Research Data Management in the life sciences

Stephan Nylinder & Elin Kronander
NBIS / SciLifeLab - ELIXIR Sweden
data-management@scilifelab.se

Data Pipeline - The Scientist Perspective

Year 0	Year 1	Year 2	Year 3
			
The ideas!	Data collection!	Results!	Paper out!

Finished!

Oh, I need to publish the data also? Ok.

I can do that!

The Data Pipeline - DM Perspective

What samples did you use,
and how are they described?
(Sample metadata)

What machines did you
run your analysis on?
(Technical metadata)

Which files to publish, and
where are they?

The Data Pipeline - Scientist perspective

Year 0

Sample metadata? What is it? Where is it saved? Did I even save it at all?
(3 years ago)

Year 1

Technical metadata? What is that, and who do I ask?
(2 years ago)

Year 2

Year 3

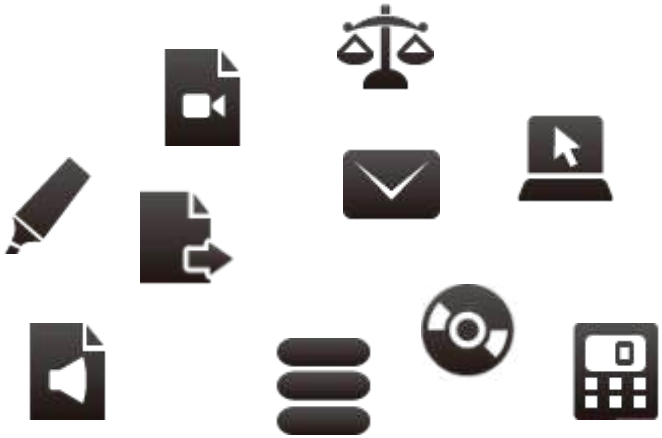
The data files? They are ALL here! At least I think so…
(1-2 years ago)

Data Management 101

Let the journey begin...

(Imagine Star Wars theme
and animation here)

Research data management



- Raw data
- Processed data
- Data about data (metadata)
- …

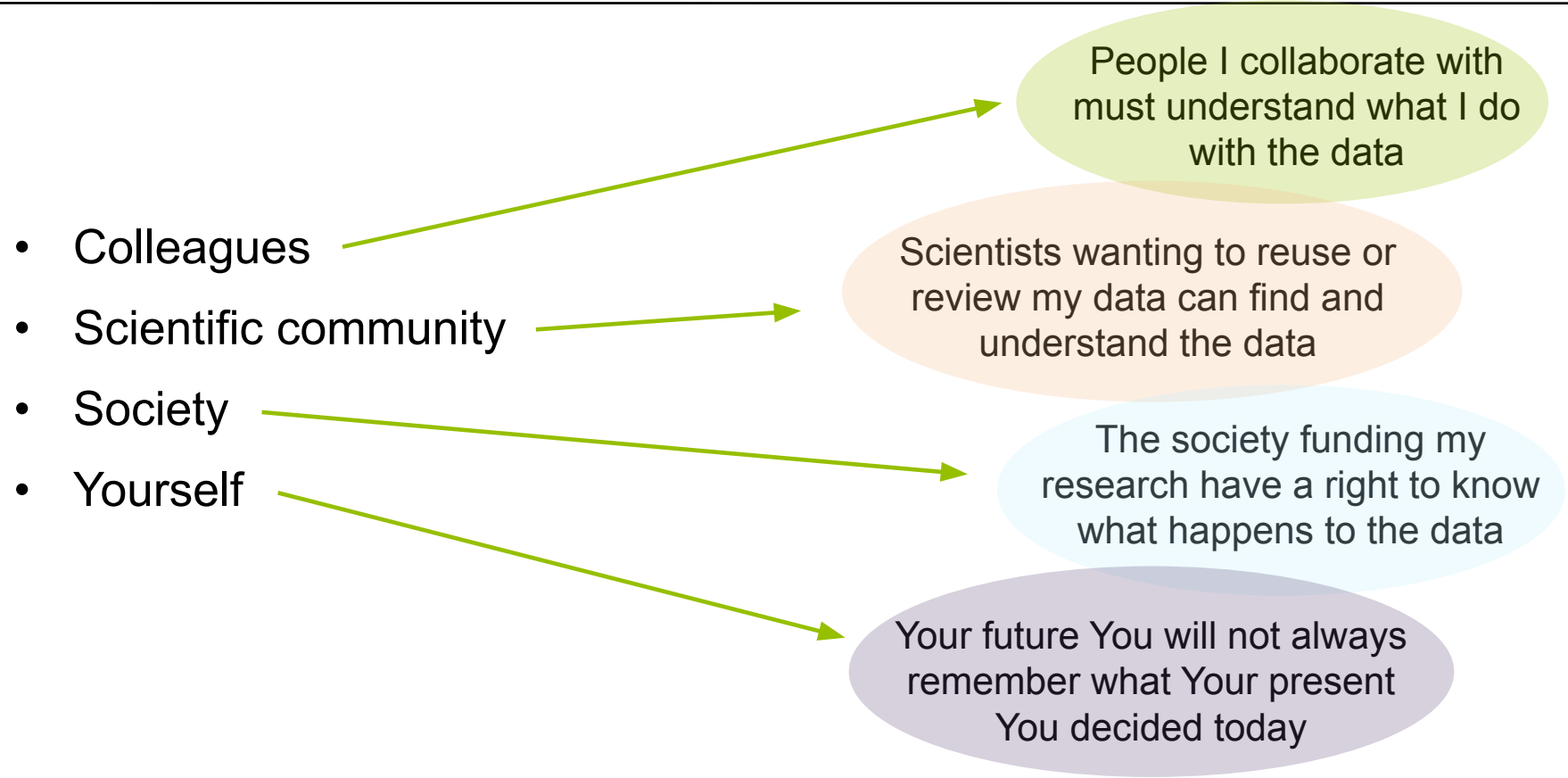
Note! Not all data is digital!

A FAIR data lifecycle

Good data management practices in all phases of research

- Research documentation
- Data organisation
- Information security
- Ethics and legislation

Data Management Recipients



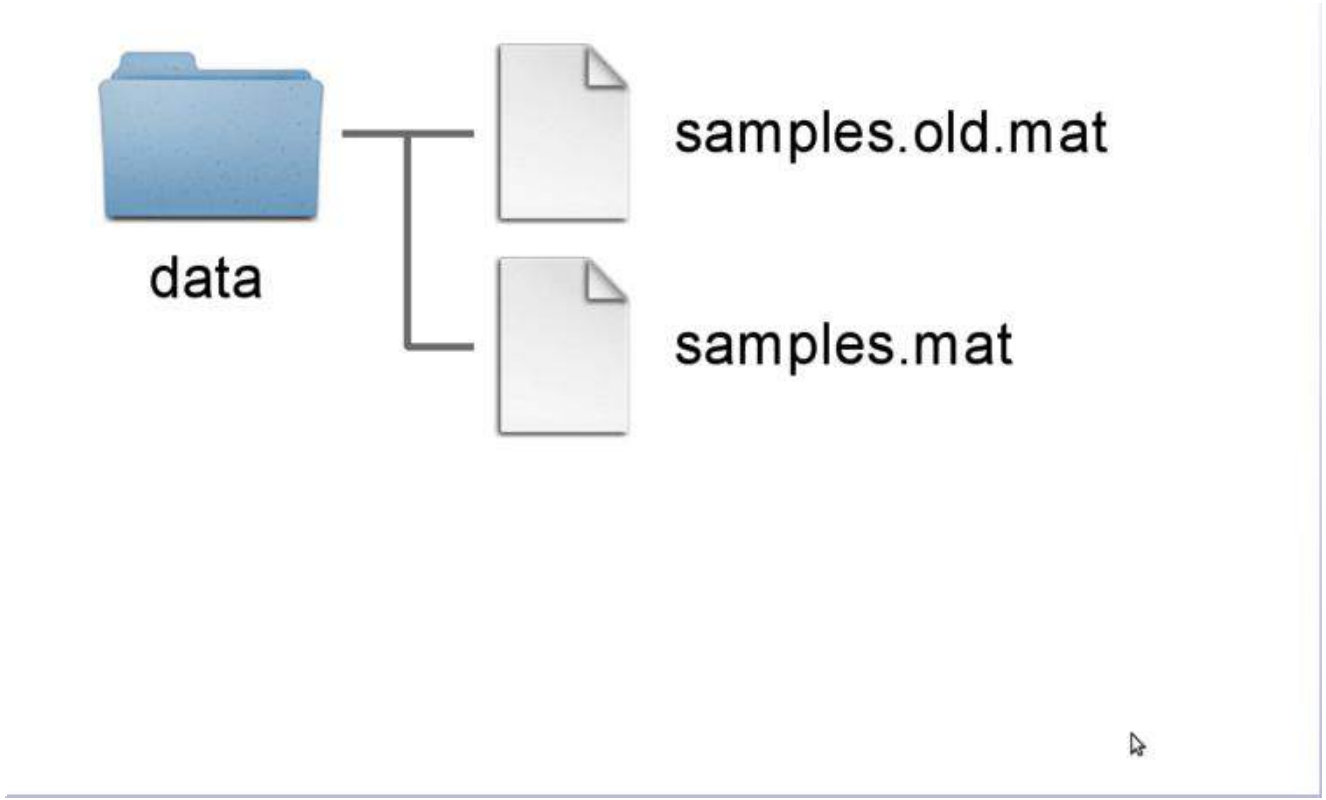
"Your primary collaborator is yourself six months from now, and your past self doesn't answer e-mails,"

-Rachael Ainsworth

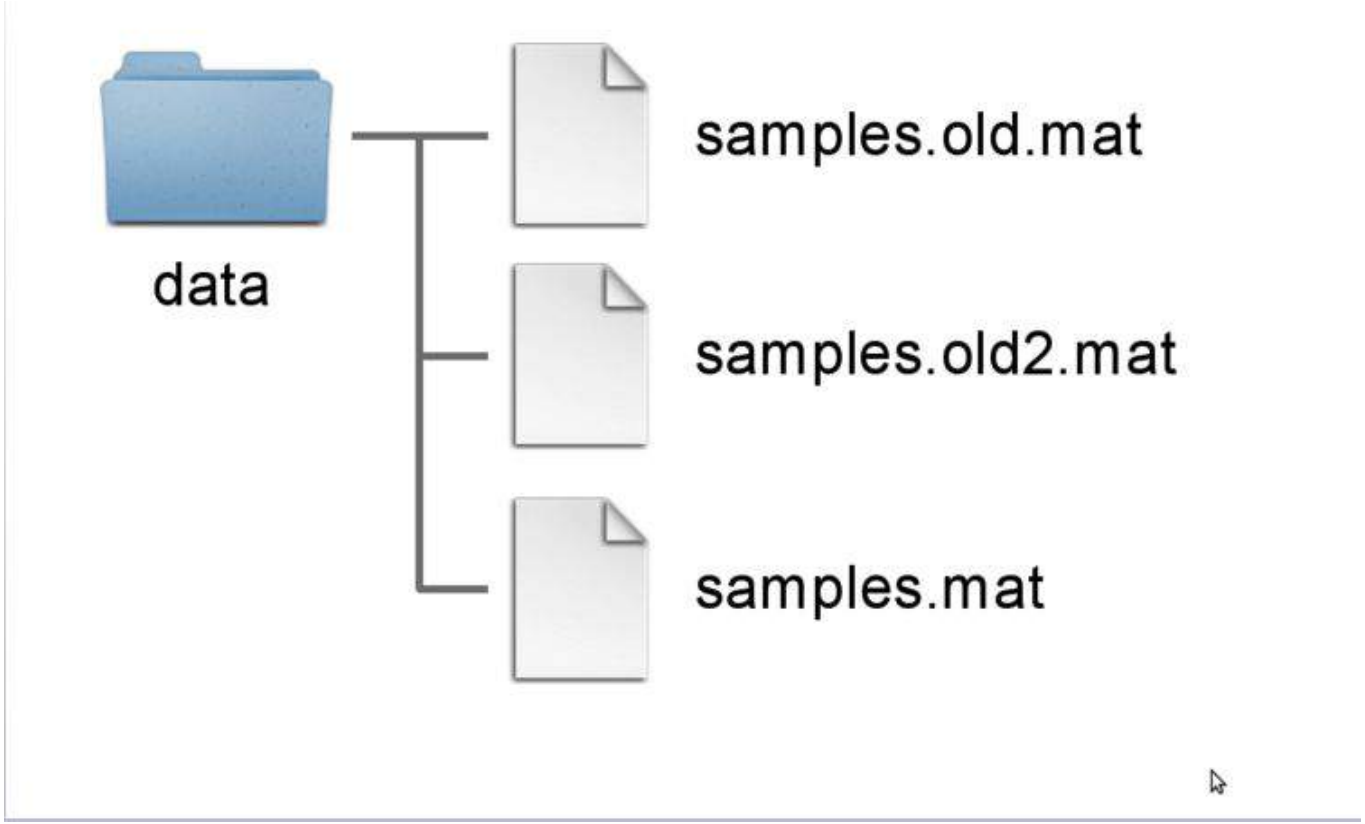
How do you know how an old result was generated?

First step - Organization

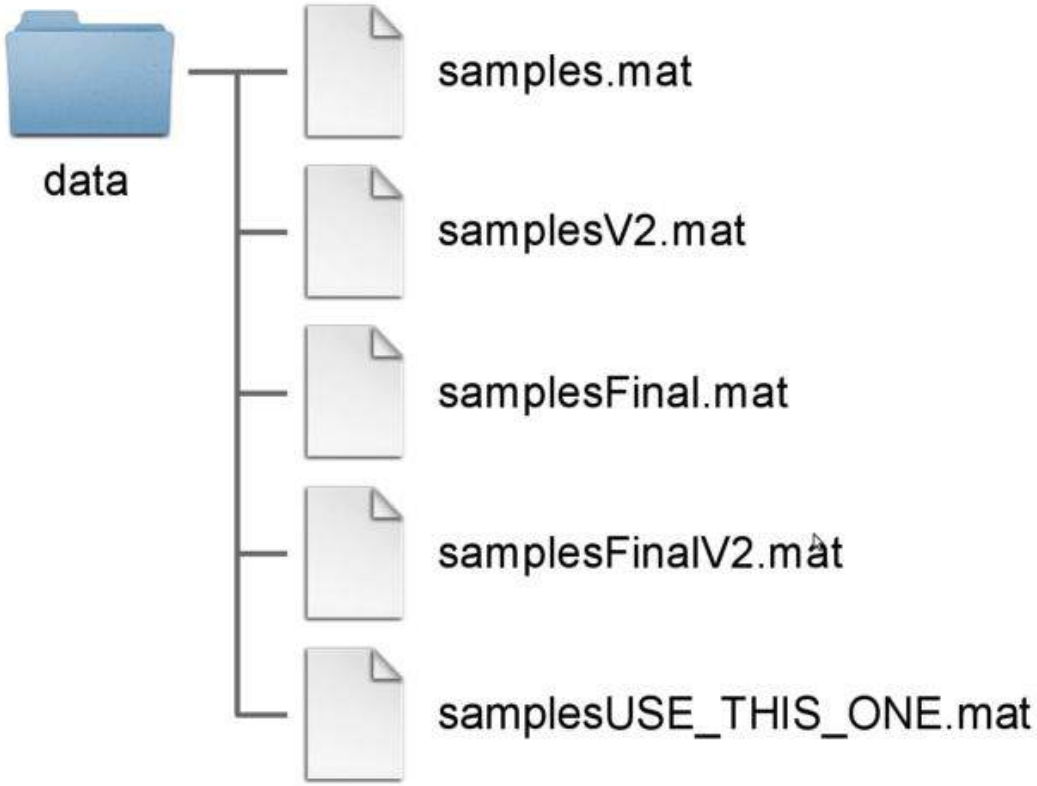
I guess this is alright



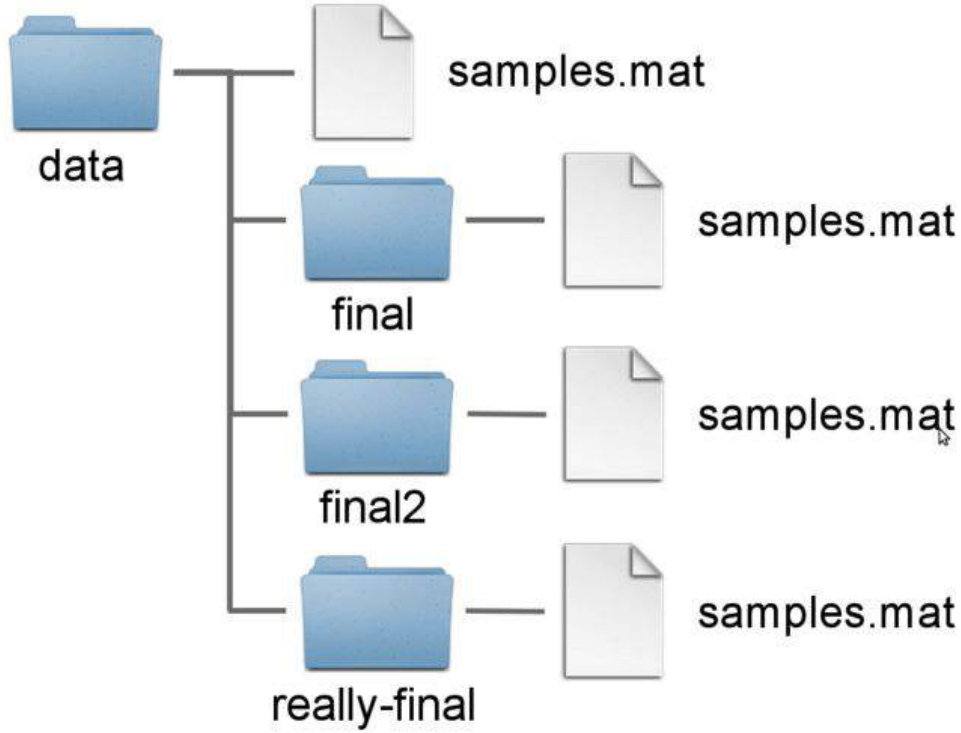
I guess this is alright



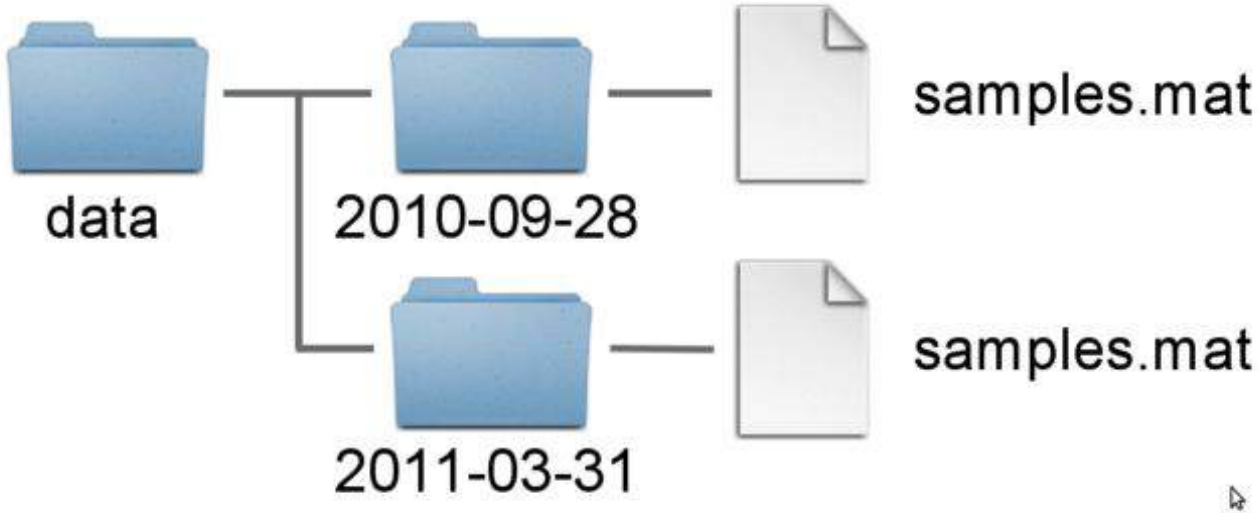
Which one is the most recent?



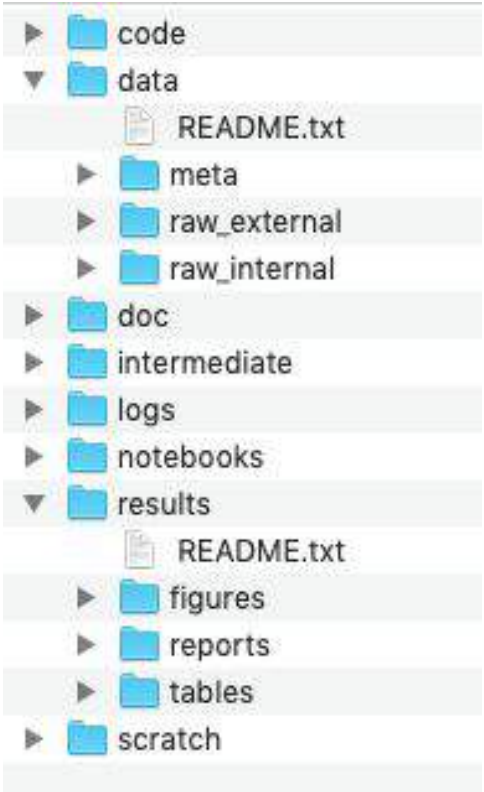
Another (bad) common approach



A possible solution



Directory structure for a sample project



all code needed to go from input files to final results

raw and primary data, essentially all input files, **never** edit!

documentation for the study

output files from different analysis steps, *can be deleted*

logs from the different analysis steps

output from workflows and analyses

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

Suggested best practices - file organisation

- There is a **folder for the raw data**, which do not get altered, or intermixed with data that is the result of manual or programmatic manipulation. I.e., derived data is kept separate from raw data, and **raw data are not duplicated**.
- **Code is kept separate from data**.
- Use a **version control system** (at least for code) – e.g. **git**
- There is a **scratch directory for experimentation**. Everything in the scratch directory can be deleted at any time without negative impact.
- There should be a **README** **in every directory**, describing the purpose of the directory and its contents.
- Use **file naming conventions** that makes it easy to find files and understand what they are (for humans and machines) and **document them**
- Use **non-proprietary formats** – *.csv* rather than *.xlsx*
- Etc…

File naming strategy

Two starting points for your file naming strategy are:

- A file name is a principal identifier of a file
- File naming strategy should be consistent in time and among different people

Principles for naming files:

1. Consider file name lengths – beware of operating system limitations and full path names!
2. Make names human readable – name describes content of file
3. Make names machine readable – Avoid spaces, punctations, accented characters etc.
4. Explain file naming strategy in associated README files (stored in the same location)

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_2.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.csv

Explanation - Time_ProjectAbbreviation_ExperimentNumber_ Location_TypeOfData_VersionNumber

File naming example

Names for files and folders should be *consistent* and *meaningful to yourself and collaborators*, allow for *easy tracking/searching*, and be *somewhat descriptive of content*.

Example: LD_phyA_off_t04_2020-08-12_norm.csv

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

But! Not obvious from the letters and words alone. Explanation is required - README!

File naming Dont's

- Using spaces (use _ or - instead)
- Dots, commas and special characters
  (e.g. ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ‘ “)
- Using language specific characters (e.g óężé), unfortunately they still cause problems with most software or between operating systems (OS)
- Long names incl. file headers (>256 characters)
- Consider repetitions in file names, e.g if directory name is Electron_Microscopy_Images, and file ELN_MI_IMG_20200101.img then ELN_MI_IMG is redundant
- Deep paths with long names (i.e. deeply nested folders with long names), as archiving or moving between operating systems may fail

File naming Do's

- For dates use the YYYY-MM-DD standard and place at the end of the file UNLESS you need to organize your files chronologically
- Include version number (if applicable), use leading zeros (i.e.: v005 instead of v5).
- make sure the end-letter file format extension is present at the end of the name (e.g. .txt, .md, .csv, .FASTQ)
- Add a README file in your top directory which details your naming convention, directory structure and abbreviations

File naming strategy

Discussion

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

File naming strategy

Discussion

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

- Easier to process - Team members 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

File naming

Names for files and folders should be *consistent* and *meaningful to yourself and collaborators*, allow for *easy tracking/searching*, and be *somewhat descriptive of content*.

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

Good file name - 20201202_HB_EXP2_HEL_DATA_V03.csv

Explanation - Time_ProjectAbbreviation_ExperimentNumber_ Location_TypeOfData_VersionNumber.txt

Explanation is required - README

README

A file usually defined as the starting point of information about something (attracts attention!)

Using them as documentation files for:

Folder level	–	Explaining folder contents, naming, file history, organisation/structure etc
Data	–	Explaining file names and contents

README - Example 1

Dataset Title: Raw Images for Experiment A, Famous Lab

Principal Investigator: Very Famous, PI, v.famous@best.university.ever

File Naming Convention: ExperimentName_InstrumentID_CaptureDateTime_ImageID.tif
The base file name is composed of the name of the experiment, the ID number of the instrument used, the date and time that the image was captured, and the unique identifier of the image.

Attributes: Also see the Codes section for a list of instruments and their ID numbers

ExperimentName	=	Name of the experiment
Instrument ID	=	Five-digit code assigned to the lab instrument
CaptureDateTime	=	Date and time at which the image set was captured, in YYYYMMDD format
Image ID	=	Three-digit unique identifier for image, such as 001, 002, 003

Codes: [List of instruments and IDs]

Examples:

File formats: daf2-age1_14052_20240412T0515_005.tif
Versioning: All changes to this dataset will be documented in a changelog in this README file

(Modified from: https://datamanagement.hms.harvard.edu/collect-analyze/documentation-metadata/readme-files)

README - Example 2

README.txt — Edited

```
README for the Honey Bee project field measurements data folder
This folder contains raw data collected manually from field measurements over several time
points.
file naming convention:
Time_ProjectAbbreviation_ExperimentNumber_Location_TypeOfData_VersionNumber
For example
20201202_HB_EXP2_HEL_DATA_V01.csv
20201202_HB_EXP2_HEL_DATA_V03.csv
20201202_HB_EXP2_HEL_DESCR_V03.csv

Time - is the date at the start of experiment YYYY-MM-DD
ProjectAbbreviation - is HB for Honey Bee
ExperimentNumber - EXP1, EXP2, EXP3 or EXP4
Location - refers to a city, HEL for Helsinki, STO for Stockholm or OSL for Oslo
TypeOfData - DATA for numeric measurements, DESCR for qualitative values
VersionNumber - Version number is increased each time point of data collection as V01, V02
and so on.
```

README

Discussion

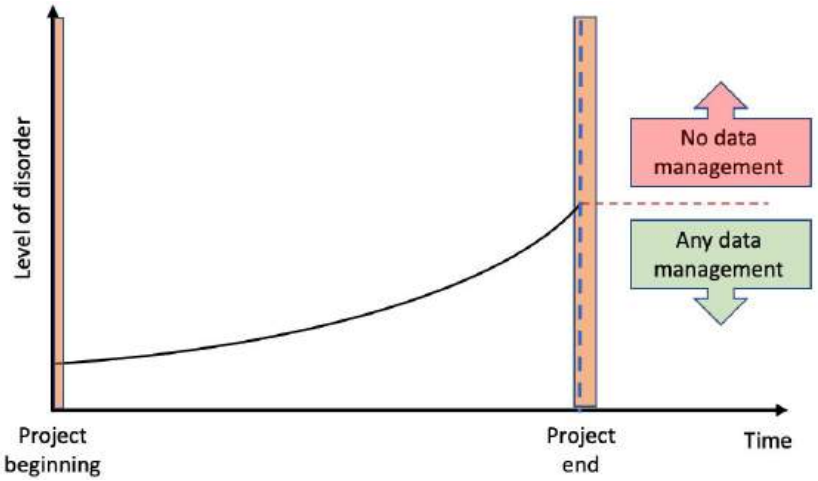
Can you think of an example where you would have benefited from having access to a README-file when working with data? Describe to the group what you would have wanted such a file to contain.

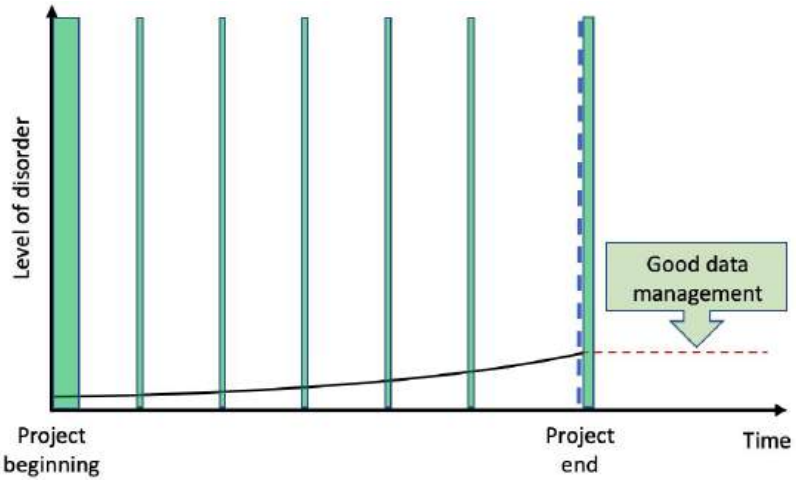
File organisation

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





The Data Pipeline - DM perspective

Year 0

The ideas!

Year 1

Data collection!

Year 2

Results!

Year 3

Paper out!

Sample metadata! Technical metadata!

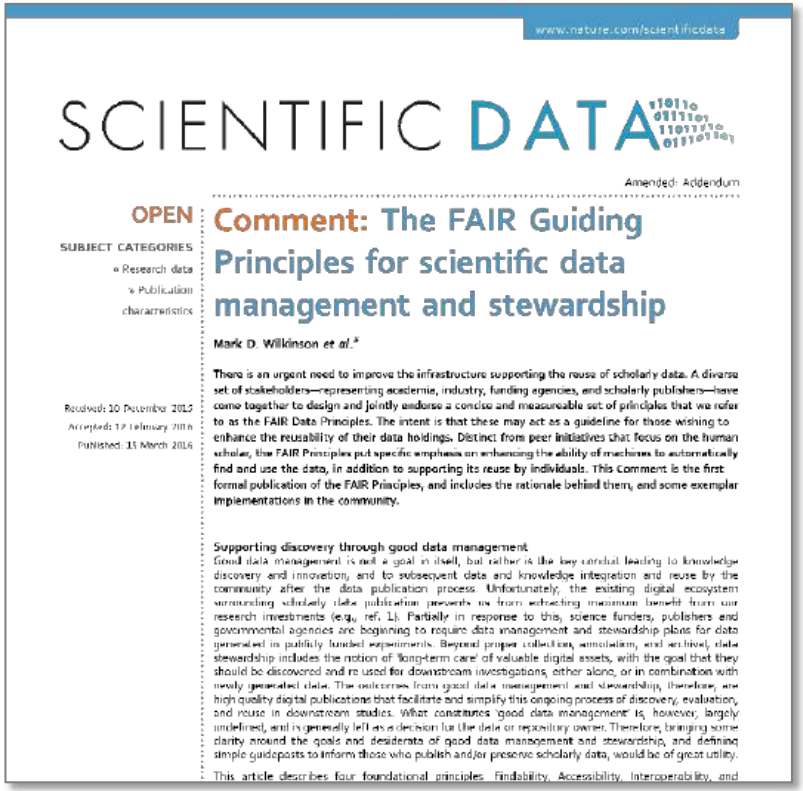
Publish early!
(with embargo if necessary)

RDM - Good practices

- **Secure/organise data & analyses**, by using folder structures, file naming conventions and README files, managing back-ups, access restrictions, versioning, docs, scripts and transcripts
- **Deposit and share data** using restricted or public access data repositories that promote FAIR data principles
- **Adhere to community standards**, such as file formats, data dictionaries, controlled vocabularies and metadata
- **Maintain a Data Management Plan,** outlining the project's data management practices

The FAIR principles

- Promote **efficient data discovery and reuse** by providing guidelines to make digital resources
  - ❑ **F**indable
  - ❑ **A**ccessible
  - ❑ **I**nteroperable
  - ❑ **R**eusable
- Address aspects **enabling software and infrastructure** to automatically find and use research data



www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
• Research data
• Publication characteristics

Received: 10 December 2015
Accepted: 17 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[*]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

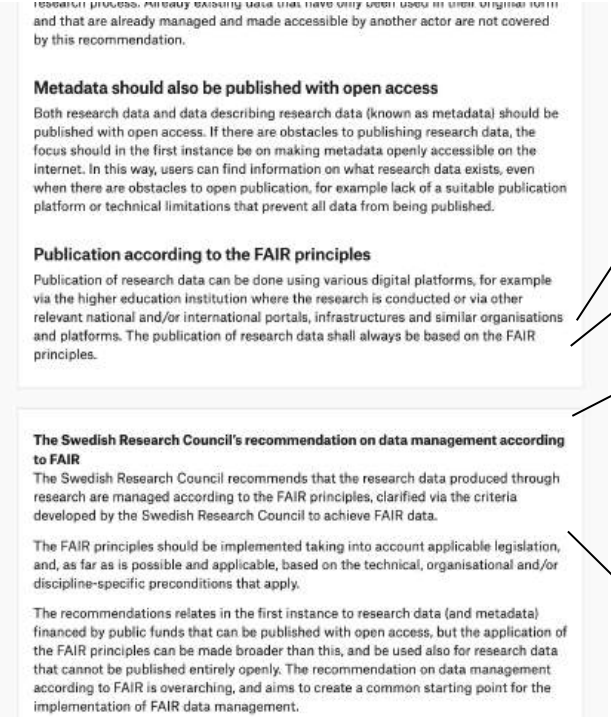
Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from achieving maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and

Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. doi:10.1038/sdata.2016.18

VR - Swedish Research Council

Vision: As open as possible, as closed as necessary by 2026



research process. Already existing data that have only been used in their original form and that are already managed and made accessible by another actor are not covered by this recommendation.

Metadata should also be published with open access

Both research data and data describing research data (known as metadata) should be published with open access. If there are obstacles to publishing research data, the focus should in the first instance be on making metadata openly accessible on the internet. In this way, users can find information on what research data exists, even when there are obstacles to open publication, for example lack of a suitable publication platform or technical limitations that prevent all data from being published.

Publication according to the FAIR principles

Publication of research data can be done using various digital platforms, for example via the higher education institution where the research is conducted or via other relevant national and/or international portals, infrastructures and similar organisations and platforms. The publication of research data shall always be based on the FAIR principles.

The Swedish Research Council's recommendation on data management according to FAIR

The Swedish Research Council recommends that the research data produced through research are managed according to the FAIR principles, clarified via the criteria developed by the Swedish Research Council to achieve FAIR data.

The FAIR principles should be implemented taking into account applicable legislation, and, as far as is possible and applicable, based on the technical, organisational and/or discipline-specific preconditions that apply.

The recommendations relates in the first instance to research data (and metadata) financed by public funds that can be published with open access, but the application of the FAIR principles can be made broader than this, and be used also for research data that cannot be published entirely openly. The recommendation on data management according to FAIR is overarching, and aims to create a common starting point for the implementation of FAIR data management.

*[...] The publication of research data **shall always be based on the FAIR principles**.[...]*

The Swedish Research Council's recommendation on data management according to FAIR

*The Swedish Research Council recommends that the research data produced through research are managed according to the FAIR principles, clarified via the **criteria developed by the Swedish Research Council to achieve FAIR data**. [...]*

https://www.vr.se/english/mandates/open-science/open-access-to-research-data/the-swedish-research-councils-recommendation.html

A FAIR data lifecycle

Data Life Cycle by RDMkit used under CC-BY

- **FAIR data ≠ Open data**
  Data can be Open without being FAIR
  Data can be FAIR without being open
  "As open as possible, as closed as necessary"
- **FAIR software/FAIR training materials**
- **Data can be more or less FAIR**

How to get started?

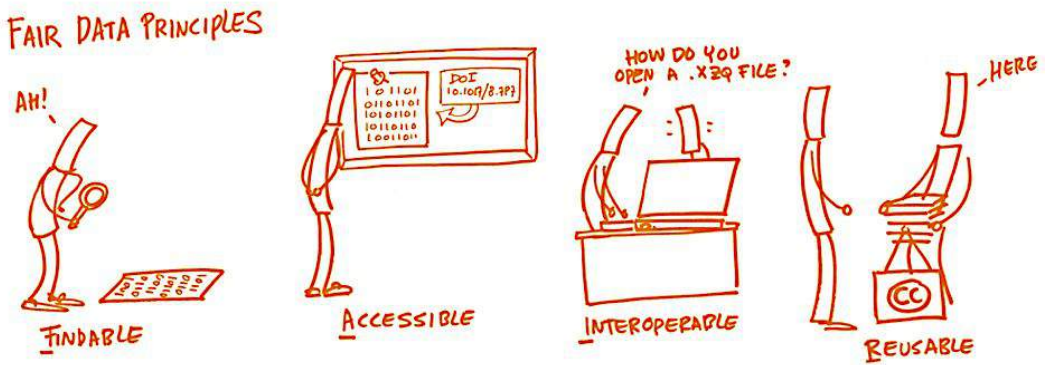


Image: https://book.fosteropenscience.eu/

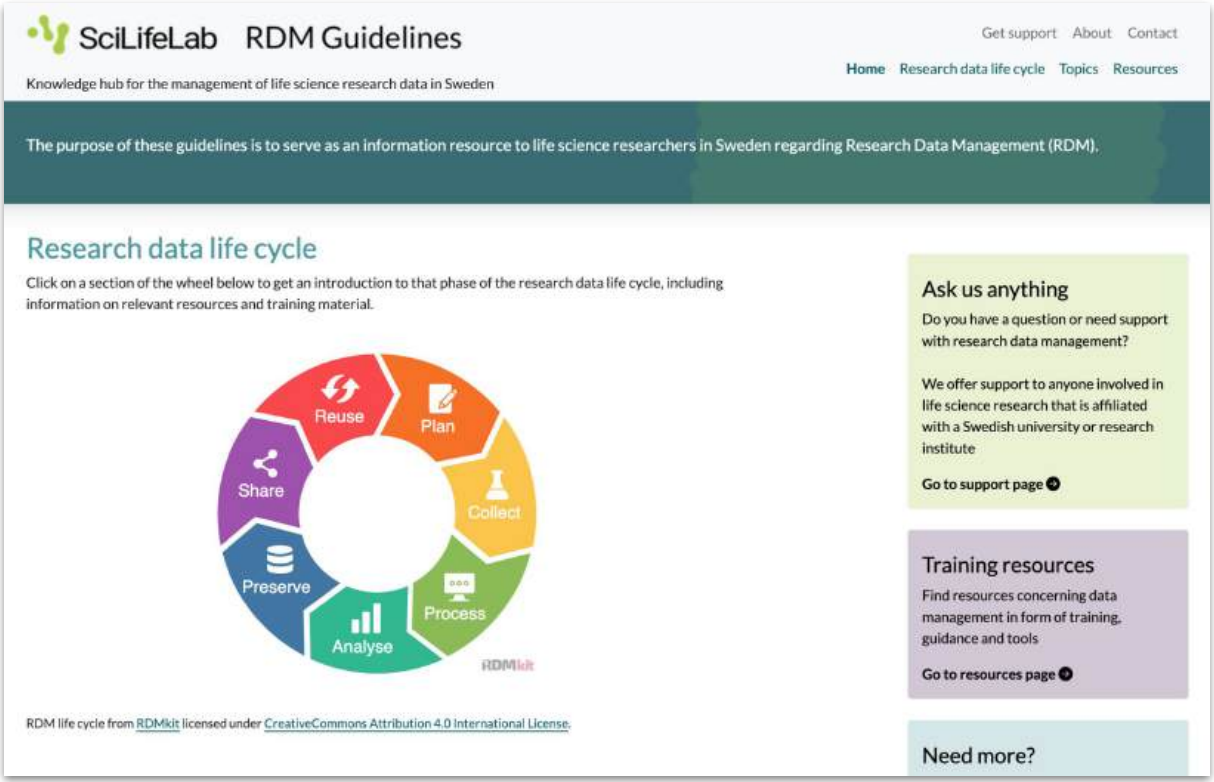
FAIRify by README - Adopting good practices for data organization, makes research data more FAIR

FAIRify by planning - Thinking ahead and continuously document your strategies in a Data Management Plan using a guiding tool https://dsw.scilifelab.se/

Deposit data in a repository (early!)

Get support from data stewards

SciLifeLab RDM Guidelines



https://data-guidelines.scilifelab.se/

data-management@scilifelab.se

NBIS Data management services

- Guide writing a data management plan
- Identify a suitable repository for publishing your data
- Assist during the submission process when publishing your data and code
- Advice on what needs to be done when working with sensitive human data
- Advice on describing data with proper metadata for documentation and publishing
- Data transfers, data organisation, backup, and security procedures

Contact us

- data-guidelines.scilifelab.se
- data-management@scilifelab.se

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

Thank you!