_02.tiff
```

Alphabetically
(depending on type of files)

```
bos_20171028_001.tiff
bos_20171028_002.tiff
bos_20171029_001.tiff
cam_20170922_001.tiff
cam_20170922_002.tiff
```

- Human readable
- Machine readable

File naming

A well suited file naming protocol should be:

1. **Human readable**

   A name describes the content of the file, connects to concept of a *slug* from semantic URLs (e.g. www.scilifelab.se/*this-is-a-slug*).

2. **Machine readable**

   Avoid spaces, deliberate punctuation, accented or odd characters, inconsistent letter casing

3. **Default ordered**

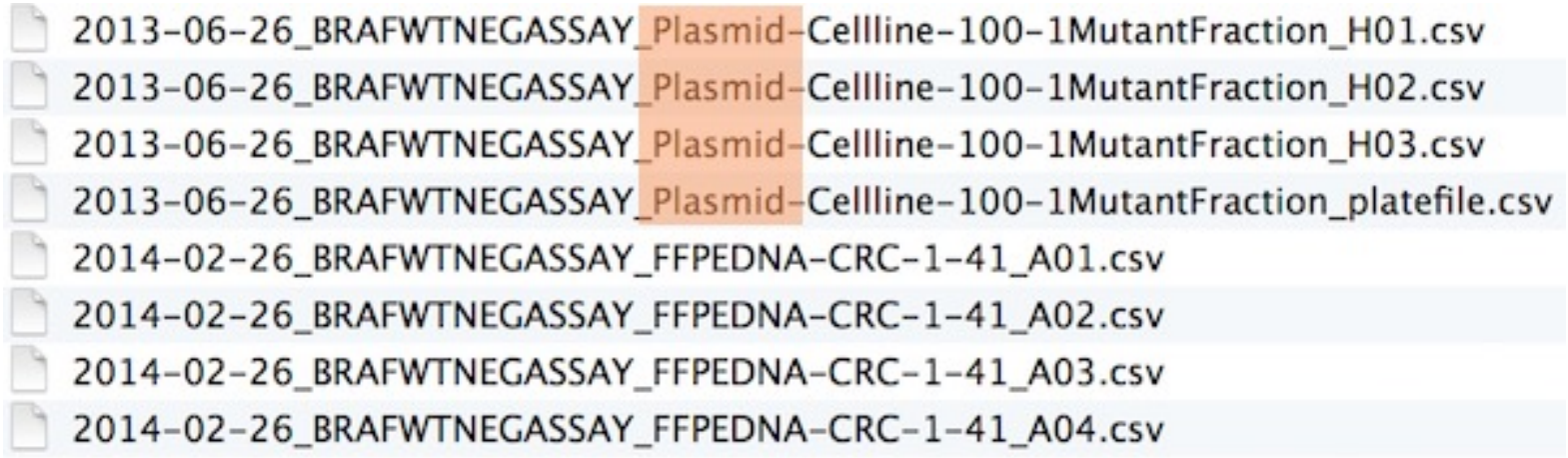
   Put something numeric first, use the ISO 8601 standard for dates (YYYYMMDD, or YYYY-MM-DD), left pad single digits with zeros (01, 02, 03... 10)

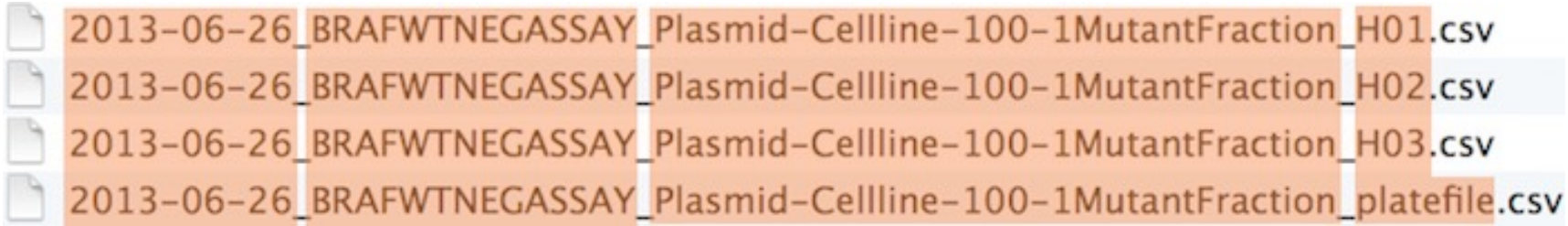
Human readable

```
01_marshal-data.md
01_marshal-data.r
02_pre-dea-filtering.md
02_pre-dea-filtering.r
03_dea-with-limma-voom.md
03_dea-with-limma-voom.r
04_explore-dea-results.md
04_explore-dea-results.r
90_limma-model-term-name-fiasco.md
90_limma-model-term-name-fiasco.r
Makefile
figure
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
tmp.txt
```

```
01.md
01.r
02.md
02.r
03.md
03.r
04.md
04.r
90.md
90.r
Makefile
figure
helper01.r
helper02.r
helper03.r
helper04.r
tmp.txt
```

Machine readable





File naming

Examples of a **poor** file name**:**

”Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020”

File name - Runnew_again_2NDTRY.xls

Explanation - N/A

File naming

Examples of a **good** file name**:**

”Honeybee project, experiment 2 done in Helsinki, data file created on the second of December 2020”

File name - 20201202_HB_EXP2_HEL_DATA_V03.xls

Explanation - Time_ProjectAbbreviation_ExperimentNumber_ Location_TypeOfData_VersionNumber

Documenting file structure

- Ensures orientation and longevity of data
- Add to top level folder in separate plain text file
- Collect all information to be documented and cross-reference where necessary

➢ Shared projects where files are handles by several individuals with different areas of responsibility

➢ Project ending to guarantee rememberance of file structure over longer time (past your present You)

Conditions and use

- Consider upgrading your file structure with an explicit **Conditions and use** agreement, and apply it to file permissions, at least for raw data.

- Edits and changes to raw data should be restricted to only a few trusted individuals with particular responsibilities.

Digitalbevaring.dk

Credit: Illustration from Digitalbevaring.dk / Jørgen Stamp (CC BY 2.5 Denmark license).

Increasing file findability

When we collect data and organise it for research purposes, findability is not necessarily our primary motivation.

- Increase the number of **file copies**

  Pros - Someone (always?) have the latest version of a file!
  Cons - Who has the latest version of a file?

- Increase the number of **shortcuts**

  Pros - Easy to create, file name of shortcut can be changed and may even increase findability when named differently.
  Cons - Shortcut may break if original file location or name is changed. Easy to lose orientation if not maintained regularly.

Increasing file findability

When we collect data and organise it for research purposes, findability is not necessarily our primary motivation.

- Increase the amount of **metadata**

Pros - Possible to enrich any file with unlimited amounts of metadata

Cons - Can be cumbersome to keep updated as number of files increase and file names are changed

Increasing file findability

Keyword tagging

```
(Metadata.txt content)

20220115_MyFile_Project1_Location_Dataiteration1_V1.xml
First version of X data from Y, with additions of Z made by A and B on 20220110 includi
ng suggestions by C.
Keywords HumptyDumpty Genome_Assembly

20220115_MyFile_Project1_Location_Dataiteration2.xml
Contains X data from Y, with additions of Z made only by A on 20220111 not including su
ggestions by C.
Keywords Published
```

Associated metadata to increase findability of files over e.g. multiple projects

Exercise 4

For the following filename, construct a metadata explanation:

20220310_GenAnn_AssemblyProj2_GOT_Dataiteration_2nd_try.xml

Exercise 5

Using the downloaded compressed file directory *Example_project_begin.zip :*

1. Create a hierarchical folder system based on the file names and contents.

2. Rename files in a consistent manner if required, such that it reflects both contents and file version. Consider number and date formats as well.

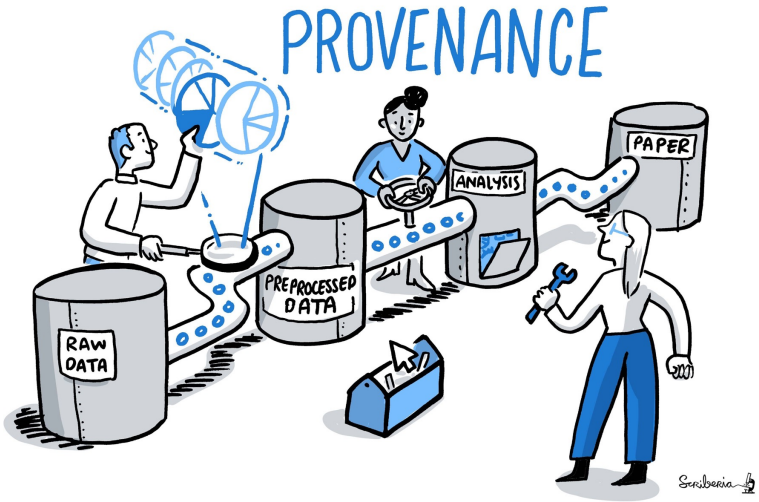
3. Optionally, create a file for tagging files with metadata and keywords in accordance with file contents.

Storage & back-up

- How to store during the project?
  - **Storage/processing locations**
    For data collection, analysis, reporting, code, transfers etc.
  - **Back-up and data recovery**
    Strategies to mitigate risks of data-loss and data corruption?
    (Beware of laptops and external storage)
  - **Technical requirements**
    Software and systems required to access / process the data?
- How to protect data?
  - **Information classification**
    Suitable storage based on the characteristics of the data?
  - **Access control**
    Who will have access to what data and how will it be enforced?
  - **Data protection procedures**
    Other strategies to mitigate risks of unwanted data disclosure or sabotage.

Data storage and processing

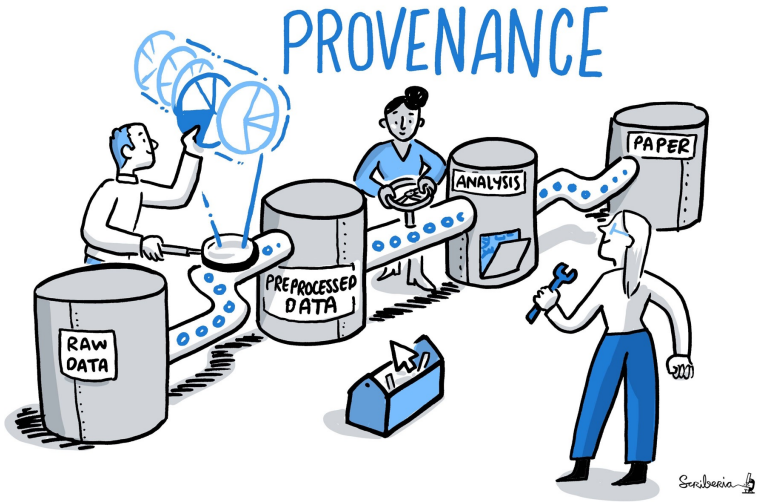
- Data has a life cycle
  - Raw (experiment) data – produce, collect, license, get access, …
  - Processed – generate, clean, aggregate, label, transform, analyse, …
  - Long-term storage – document, select, convert, package, submit, …
  - Published – FAIRify, promote reuse, …
- Maintain data integrity and authenticity
- Plan a storage strategy
- Plan a backup and disaster recovery strategy



Discussion 1

What storage categories do you use and what factors do you consider when selecting which category to use or not to use?

- **Portable devices** – Laptops, tablets, external hard-drives, flash drives and Compact Discs
- **Cloud storage** - E.g. Google Drive, OneDrive, Dropbox, a University's OwnCloud, Open Science Framework and Tresorit
- **Local storage** – Desktop computers and personal laptops
- **Networked drives** – Shared drives on university servers, NAS servers (Network Attached Storage) or infrastructures (such as SNIC)



Portable devices

- **Temporary, short-term storage for non-sensitive data**, e.g. in the field or to transport data and files when online transmission is not possible.
- In combination with **encryption and strong password protection**, especially if working with sensitive information.

✓ Conduct **regular checks** to ensure your device is working and that files are accessible.

❖ **Not for long-term storage or master copies** of your data

Cloud storage

- Granting **shared, remote and easy access to data and other files** to all involved in the project
- **Read the terms of service**. Especially focus on rights to use content given to the service provider
- **Opt for European, national, or institutional** cloud services which store data in Europe if possible

- **Not your only storage and backup solution**
- **Not for unencrypted (sensitive) personal data**

Local storage

- When working on different (local) workstations, e.g. laptop at home and the desktop in the office:
  - **always make sure that you are working on the most current version**, for example with the help of versioning software or guidelines
  - make sure that the most **current version is always backed up somewhere else**

✓ Suitable as a primary storage for projects involving only very few people

❖ Avoid if data will be moved back and forth between personal computers frequently

Networked drives

- **Use in projects involving many people** who need access to data and files
- Use a suitable security strategy to **protect data and files against unauthorised access**
- **Agree on rules for versioning files and data** to ensure that everyone can locate and access what they need
- **Long-term store complete data** that has been analysed, which can be cost efficient and offer increased security
- Restrict access where possible using rights and permissions, e.g. **write protect a master copy** and only grant access to specific files/folders when necessary

Backup and disaster recovery

A minimalist strategy

- ❑ There are at least
  - ➢ Three copies of the data, of which…
    - ✓ Two are kept on different types of storage media
    - ✓ Two are at different locations
    - ✓ One is located off-site
- ❑ All copies are checked regularly to make sure that they work
- ❑ The process is known and applied in the project (automated)

… also determine what you want to back up, and find out whether your institution already has a backup strategy.

Discussion 2

What are examples of potential causes for data loss in a research project?