NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN

SciLifeLab

Introduction to Data Management Practices course

NBIS DM Team
data-management@nbis.se

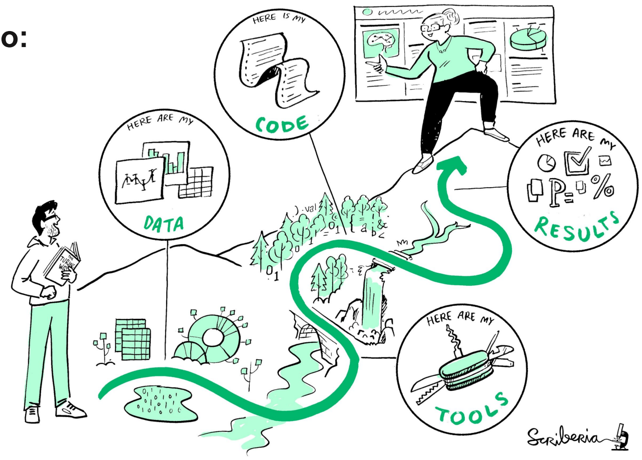
Data Organization practices

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

Session Objectives

After this session, you should be able to:

- Understand why good data organisation and records matter
- Learn practical ways to organise files, folders, and tabular data
- Adopt habits that support reuse and FAIR research (**FAIR-ification**)



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

Welcome to Science!

Real Life scenario…

Image: SciLifeLab for Press and Media: https://www.scilifelab.se/contact/for-press-and-media

You have been recruited to the "Famous lab"!

Your research project is a continuation of previous work by PhD, Emily Johnson

You inherit a zipped folder, and a digital copy of the lab notes.

The road to success is open!

Welcome to Science!

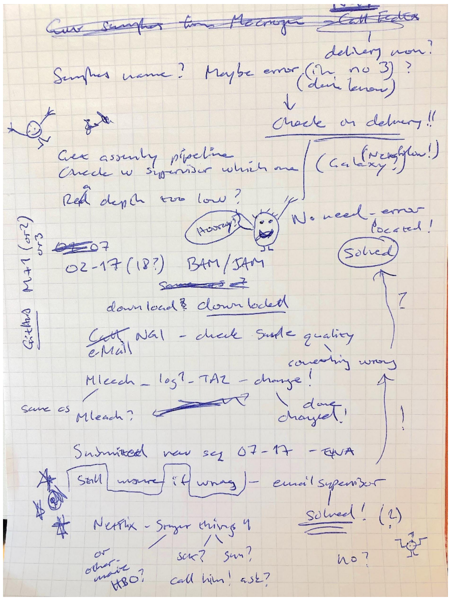
… And this is what you get…

Exercise 1

Can you list at least five major issues with the lab documentation in the image?

Exercise 2

What kind of general questions the note and Terry's answer raise about the work done in the lab.
(Who is responsible?)



~~Get samples from Macrogen~~ ~~Call FedEx~~

delivery now?

Samples name? Maybe error (in no 3) ?
(don't know)

Check on delivery!!

Get assembly pipeline
Check w supervisor which one (Galaxy?) (Nextflow!)

Read depth too low?

Hooray!

No need - error located!

Solved

M+1 (or 2) or 3

~~01~~ 07

02-17 (18?) BAM/JAM

~~Same as ?~~

download? downloaded

~~Call~~ eMail NGI - check Stude quality

something wrong

?

Mleech _ log? _ TAZ - change!

same as Mleach?

done changed!

!

Submitted new seq 07-17 - ENA

Still unsure if wrong - email supervisor

solved! (?)

Netflix - Stranger things 4

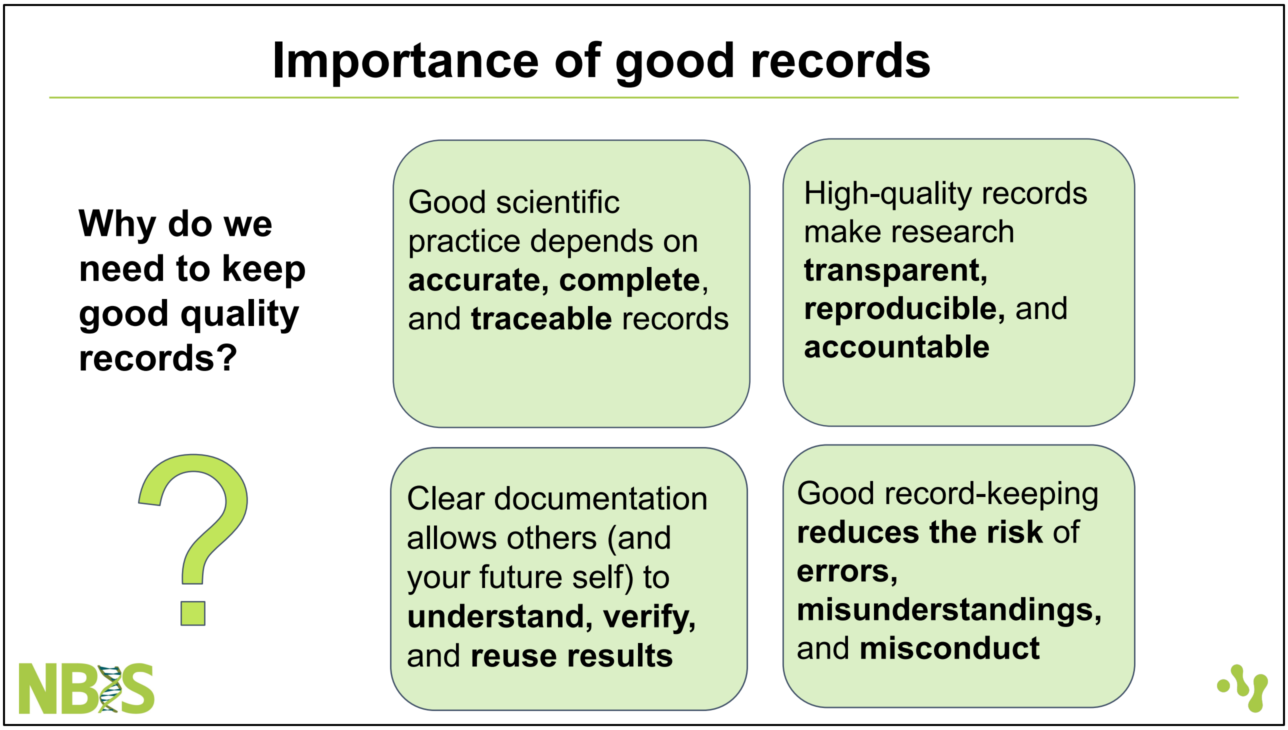
or other movie HBO?

sck? sun?

call him! ask?

no?

Github



Importance of good records

Why do we need to keep good quality records?

?

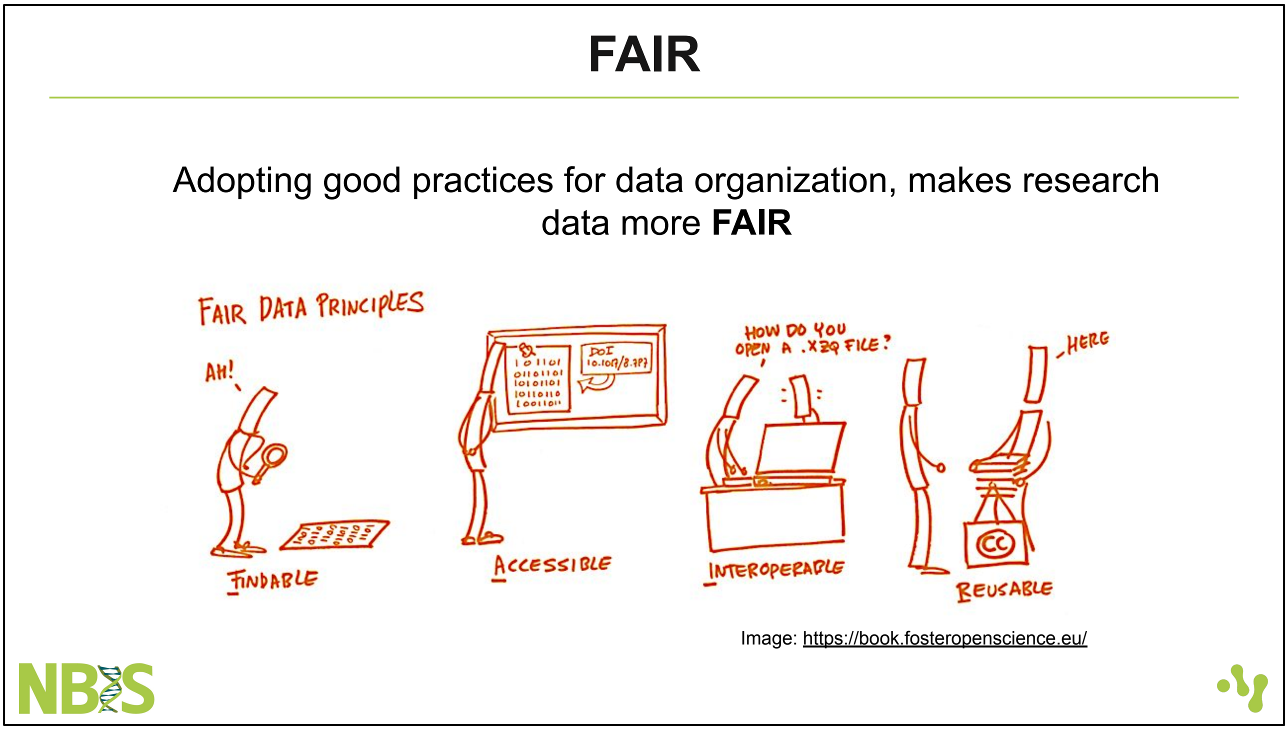
Good scientific practice depends on **accurate, complete**, and **traceable** records

High-quality records make research **transparent, reproducible,** and **accountable**

Clear documentation allows others (and your future self) to **understand, verify,** and **reuse results**

Good record-keeping **reduces the risk** of **errors, misunderstandings,** and **misconduct**

Science is cumulative — small documentation gaps can become big problems over time.



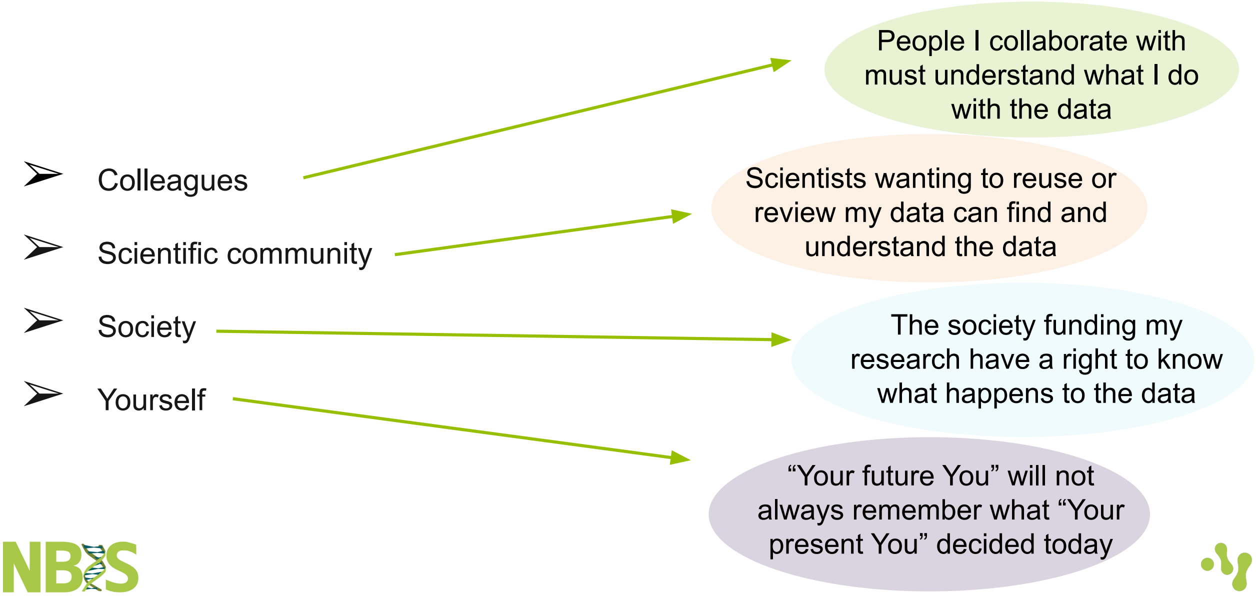
FAIR

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

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

Good-quality records are a foundation for FAIR data and documentation

Data 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

Data, metadata, documentation, and records — how they relate

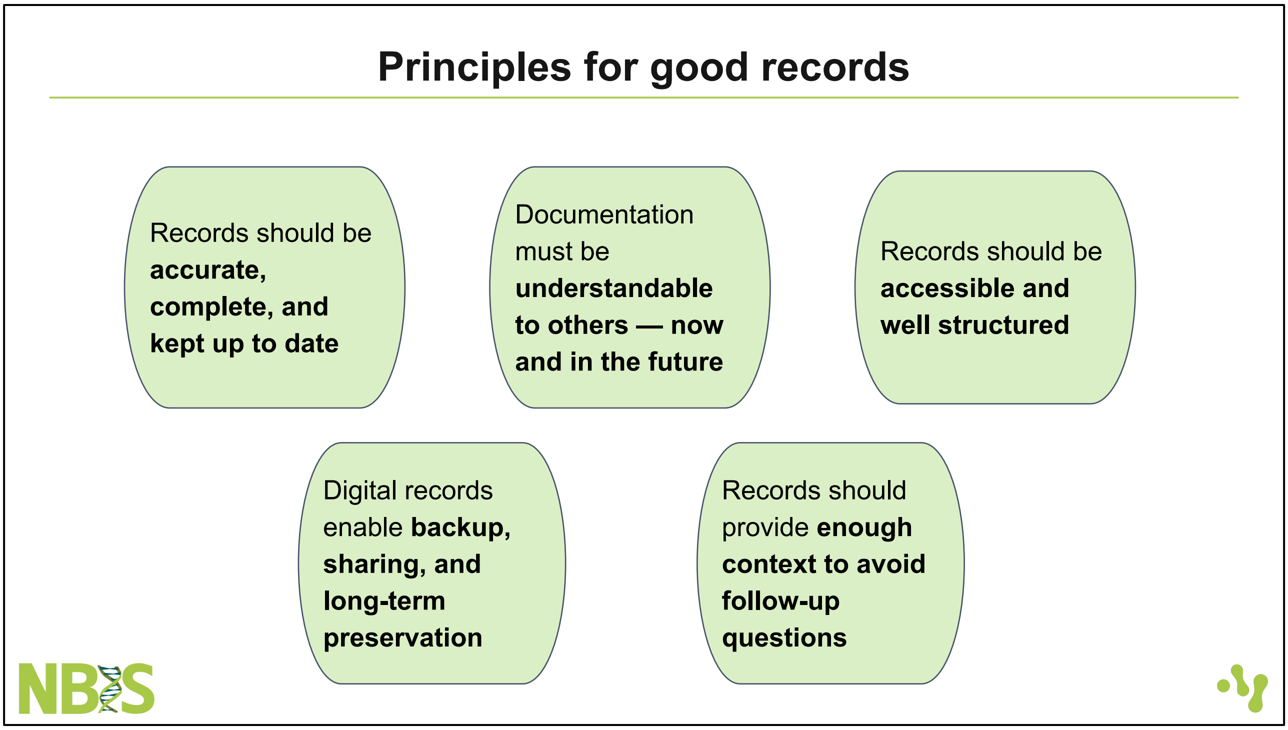
➔ **Data**
The actual measurements or files generated
(e.g. images, sequences, tables)

➔ **Metadata**
Information describing the data
(e.g. sample ID, instrument, date, settings)

➔ **Documentation**
Text explaining how data were generated and processed

➔ **Records**
The complete context: data, metadata, protocols, notes, and decisions

Good record-keeping keeps data, metadata, and documentation connected and interpretable over time.



Principles for good records

Records should be **accurate, complete, and kept up to date**

Documentation must be **understandable to others — now and in the future**

Records should be **accessible and well structured**

Digital records enable **backup, sharing, and long-term preservation**

Records should provide **enough context to avoid follow-up questions**

Good records allow someone else to understand what was done — without asking you.

Principles for good records

(*Protocols vs Lab notes*)

Contents in **protocols** can include

- Name, affiliation and contact information
- Originator of protocol (if not you)
- Information on ***why*** and ***how experiment*** was done
- Health and safety advice (and technical advice)
- Required software, materials and instruments
- Being self-explanatory
- Describe mistakes (for others to avoid repeating)
- Reference ethical application (if applicable)

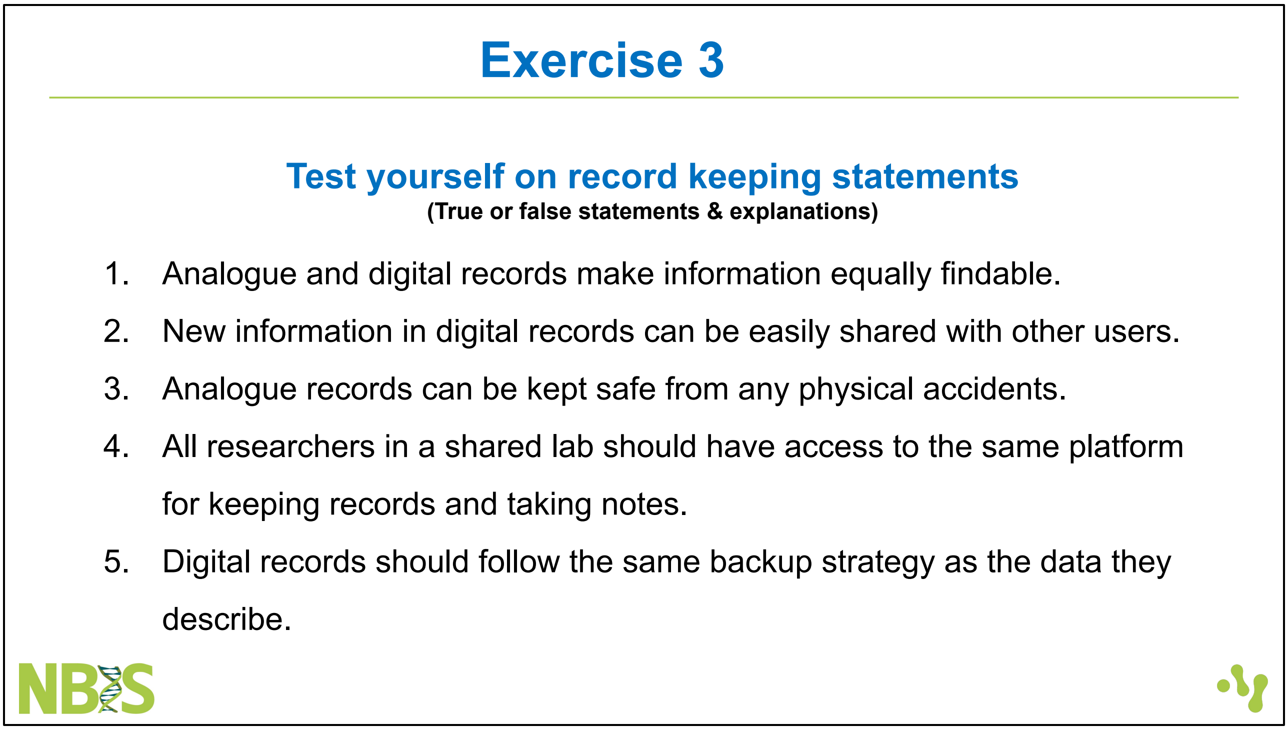
Protocols and lab notes should both be…

- **Detailed**
- **Up-to-date**
- **Accurate**
- **Easy to understand**

Contents in **lab notes** can include

- Name and affiliation
- Details on ***what, when*** and ***how***
- What project the experiment is part of
- Lot and batch numbers for consumables
- Information on ***metadata*** collected
- Interpretation of outcome and outlook/plans
- Post-outcome treatment of data

Electronic Lab Notebook Comparison Matrix

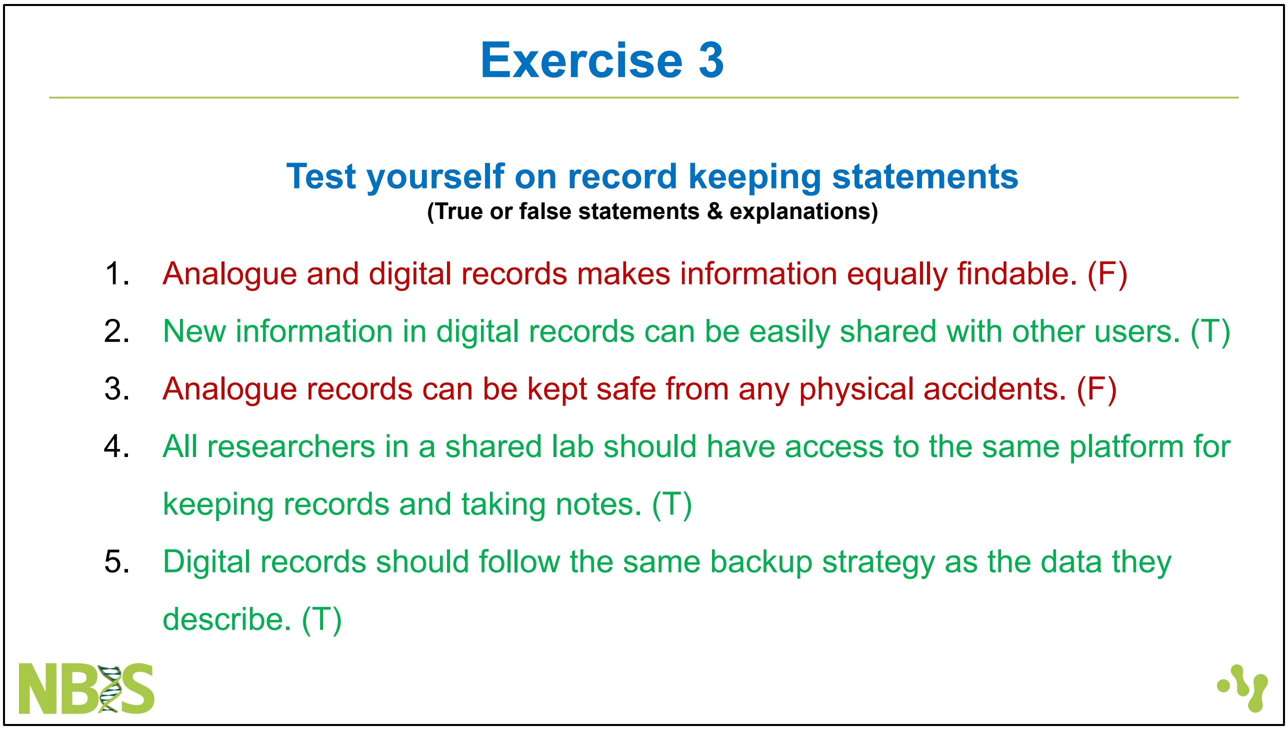


Exercise 3

Test yourself on record keeping statements

(True or false statements & explanations)

1. Analogue and digital records make information equally findable.
2. New information in digital records can be easily shared with other users.
3. Analogue records can be kept safe from any physical accidents.
4. All researchers in a shared lab should have access to the same platform for keeping records and taking notes.
5. Digital records should follow the same backup strategy as the data they describe.



Exercise 3

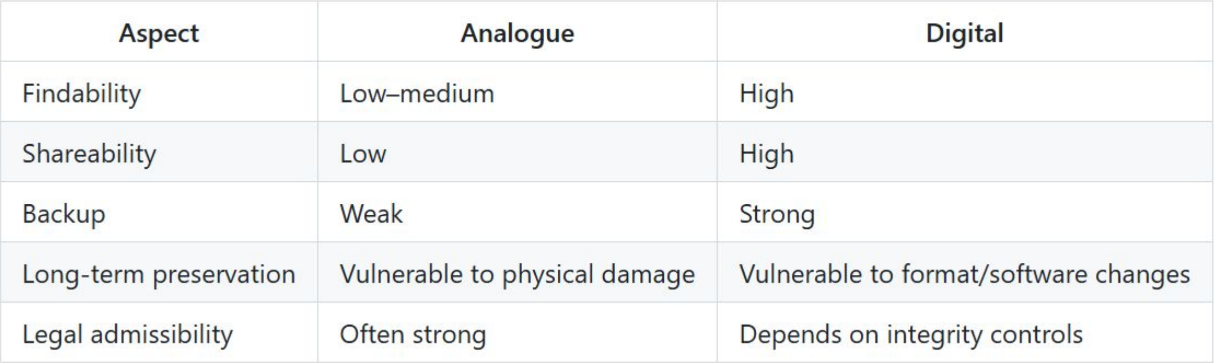
Test yourself on record keeping statements

(True or false statements & explanations)

1. Analogue and digital records makes information equally findable. (F)
2. New information in digital records can be easily shared with other users. (T)
3. Analogue records can be kept safe from any physical accidents. (F)
4. All researchers in a shared lab should have access to the same platform for keeping records and taking notes. (T)
5. Digital records should follow the same backup strategy as the data they describe. (T)

Analogue vs Digital records

Analogue vs. Digital Records – A Practical Comparison



Aspect	Analogue	Digital
Findability	Low–medium	High
Shareability	Low	High
Backup	Weak	Strong
Long-term preservation	Vulnerable to physical damage	Vulnerable to format/software changes
Legal admissibility	Often strong	Depends on integrity controls

Digital records reduce many risks but introduce new ones, such as file format obsolescence, software dependency, and access control.

Effective Record Keeping in Research

Choose the **Right Format**

digital or analogue

Organise **Systematically**

maintain consistent structure

Ensure **Accessibility**

use platforms or systems that are accessible

Enable **Sharing**

systems allow secure, controlled sharing

Follow **Backup** Best Practices

Prioritize **Security**

Backup

Data and hardware failure is always a threat.

Plan early (have a backup strategy) for potential failure!

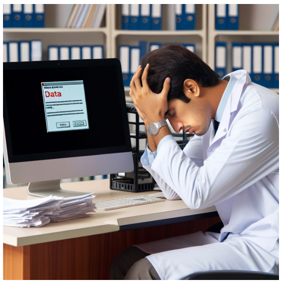


Image: generated by ChatOpen

Good to know for **backup** planning purposes:

- ➔ Data sensitivity
- ➔ Ease of access
- ➔ File sizes
- ➔ Overall data volumes
- ➔ Data life cycle in project

Backing up your research data is essential to ensure its **safety and longevity**. Research data represents countless hours of effort, and losing it due to **hardware failure, accidental deletion, or cyber threats** can be devastating. A robust backup strategy—using multiple locations, such as cloud storage, external drives, or institutional servers—provides a **safety net,** protecting your work from unforeseen events. By **regularly backing up your data,** you not only safeguard your progress but also maintain the **integrity and reproducibility** of your research.

Backup

Nearly all data, metadata and project information necessary to understand your analysis and results **require some sort of backup** strategy

Try to keep backup in **three separate copies,** on at least **two different kinds** of media (server, portable hard drive, cloud). Consider **off-site backup.** **3-2-1 redundancy**

Never backup your data on **portable drives only**, and particularly not on USB sticks!

Robust backups need to be **automated!**

Exercise 4

Discussion

Discuss the validity of the following statements on data backup:

1. I have my most important data backed up on my laptop. I have never experienced a hard drive failure, and my current laptop has a new state-of-the-art hard drive. Therefore, I don't need external backups.
2. All my data is stored in a cloud service.
3. My data is on a portable hard drive. There is a backup of the most important files on a shared USB stick in my research group.
4. My data is on a departmental backup administered by my University. Additionally, we have a server for all the data stored in our project.
5. We have no shared backup at all. All members in our research group are responsible for their own data.

Exercise 4

Discussion

Discuss in pairs the validity of the following statements on data backup:

1. Unsafe and not recommended. All hard drives can be subject to failure. In case of failure, all data will be lost.
2. Cloud services can be sufficient as backup, but are not fail safe. It can be sufficient in combination with a secondary backup on e.g. a shared server. For certain types of data (e.g. sensitive information), a cloud service may be outright inappropriate.
3. Not a good solution. Both portable hard drives as well as USB sticks are prone to failure.
4. A good solution in general. Data is stored independently in two separate systems. Centrally administered services are usually organised in such a way that partial failures does not affect the users.
5. Worst possible alternative. A disaster waiting to happen!

Backup

Creating a **backup strategy** in 10 steps

- Find out whether **your institution** has a backup strategy
- Determine **what** you want to back up
- Decide **how many backups** you will need and **how frequently** to backup
- Decide **where** backups will be **stored**
- Determine **how much storage capacity** will be needed
- Determine if there are tools you could use to **automate backup**
- Determine **how long** backups **will be kept** and how they will be destroyed
- Determine how **personal data** will be protected
- Devise a disaster **recovery plan**
- Assign **responsibilities**

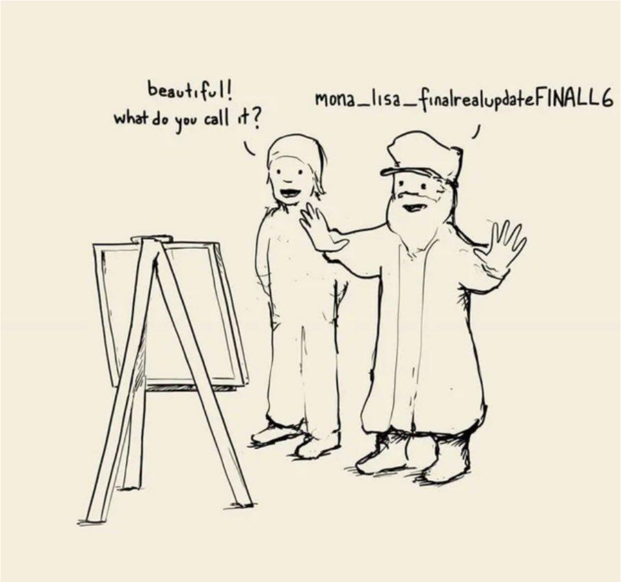
Typical storage options at Swedish universities

- Managed project storage / file service (backup + access control)
- Department/lab servers (routines vary—verify backup & restore)
- Secure platforms for sensitive / GDPR data (often separate environment)

Check: backup frequency • restore time/process • responsibility/ownership

Files and Folders

Why is file organisation important for data management?


beautiful! what do you call it?
mona_lisa_finalrealupdateFINALL6

https://twitter.com/nathanwpyle/status/1108902487203958784

What level of data organisation will work for me and my project/ team?

In the following slides, we are going to talk about some **good practices** for **organising files and folders**…We will look into practices for **classifying** and **structuring files** and folders to make them more useful.

Your guiding principle should be that someone unfamiliar with your project would be able to look at your files and understand, in detail, what you did and why. This someone could be

1) a researcher who wants to reproduce the results in your article,

2) a new collaborator who needs to understand the details of your experiments, or more commonly,

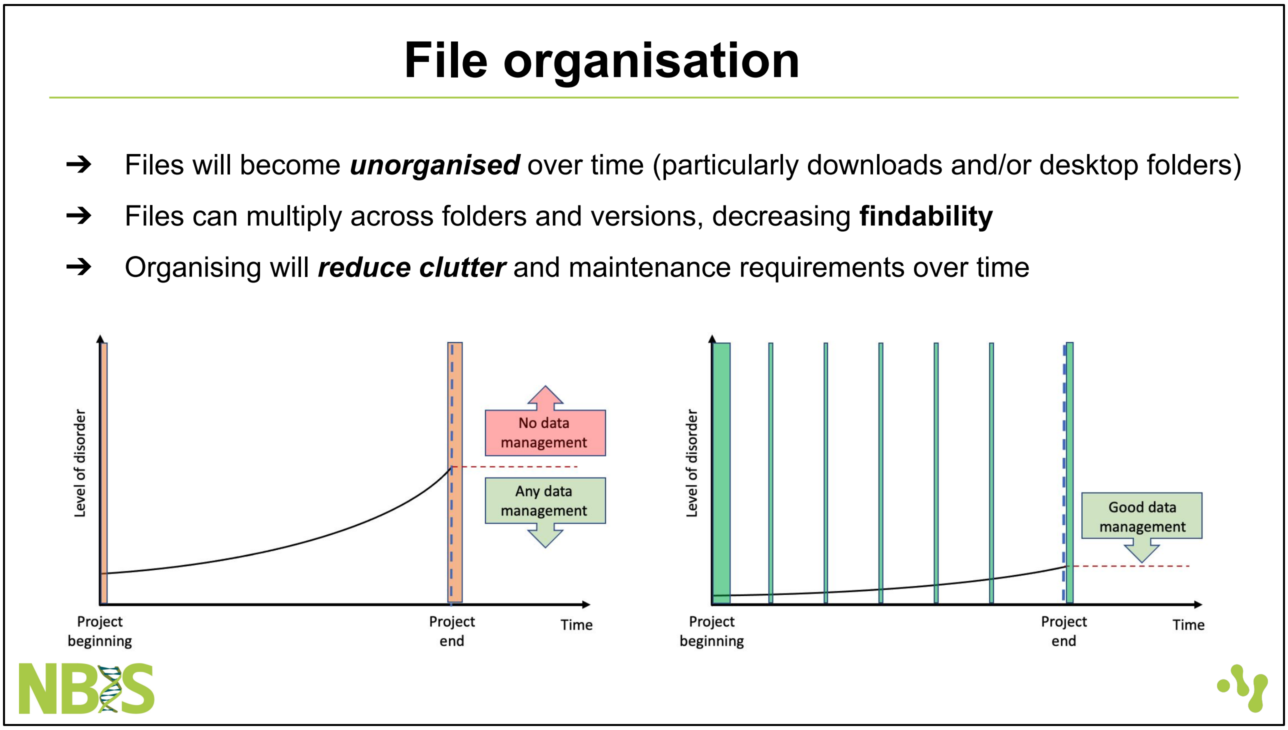
3) that someone could be your future self not remembering why you created a particular set of files.

Poor organisation practices can lead to significantly slower research progress and you may end up having to spend significant time re-orienting yourself among files and contents you once knew.

File organisation

Benefits of systematically organising research and data files in your project :

- Easier to **locate** a file
- Find **similar files** together
- Easy to **identify** which files you want to **back up**
- **Moving files** becomes much easier
- Increases **productivity**
- Useful to keep and **maintain a record** of the project
- Projects can **easily** be **understood** by others (including your future self)
- Keep **organised** in the long-run



File organisation

- ➔ Files will become ***unorganised*** over time (particularly downloads and/or desktop folders)
- ➔ Files can multiply across folders and versions, decreasing **findability**
- ➔ Organising will ***reduce clutter*** and maintenance requirements over time

It is natural for some of your files to become unorganised from time to time (e.g. downloads or desktop folder), and in those cases, there may be multiple copies and versions of files, making it challenging to find what you are looking for. You can avoid this clutter by organising your files ahead of time, and any system is better than none.

File and Folder naming

Names for files and folders should be *consistent* and *meaningful* to yourself and collaborators

Example: LD_phyA_off_t04_2020-08-12_norm.xlsx

Based on the name, the file could contain information about:

LD	Long day sampling, of the
phyA	Phytochrome A genotype, in a
off	Medium without sucrose, at
t04	Time point 4
2020-08-12	Sampled on Aug 12th, 2020, with
norm	Normalised data

File and Folder naming

Names for files and folders should be *consistent* and *meaningful* to yourself and collaborators

Example: LD_phyA_off_t04_2020-08-12_norm.xlsx

Based on the name, the file could contain information about:

LD	Long day sampling, of the

Not obvious from the letters and words alone!
Explanation is required!

2020-08-12	Sampled on Aug 12th, 2020, with
norm	Normalised data

Exercise 5

Group discussion

The following example contain files from an imaginary project

- ***phyA/phyB*** - genotypes
- ***sXX*** - sample number
- ***LD/SD*** - light conditions (Long Day, Short Day)
- ***on/off*** - different growth media (on sucrose, off sucrose)
- ***date format*** - sample date
- ***tXX*** - sample time point
- ***raw, norm*** - raw or normalised data

```
2020-07-14_s12_phyB_on_SD_t04.raw.xlsx
2020-07-14_s1_phyA_on_LD_t05.raw.xlsx
2020-07-14_s2_phyB_on_SD_t11.raw.xlsx
2020-08-12_s03_phyA_on_LD_t03.raw.xlsx
2020-08-12_s12_phyB_on_LD_t01.raw.xlsx
2020-08-13_s01_phyB_on_SD_t02.raw.xlsx
2020-7-12_s2_phyB_on_SD_t01.raw.xlsx
AUG-13_phyB_on_LD_s1_t11.raw.xlsx
JUL-31_phyB_on_LD_s1_t03.raw.xlsx
LD_phyA_off_t04_2020-08-12.norm.xlsx
LD_phyA_on_t04_2020-07-14.norm.xlsx
LD_phyB_off_t04_2020-08-12.norm.xlsx
LD_phyB_on_t04_2020-07-14.norm.xlsx
SD_phyB_off_t04_2020-08-13.norm.xlsx
SD_phyB_on_t04_2020-07-12.norm.xlsx
SD_phya_off_t04_2020-08-13.norm.xlsx
SD_phya_ons_t04_2020-07-12.norm.xlsx
ld_phyA_ons_t04_2020-08-12.norm.xlsx
```

Exercise 5

1. Should dates be put first, and if not, why?
2. What is the difference between using leading 0 (zero) and not?
3. Is there a difference between using upper and lower case letters?
4. What is the difference between using two letters for *on* compared to three letters *ons*?
5. What are the effects if we, as in the above example, mix naming conventions?

- ***phyA/phyB*** - genotypes
- ***sXX*** - sample number
- ***LD/SD*** - light conditions (Long Day, Short Day)
- ***on/off*** - different growth media (on sucrose, off sucrose)
- ***date format*** - sample date
- ***tXX*** - sample timepoint
- ***raw, norm*** - raw or normalised data

```
2020-07-14_s12_phyB_on_SD_t04.raw.xlsx
2020-07-14_s1_phyA_on_LD_t05.raw.xlsx
2020-07-14_s2_phyB_on_SD_t11.raw.xlsx
2020-08-12_s03_phyA_on_LD_t03.raw.xlsx
2020-08-12_s12_phyB_on_LD_t01.raw.xlsx
2020-08-13_s01_phyB_on_SD_t02.raw.xlsx
2020-7-12_s2_phyB_on_SD_t01.raw.xlsx
AUG-13_phyB_on_LD_s1_t11.raw.xlsx
JUL-31_phyB_on_LD_s1_t03.raw.xlsx
LD_phyA_off_t04_2020-08-12.norm.xlsx
LD_phyA_on_t04_2020-07-14.norm.xlsx
LD_phyB_off_t04_2020-08-12.norm.xlsx
LD_phyB_on_t04_2020-07-14.norm.xlsx
SD_phyB_off_t04_2020-08-13.norm.xlsx
SD_phyB_on_t04_2020-07-12.norm.xlsx
SD_phya_off_t04_2020-08-13.norm.xlsx
SD_phya_ons_t04_2020-07-12.norm.xlsx
ld_phyA_ons_t04_2020-08-12.norm.xlsx
```

Exercise 5

1. Should dates be put first, and if not, why?
2. What is the difference between using leading 0 (zero) and not?
3. Is there a difference between using upper and lower case letters?
4. What is the difference between using two letters for *on* compared to three letters *ons*?
5. What are the effects if we, as in the above example, mix naming conventions?

```
2020-07-14_s12_phyB_on_SD_t04.raw.xlsx
2020-07-14_s1_phyA_on_LD_t05.raw.xlsx
2020-07-14_s2_phyB_on_SD_t11.raw.xlsx
2020-08-12_s03_phyA_on_LD_t03.raw.xlsx
2020-08-12_s12_phyB_on_LD_t01.raw.xlsx
2020-08-13_s01_phyB_on_SD_t02.raw.xlsx
2020-7-12_s2_phyB_on_SD_t01.raw.xlsx
AUG-13_phyB_on_LD_s1_t11.raw.xlsx
JUL-31_phyB_on_LD_s1_t03.raw.xlsx
LD_phyA_off_t04_2020-08-12.norm.xlsx
LD_phyA_on_t04_2020-07-14.norm.xlsx
LD_phyB_off_t04_2020-08-12.norm.xlsx
LD_phyB_on_t04_2020-07-14.norm.xlsx
SD_phyB_off_t04_2020-08-13.norm.xlsx
SD_phyB_on_t04_2020-07-12.norm.xlsx
SD_phya_off_t04_2020-08-13.norm.xlsx
SD_phya_ons_t04_2020-07-12.norm.xlsx
ld_phyA_ons_t04_2020-08-12.norm.xlsx
```

1.Using dates as leading information in file names makes finding data quickly harder as the more interesting information may be samples or timepoints (unless date is crucial to data)

2. Without leading zeros, sorting will make 10 and 11 appear before 2

3. Upper and lower cases may sort differently

4. Comparing files is easier if the file name lengths are uniform

5. Mixed naming conventions can make it difficult to locate particular files, and/or sort a large number of files

File naming strategy

Two starting points for your file naming are:

A file name is a principal identifier of a file

- ➔ useful clues to the content
- ➔ status and version of a file
- ➔ help in classifying and sorting files
- ➔ facilitate searching and discovering files

File naming strategy should be consistent in time and among different people

- ➔ systematic and consistent across all files in the study
- ➔ a group of cooperating researchers should follow the same file naming strategy
- ➔ file names should be independent of the location of the file on a computer

File naming Do's

For dates use the **YYYY-MM-DD standard** at the end of the file UNLESS you need to organize your files chronologically

Version number (if applicable), use **leading zeros** (i.e.: v005 instead of v5)

Make sure the **end-letter file format extension** is present (e.g. .doc, .xls, .mov, .tif)

Add a **file in your top directory** which details your naming convention, directory structure and abbreviations

File naming Dont’s

Using spaces (use _ or - instead)

Long names

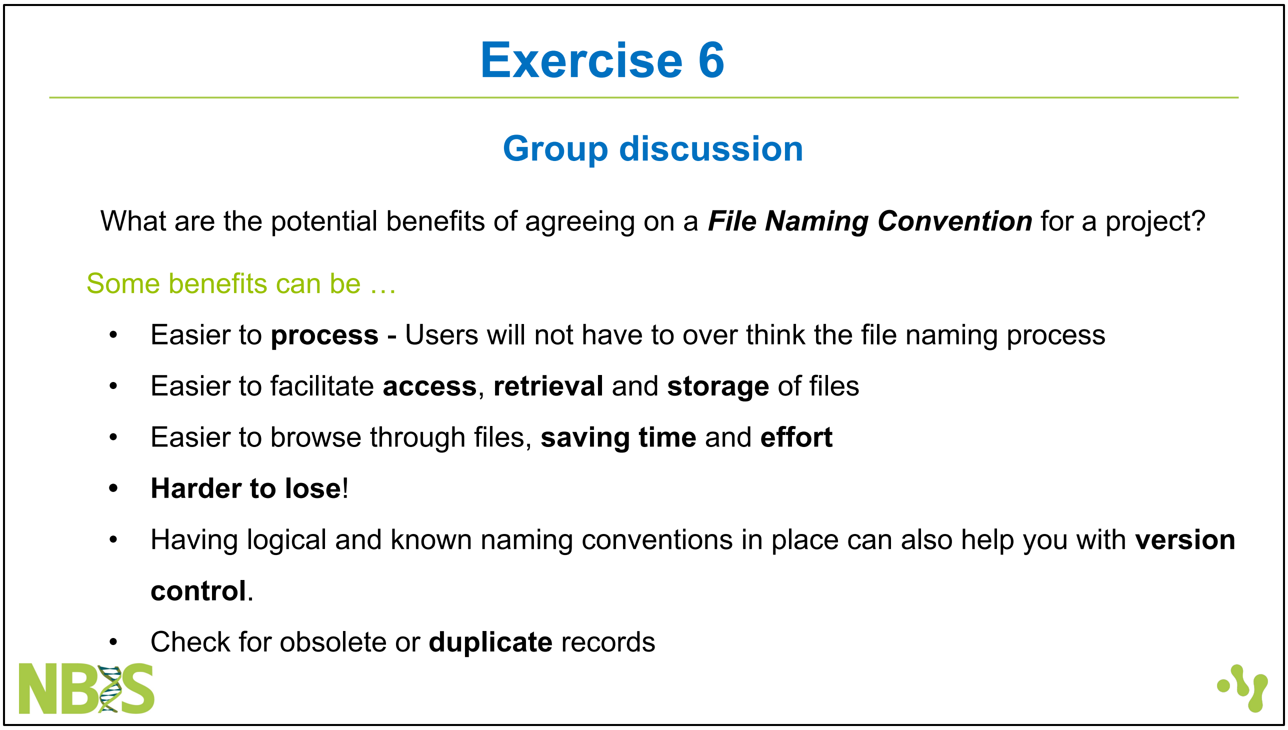
Repetition, e.g if directory name is Electron_Microscopy_Images, file ELN_MI_IMG1_S01_20200101.img then ELN_MI is redundant*

Dots, commas and special characters (e.g. ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “)

Using language specific characters (e.g óężé)

Deep paths with long names (i.e. deeply nested folders with long names), as archiving or moving between OS may fail

* applicable if the file is not going to be accessed independently of the location



Exercise 6

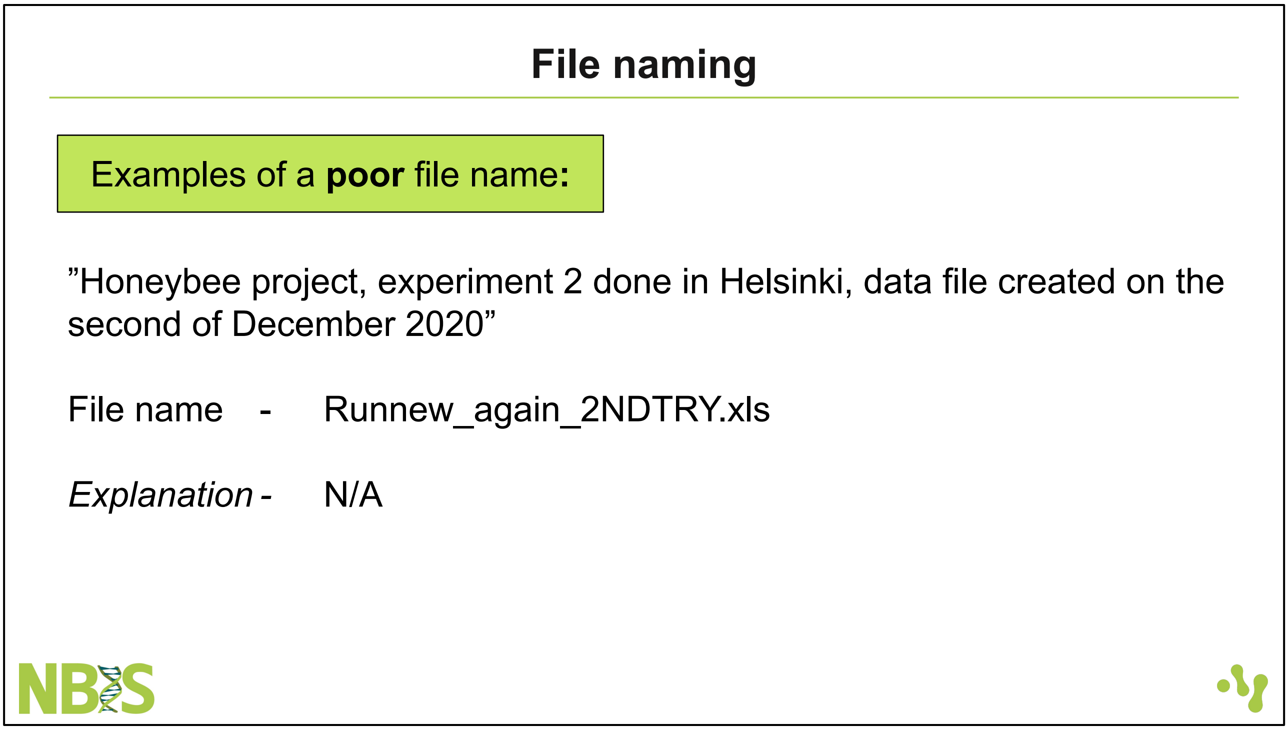
Group discussion

What are the potential benefits of agreeing on a ***File Naming Convention*** for a project?

Some benefits can be …

- Easier to **process -** Users will not have to over think the file naming process
- Easier to facilitate **access**, **retrieval** and **storage** of files
- Easier to browse through files, **saving time** and **effort**
- **Harder to lose**!
- Having logical and known naming conventions in place can also help you with **version control**.
- Check for obsolete or **duplicate** records

A File Naming Convention is a framework, or protocol if you like, for naming in a way that describes what the files and folders contain and, importantly, how they relate to each other.



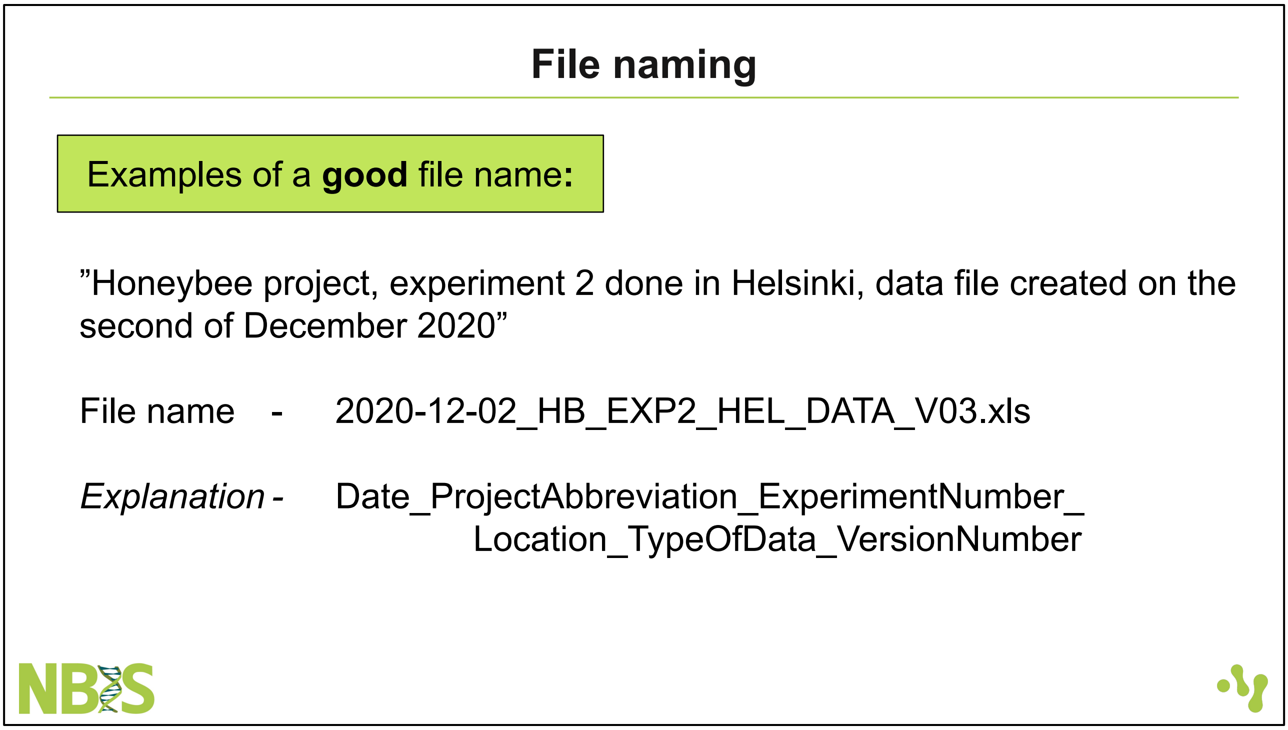
File naming

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 - 2020-12-02_HB_EXP2_HEL_DATA_V03.xls

Explanation - Date_ProjectAbbreviation_ExperimentNumber_Location_TypeOfData_VersionNumber

There are more examples of good and bad file names in the canvas module

File naming convention

Want to create your own File Naming Convention? Consider…

What **group of files** will this naming convention cover?

What **metadata is important** about these files and makes each file distinct?

Do you need to **abbreviate** any of the metadata or encode it?

What is the **order** for the **metadata** in the file name?

What **characters** will you use **to separate** each piece of metadata in the file name?

Will you need to **track** different **versions** of each file?

Write down your naming convention pattern

Document this convention in an introductory **file** and keep it with your files

Briney, Kristin A. (2020) File Naming Convention Worksheet. [Teaching Resource]

File versioning

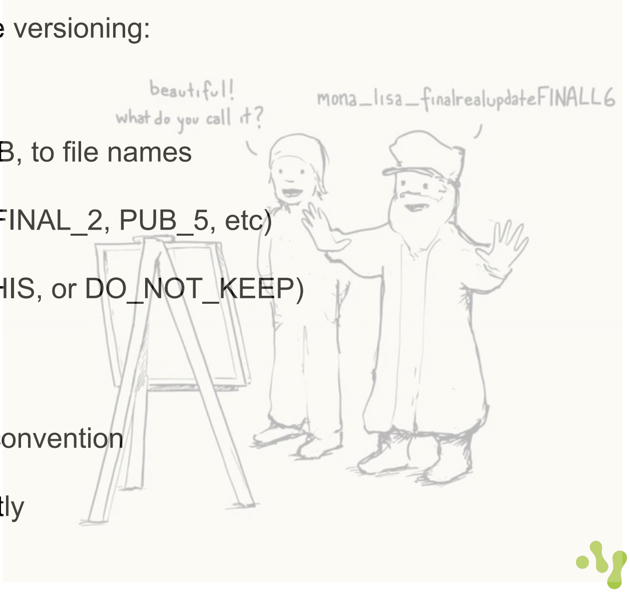
The simple yet powerful **Dont's** and **Do's** of file versioning:

Dont's

- Add suffixes like FINAL, THIS_ONE, or PUB, to file names
- Add numbers to already bad suffixes (e.g. FINAL_2, PUB_5, etc)
- Add negative information (e.g. DELETE_THIS, or DO_NOT_KEEP)

Do's

- Explicitly include versioning in file naming convention
- Use version numbers, preferably consistently

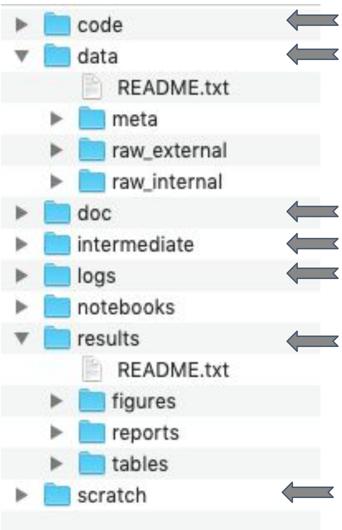


Organising files and folders — key principles

- **Plan early:** decide how files will be organised and found over time
- **Be consistent:** all collaborators use the same structure and naming
- **Keep projects separate:** one top-level folder per project
- **Use clear names:** short, descriptive folder names (project + ID + date)
- **Choose a sensible hierarchy:** shallow or deep, depending on the project
- **Group meaningfully:** organise by data type, experiment, method, or time point
- **Design for growth:** start simple and expand as the project evolves

A good structure supports collaboration, reuse, and long-term access.

Directory structure for a sample project



```
▶ code
▼ data
      README.txt
   ▶ meta
   ▶ raw_external
   ▶ raw_internal
▶ doc
▶ intermediate
▶ logs
▶ notebooks
▼ results
      README.txt
   ▶ figures
   ▶ reports
   ▶ tables
▶ scratch
```

- code ⟵ all code needed to go from input files to final results
- data ⟵ raw and primary data, essentially all input files, **never** edit!
- doc ⟵ documentation for the study
- intermediate ⟵ output files from different analysis steps, *can be deleted*
- logs ⟵ logs from the different analysis steps
- results ⟵ output from workflows and analyses
- scratch ⟵ temporary files that can be safely *deleted or lost*

Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424.
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000424

Break

Thank you for your attention and participation!

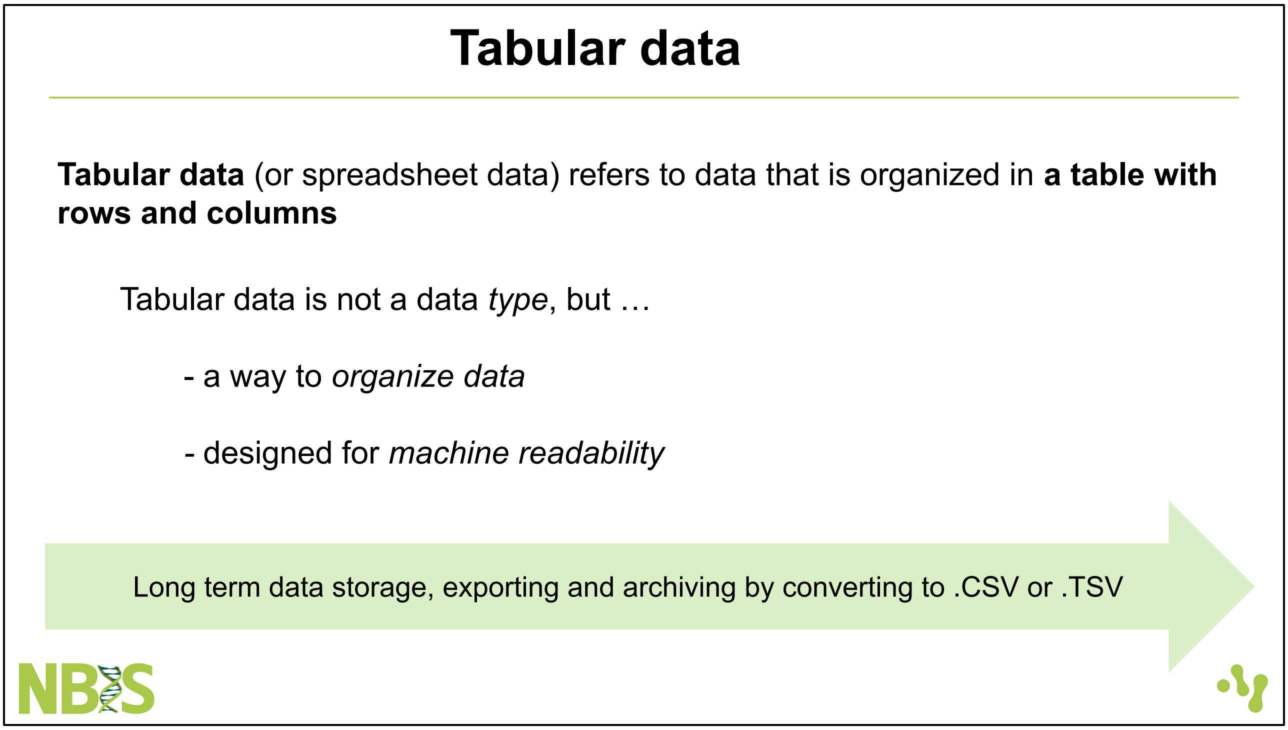
Caffeine is a stimulant

Exercise 7

We are back in the Famous lab!

- Considering the very limited metadata we have access to, and the inherited files, what can we do in order to increase the level and quality of data organization?
- Download the zip-file containing the inherited data structure
- Consider the following:
  - File names
  - Folder structure
  - Documentation

★ Work in pairs or in smaller groups
★ Focus on the discussion more than finishing the exercise
★ Consider your own data and files from a third-person-view



Tabular data

Tabular data (or spreadsheet data) refers to data that is organized in **a table with rows and columns**

Tabular data is not a data *type*, but …

- a way to *organize data*
- designed for *machine readability*

Long term data storage, exporting and archiving by converting to .CSV or .TSV

We tend to organize data in spreadsheets as we humans want to work with the data, but computers and humans see it in different ways. In order to use tools that **make computation** more efficient, we need to structure our data the way that computers need it.
CSV- comma separated values
TSV- tab separated values

Tabular data

Good practice for structuring tabular data is to…

Think about **how** to **organize** your data both from a data entry and data analysis point of view from the beginning

Adopt good **metadata standards** and **column header formats** early in the data collection phase

Setting up **well-formatted tables** early in the research process is extremely important – before you even start entering data

Separate r**aw data** from the **analysed data** can have different layout/format

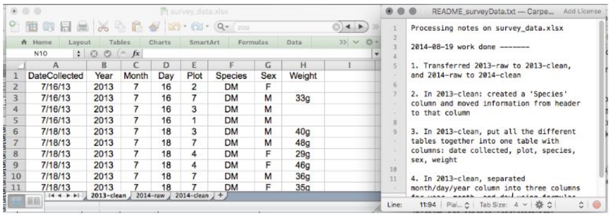
Leave the original (raw) data raw!!!

Keeping track of your analyses

When working with spreadsheets during data clean up or analyses you **must:**

- **…create a new file or tab with your cleaned or analyzed data.**
  Do not modify the original dataset, or you will never know where you started!

- **…keep track of the steps you took in your clean up or analysis.**
  You can do this in another text file, or a good option is to create a new tab in your spreadsheet with your notes.



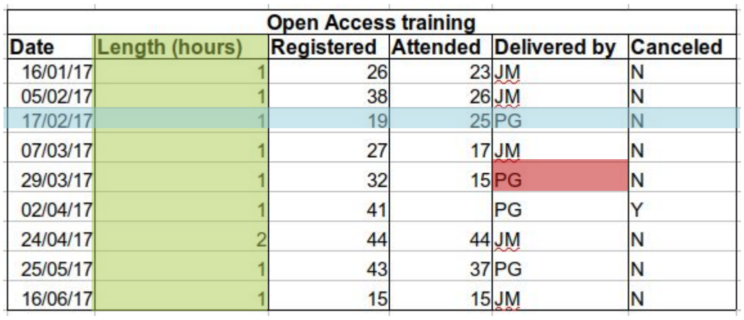
	A	B	C	D	E	F	G	H	I
1	DateCollected	Year	Month	Day	Plot	Species	Sex	Weight	
2	7/16/13	2013	7	16	2	DM	F		
3	7/16/13	2013	7	16	7	DM	M	33g	
4	7/16/13	2013	7	16	3	DM	M		
5	7/16/13	2013	7	16	1	DM	M		
6	7/18/13	2013	7	18	3	DM	M	40g	
7	7/18/13	2013	7	18	7	DM	M	48g	
8	7/18/13	2013	7	18	4	DM	F	29g	
9	7/18/13	2013	7	18	4	DM	F	46g	
10	7/18/13	2013	7	18	7	DM	M	36g	
11	7/18/13	2013	7	18	7	DM	F	35g	

```
Processing notes on survey_data.xlsx

2014-08-19 work done -------

1. Transferred 2013-raw to 2013-clean,
and 2014-raw to 2014-clean

2. In 2013-clean: created a 'Species'
column and moved information from header
to that column

3. In 2013-clean, put all the different
tables together into one table with
columns: date collected, plot, species,
sex, weight

4. In 2013-clean, separated
month/day/year column into three columns
for year, month, and day using formulas
```

How to structure data tables

The cardinal rules of using spreadsheet programs for 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

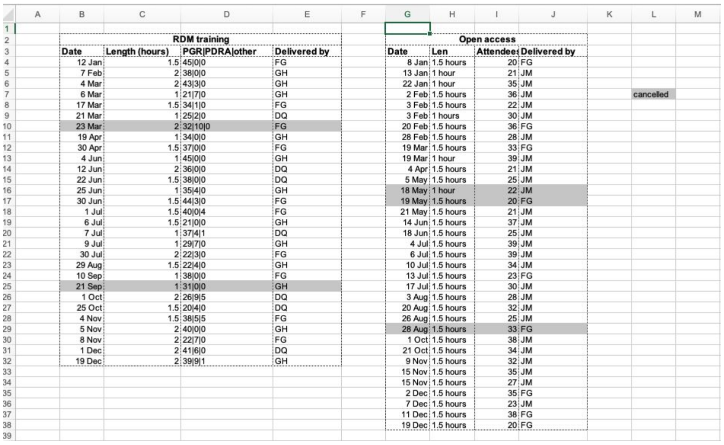
Tidy data tables

- One cell–one value
- One column–one variable
- One row–one observation

Tabular data

Do not:

- ✖ create multiple data tables within one spreadsheet tab
- ✖ combine values in cells
- ✖ merging cells
- ✖ use colors
- ✖ write comments in cells
- ✖ mix metadata and data
- ✖ use special characters
- ✖ use different date formats



	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													

Missing data

Zero vs. Missing data - How do you make explicit something that do not exist?

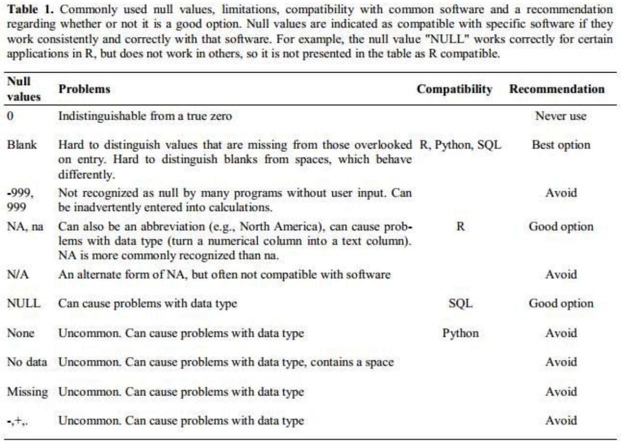
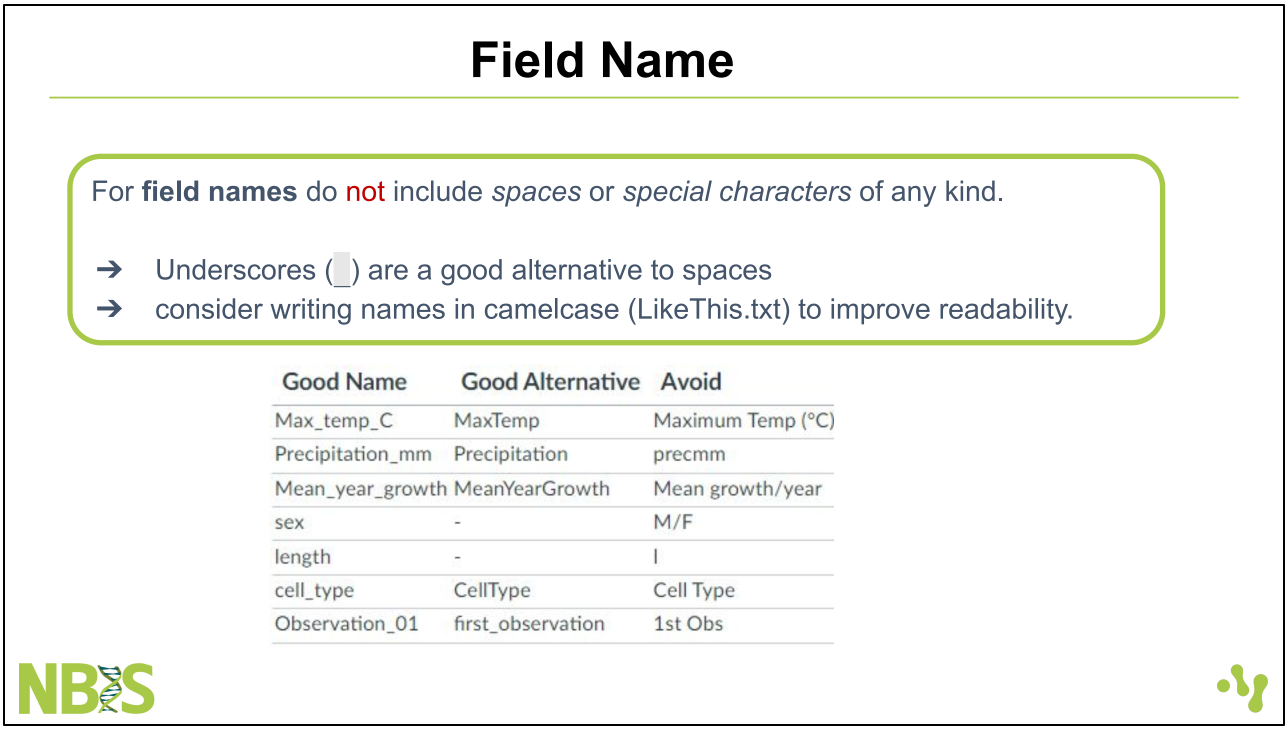


Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

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

White et al, 2013, Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution

To a computer, a zero is data. You measured or counted it, and it was zero. A blank cell means no measurement at all, and the computer will interpret it as a null value. Leaving zero data blank is not ideal in a written format, but it is NEVER acceptable when you move your data into a digital format.



Field Name

For **field names** do not include *spaces* or *special characters* of any kind.

- ➔ Underscores (_) are a good alternative to spaces
- ➔ consider writing names in camelcase (LikeThis.txt) to improve readability.

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	-	M/F
length	-	l
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

NBIS

If possible, decide on a pre-defined controlled vocabulary prior to collecting your data. Doing so will make later data publications much easier since your data is pre-adapted to the submission requirements.

Wrap-up: Organising Research Data

3 key takeaways

1. **Good organisation saves time and reduces risk**
   Clear structure, consistent file names, and basic documentation make your data reusable — by you and others.

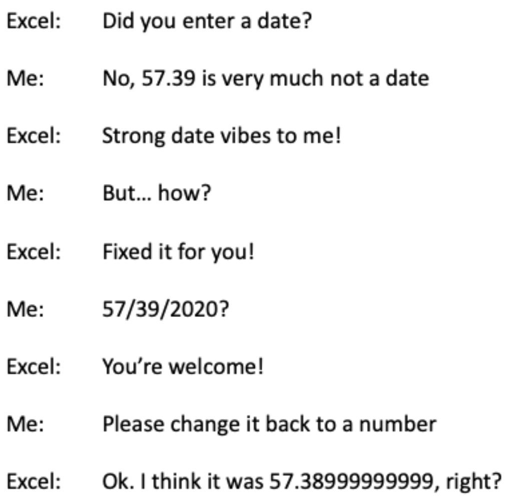
2. **Organisation is part of research quality, not admin**
   Well-organised data supports reproducibility, collaboration, and FAIR research practices.

3. **Small habits make a big difference**
   Simple rules for folders, file names, versions, and tables prevent problems before they appear.

What to do on Monday morning

- Check **where your project data is stored** and how it is backed up
- Create or clean up a **project folder structure**
- Agree on **file naming and versioning rules** in your group
- Add a short **README file** explaining what the data are and how to use them

Exercise 8



Excel: Did you enter a date?

Me: No, 57.39 is very much not a date

Excel: Strong date vibes to me!

Me: But… how?

Excel: Fixed it for you!

Me: 57/39/2020?

Excel: You’re welcome!

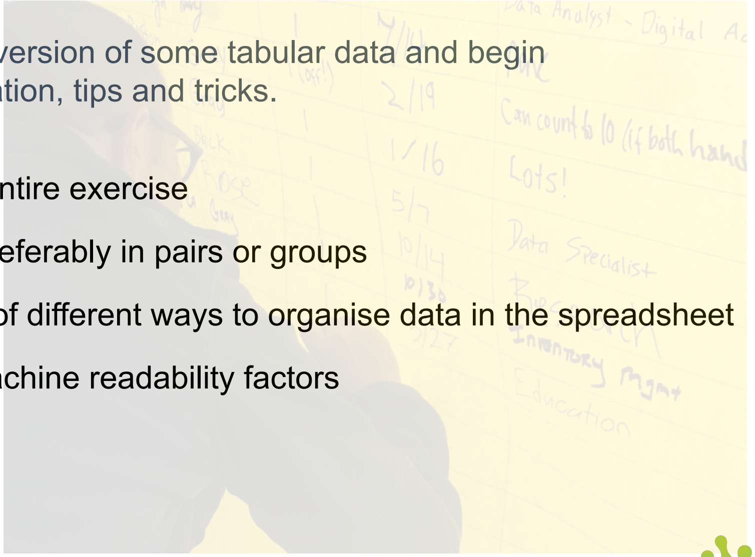
Me: Please change it back to a number

Excel: Ok. I think it was 57.38999999999, right?

Exercise 8

We are going to take a messy version of some tabular data and begin cleaning it up using the information, tips and tricks.

- Not important to finish the entire exercise
- Work at your own speed, preferably in pairs or groups
- Discuss the pros and cons of different ways to organise data in the spreadsheet
- Consider the Human vs. Machine readability factors



Break

Thank you for your attention and participation!