

Data Management in the life sciences

What is FAIR, and why should you care?

Elin Kronander & Niclas Jareborg
NBIS / SciLifeLab - ELIXIR Sweden
data@nbis.se

The SciLifeLab bioinformatics platform



~100 staff at six different sites across Sweden with expertise in many different omics-related areas



Core activities

'To help the Swedish research community build knowledge in analyzing large and complex omics data and to make bioinformatics easily accessible for life science researchers'



Support

Support services ranging from short consultations to long-term embedded bioinformaticians.



Infrastructure

Providing infrastructure in the form of services, computational resources, tools and guidelines to the life science community.



Training

Training events in advanced and applied bioinformatics.

Data management

Data management plans Data publishing FAIR data
Sensitive data - EGA helpdesk Good DM practice Data handling
Courses Guidelines Templates & Tools



Niclas

- Data stewards support researchers
- Collaboration with SciLifeLab Data Centre (supports platforms)

Team vision

“Swedish Life Science researchers **apply good Data Management practices** so that the **research outputs** produced are **available** to the global research community, and to society at large, according to the principles of *Open Science, Reproducible Research, and FAIR*”

Team mission

Support – Training – Collaboration



Niclas Jareborg
Data Manager



Markus Englund
Data Steward



Yvonne Kallberg
Data Steward



Elin Kronander
Data Steward



Stephan Nylinder
Data Steward



Maja Pelve
Data Steward



Mattias Strömberg
Data Steward

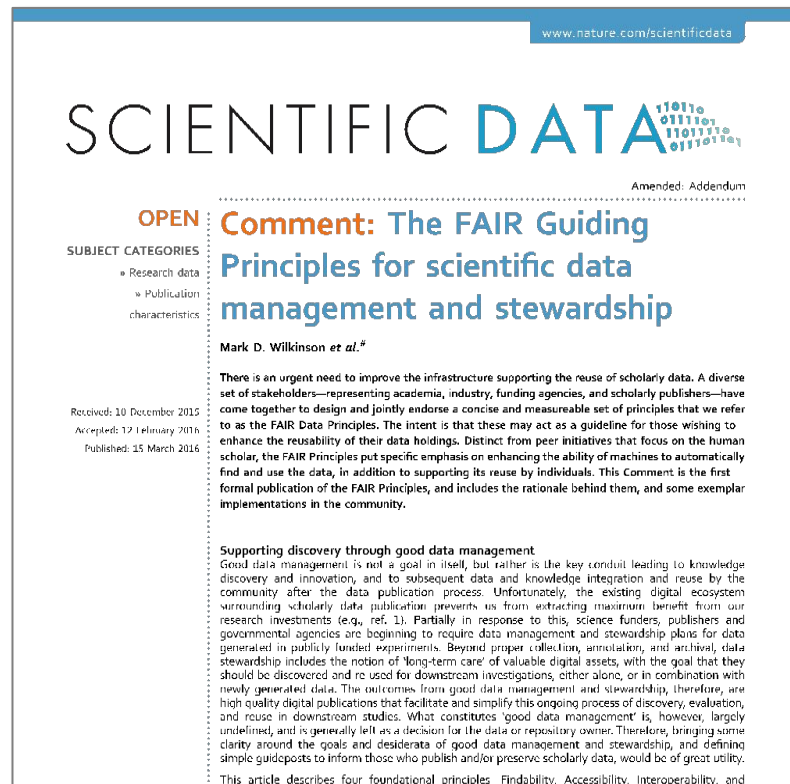


Wolmar N. Åkerström
Data Steward



Erik Hedman
Data Steward

- Promote **efficient data discovery and reuse** by providing guidelines to make digital resources
 - ☐ Findable
 - ☐ Accessible
 - ☐ Interoperable
 - ☐ Reusable
- Address aspects **enabling software and infrastructure** to automatically find and use research data



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

Swedish Research Council recommends open access to research data

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. [...]



Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

This document is a translation. The official version is in Swedish.

GOVERNING DOCUMENT SLU.ua: 2022.1.1.1-3278

Subject area: Research and doctoral education

Document type: Policy
Decision-maker: Vice-Chancellor
Organisational unit: Division of Planning
Reference: Sofia Wretblad

Decision date: 09/21/2022
Effective as of: 09/21/2022
Valid until: Further notice
To be updated by: 2023-12-31

Document(s) repealed:

Annex to: Vice-Chancellor's decision on Policy on the management of and open access to research and environmental monitoring and assessment data at SLU

Policy on the management of and open access to research and environmental monitoring and assessment data at SLU

<https://internt.slu.se/globalassets/mw/org-styr/styr-dok/forskning-forskarutb/data-management-policy-slu-eng-v-221019.pdf>

“ [...]

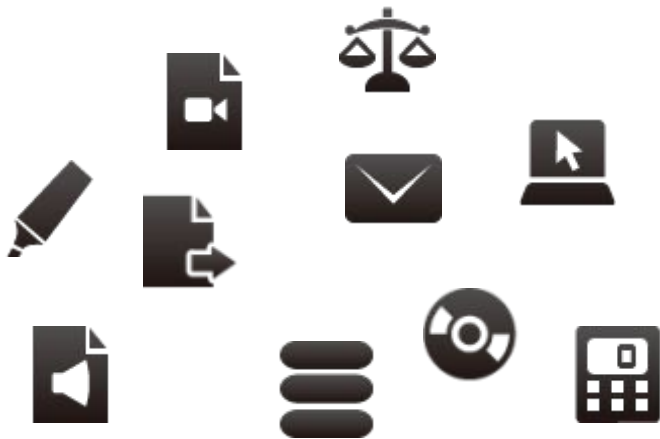
SLU endorses the FAIR data sharing principles (Findable, Accessible, Interoperable, and Reusable) and considers properly managed, accessible, and reusable data as a valuable and necessary resource for conducting research, teaching and environmental monitoring and assessment of high quality. [...] “

“ [...]

Good management of data throughout its entire lifecycle is a prerequisite for being able to comply with the FAIR data sharing principles to the greatest extent possible. [...] “

- **FAIR data \neq Open data**

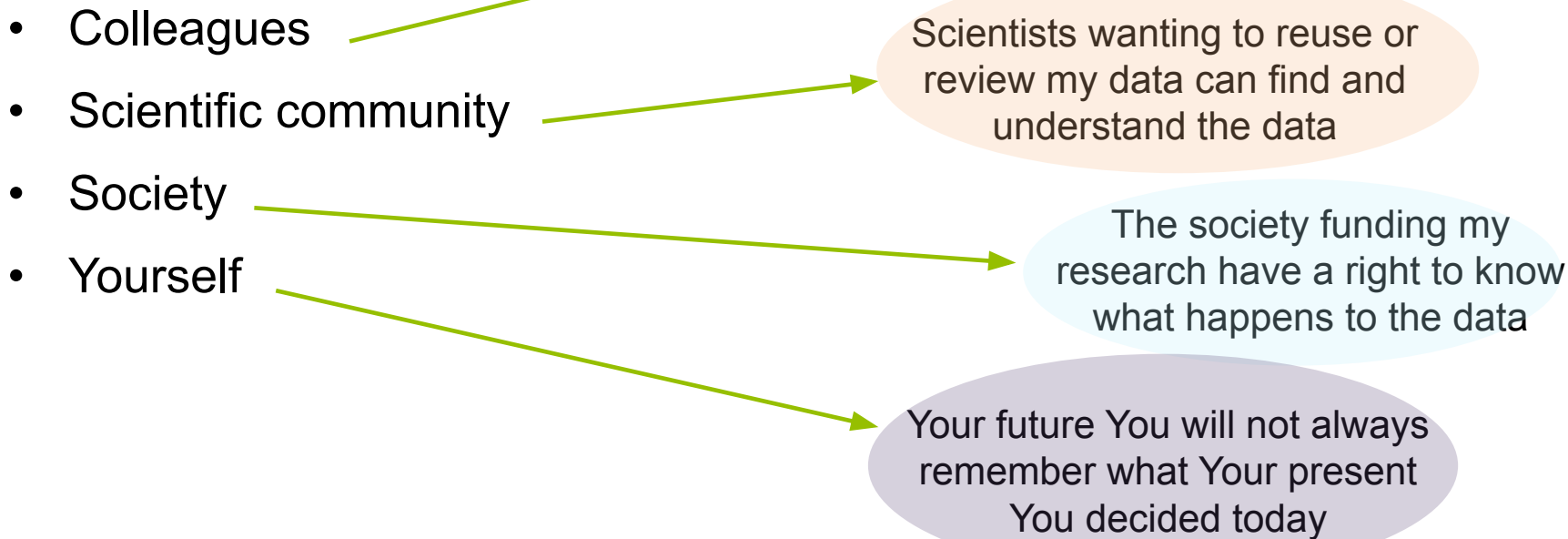
Data can be Open without being FAIR
Data can be FAIR without being open
“As open as possible, as closed as necessary”



A FAIR data lifecycle

- The FAIR principles relies on **good data management practices** in all phases of research
 - Research documentation
 - Data organisation
 - Information security
 - Ethics and legislation



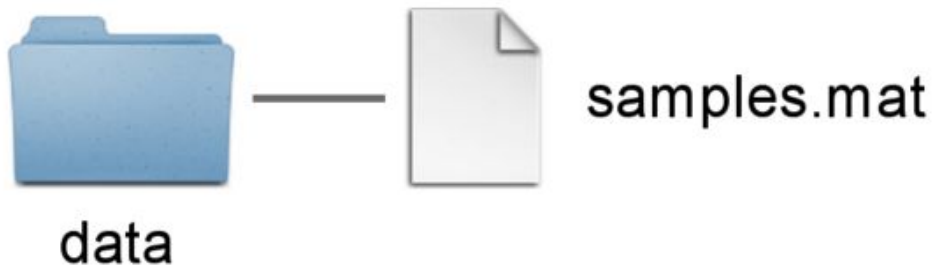


*“Your primary
collaborator is
yourself six months
from now, and your
past self doesn’t
answer e-mails,”*

-Rachael Ainsworth

How do you know how an old result was generated?

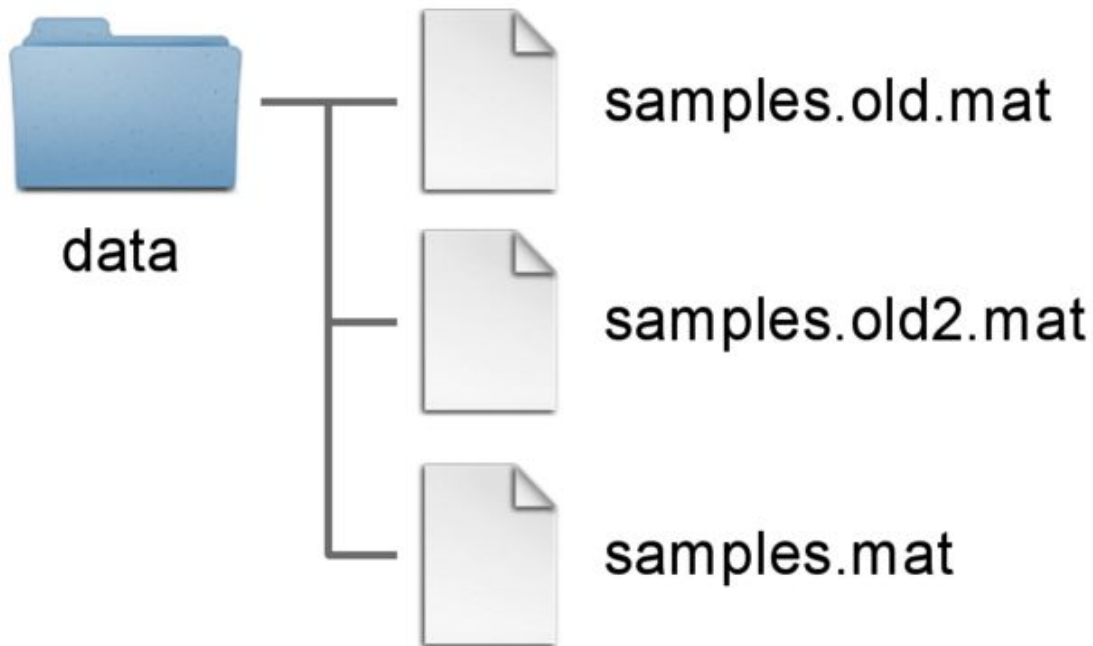
First step - Organization



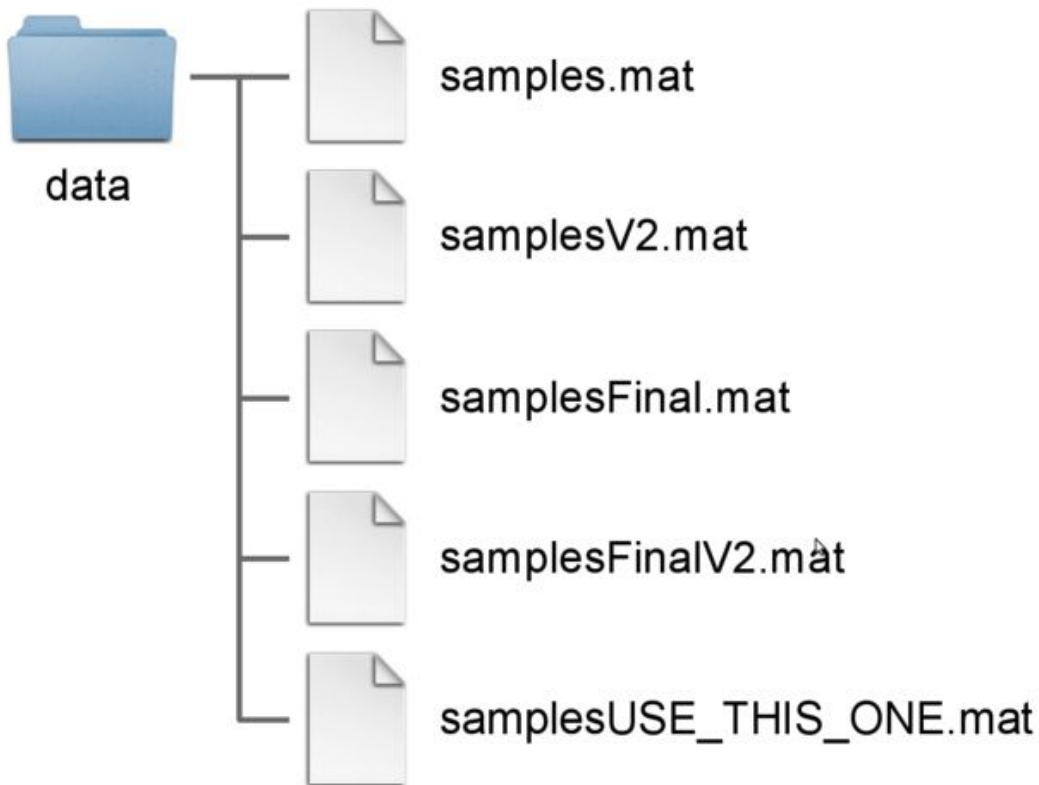
I guess this is alright



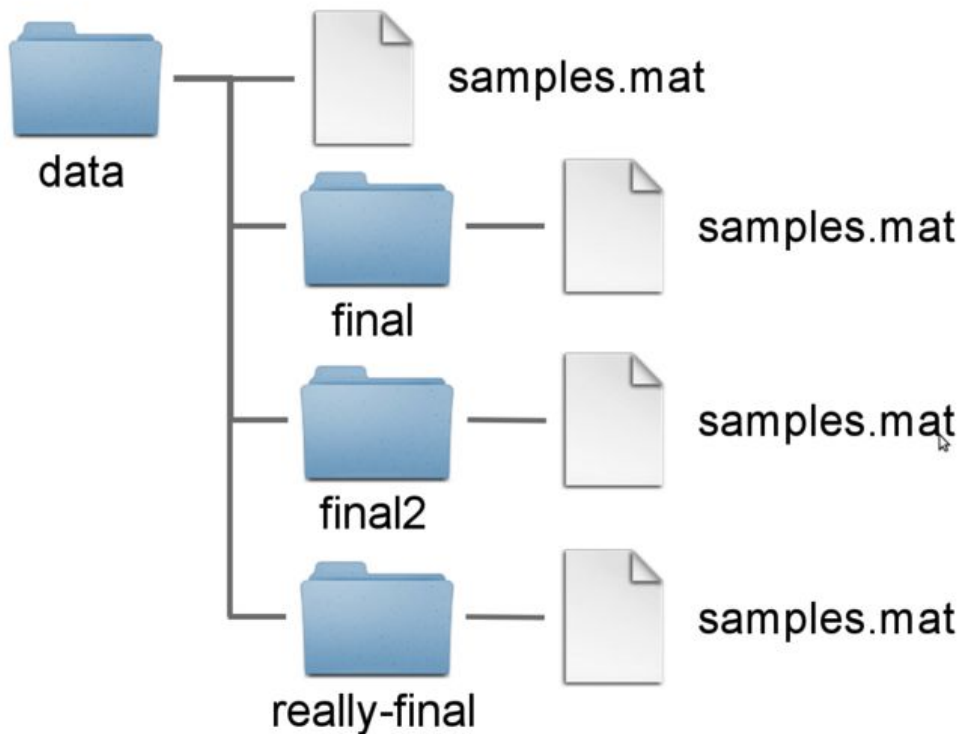
I guess this is alright



Which one is the most recent?



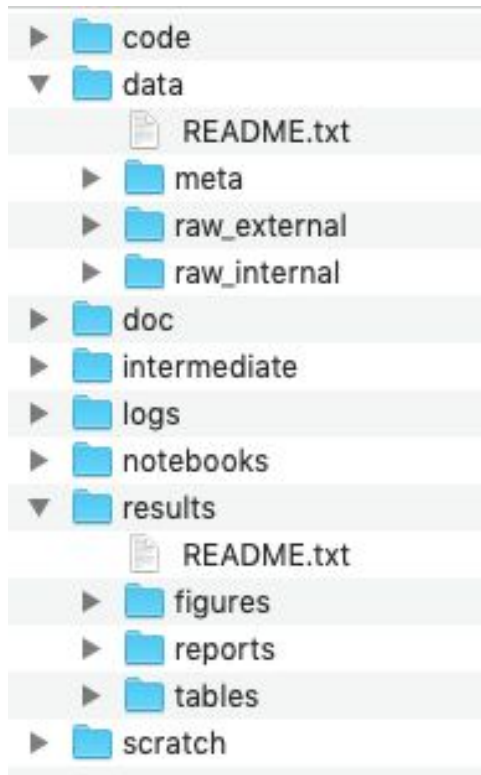
Another (bad) common approach



A possible solution



- 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 schemes** 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...



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*

- 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 schemes** 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...

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

Group 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

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

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

- 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 zeroes (i.e.: v005 instead of v5). make sure the end-letter file format extension is present at the end of the name (e.g. .doc, .xls, .mov, .tif)
- Add a README.md (or PROJECT_STRUCTURE.md) file in your top directory which details your naming convention, directory structure and abbreviations

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 including 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 suggestions by C.

Keywords Published

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

- 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
- Repetition, 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 OS may fail

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.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 |

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

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

FAIRify your research by using them as documentation files for:

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

Data – Explaining file names and contents

README in Markdown (.md)

- Allows text and content formatting without interference
- Highly compatible with e.g. GitHub
- Allows inclusion of comments without having to visualize them
- Easily editable and versatile
- Does not require particular skills

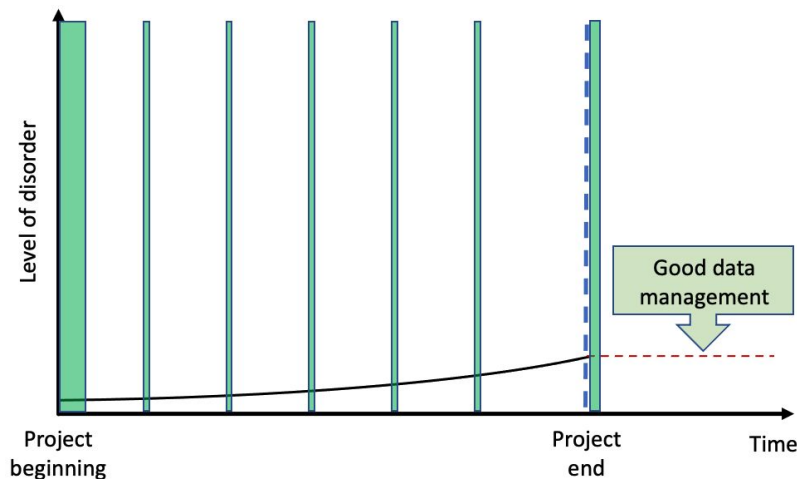
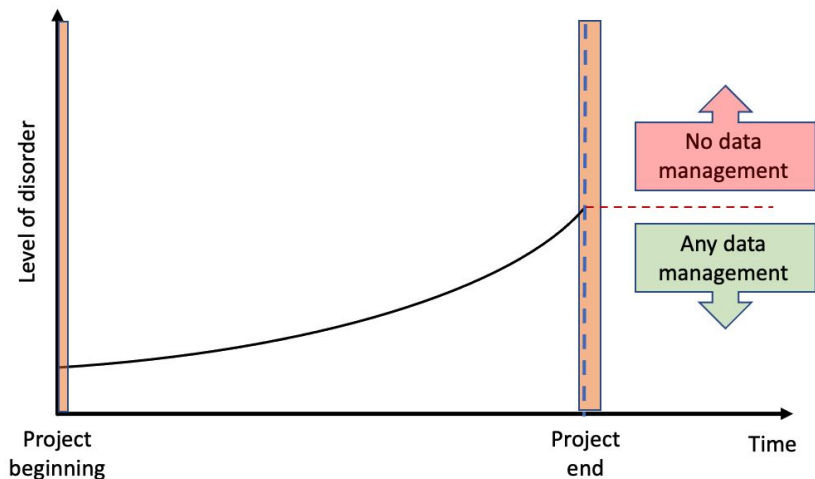
Discussion

Think of an example where you would have benefited from having access to a README-file when working with data. Describe to your neighbor what you would have wanted such a file to contain.

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



A FAIR data lifecycle

- The FAIR principles relies on **good data management practices** in all phases of research

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

- ☐ **Maintain a Data Management Plan**, outlining the project's data management practices

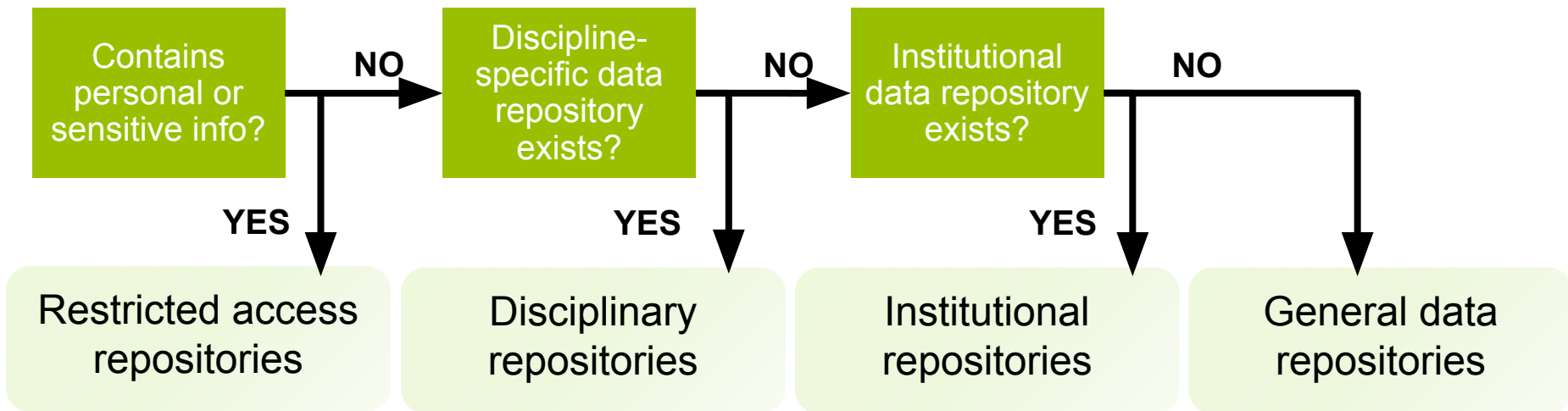


What to do?

Deposit the data in a data repository!

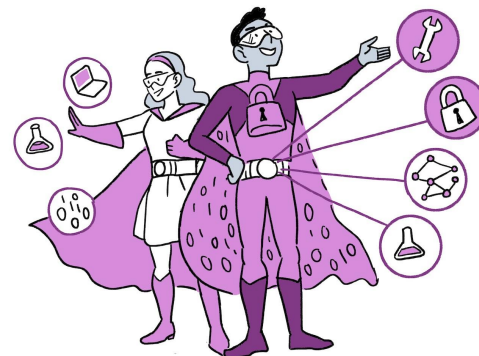


Selecting a data repository



- ❑ **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

- 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



ScilifeLab

Contact us

- nbis.se/support/supportform
- data-management@scilifelab.se

The purpose of these guidelines is to serve as an information resource to life science researchers in Sweden regarding research data management.

Research data management (RDM) concerns the organisation, storage, preservation, and sharing of data that is collected or analysed during a research project. Proper planning and management of research data will make project management easier and more efficient while projects are being performed. It also facilitates sharing and allows others to validate as well as reuse the data.

Research data life cycle

The research data life cycle can be divided into several phases as seen in the wheel below; **plan, collect, process, analyse, preserve, share** and **reuse**. Click on a section of the wheel below to get an introduction to that phase of the research data life cycle, including information on relevant resources and training material.



Get Support

Do you need support with research data management?

We offer support to anyone involved in life science research that is affiliated with a Swedish university or research institute.

[Click here to get support](#)

Meet a Data Steward

Join SciLifeLab Data Centre and NBIS get data management support. Each event consists of a 15 minutes mini-lecture and a 45 minutes Q&A.

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

data-management@scilifelab.se

Meet a data Steward Feb 7, 15.00-16.00

15.00-15.15 Mini-lecture: “Need to setup a DMP? Join us for a session about DMPs and the SciLifeLab Data Stewardship Wizard tool.” - Yvonne Kallberg, Data Steward from NBIS is presenting.

15.15-16.00 Bring your IT and RDM needs and meet our SciLifeLab Data Centre & NBIS experts.

The Spring 2023 “Meet a Data Steward” events will be Feb 7, March 14, April 18 and May 23. More info here: <https://www.scilifelab.se/event/meet-a-data-steward/>



- The **R**esearch **D**ata **M**anagement toolkit for Life Sciences

[Home](#)[About](#)[Contribute](#)[Contact](#)[GitHub](#)[Data life cycle](#)[Your role](#)[Your domain](#)[Your tasks](#)[Tool assembly](#)[All tools and resources](#)[All training resources](#)

Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing good data management practices applicable to research projects from the beginning to the end. Developed and managed by people who work every day with life science data, the RDMkit has guidelines, information, and pointers to help you with problems throughout the data's life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable and Reusable — by-design, from the first steps of data management planning to the final steps of depositing data in public archives.

The RDMkit organises information into the six sections displayed below, which are interconnected but can be browsed independently.

Data life cycle

Start here to get an overview of research data management. Click on a section of the diagram below to get an introduction to that stage of the data management life cycle.



Your role

Identify your role in research data management, find data management resources relevant for your role in research data management.

Your domain

Learn about the data management problems that affect your domain or research community, and the solutions and best practices available.

- RDM best practices and guidelines
- Links to tools/resources and training material given in specific DM context
- Examples of combination of tools for RDM

<https://rdmkit.elixir-europe.org/>

Introduction to Data Management Practices

[18-20 April, 2023, Uppsala](#)

Course name: Introduction to Data Management Practices

Contact: edu.intro-dm@nbis.se

This course will introduce important aspects of Research Data Management through a series of lectures and hands-on computer exercises. The course is intended for researchers that want to take the *first* steps towards a more systematic and reproducible approach to analysing and managing research data.

Course content

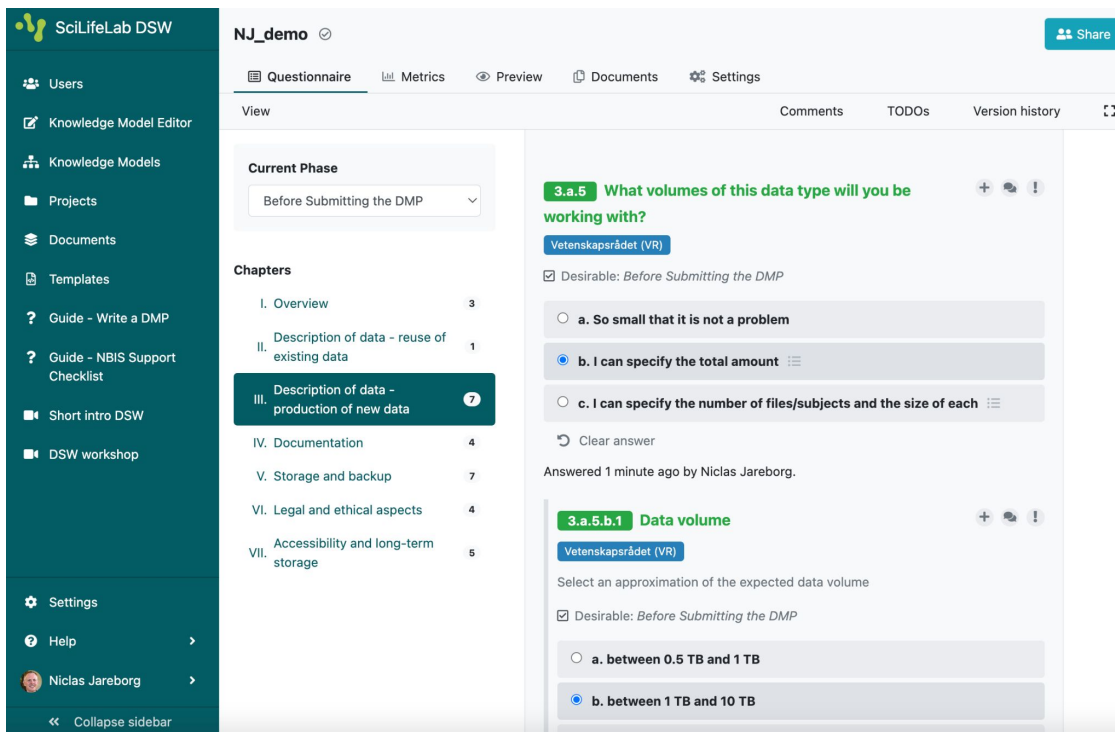
Topics covered will include:

- Open Science and FAIR in practice
- Organising data, files and folders in research projects
- Describing data with metadata
- Publishing data to public data repositories
- Cleaning tabular data and metadata with OpenRefine
- Writing basic recipes for data analysis and visualisation with R
- Versioning data, documents and scripts
- Writing Data Management Plans

<https://uppsala.instructure.com/courses/48087/pages/introduction-to-data-management-practices>

DSW - Data Stewardship Wizard

<https://dsw.scilifelab.se/>



SciLifeLab DSW

NJ_demo

Questionnaire Metrics Preview Documents Settings

View Comments TODOs Version history

Current Phase
Before Submitting the DMP

Chapters

- I. Overview 3
- II. Description of data - reuse of existing data 1
- III. Description of data - production of new data 7
- IV. Documentation 4
- V. Storage and backup 7
- VI. Legal and ethical aspects 4
- VII. Accessibility and long-term storage 5

3.a.5 What volumes of this data type will you be working with?

Vetenskapsrådet (VR)

☒ Desirable: Before Submitting the DMP

☐ a. So small that it is not a problem

☒ b. I can specify the total amount

☐ c. I can specify the number of files/subjects and the size of each

Clear answer

Answered 1 minute ago by Niclas Jareborg.

3.a.5.b.1 Data volume

Vetenskapsrådet (VR)

Select an approximation of the expected data volume

☒ Desirable: Before Submitting the DMP

☐ a. between 0.5 TB and 1 TB

☒ b. between 1 TB and 10 TB

Data Management Support

SLU Data Management Support (DMS) assists SLU employees with data management - from planning, including data management plans, to publishing, long-term preservation, archiving and the reuse of research and environmental assessment data.

How can we help?

Contact



SLU Data Management Guide

Here you will find resources to help you plan your data management, write a data management plan, publish data, find already published data and much more.



Manage monitoring data

Support for the management of data from environmental monitoring and assessment, including SLU's systematic quality work.



SND
Svensk nationell datajänst

Swedish National Data Service (SND)

SND is a network of around 40 universities, including SLU. Among SND's tasks is to offer researcher the possibility to publish data in its research data catalogue.

Shortcuts

- [Book a data date](#)
- [Workshops and webinars](#)
- [Data Management Plans \(DMPs\)](#)
- [Share and publish data](#)
- [FAIR data](#)
- [SLU's data management policy](#)
- [Frequently asked questions](#)

www.slu.se/dms

SLU Data Management Guide

Here you will find resources to help you plan your data management, write a data management plan, publish data, find already published data and much more.



Introduction to data management

What is research data management?



Plan data management

Make research more efficient, and create more value for data.



Collect, organise, and store data

Organise, document and describe data systematically.



Process and analyse data

Enable reproducibility, preservation and publishing of data.



Archive and preserve data

Ensure access to public records, cultural heritage and research needs.



Share and publish data

Publish data in a repository to ensure impact and visibility of research.



Discover, reuse, and cite data

Find research data and read about issues to consider when re-using data.

Shortcuts

- [How can we help?](#)
- [Book a data date](#)
- [Workshops and webinars](#)
- [Data Management Plans \(DMPs\)](#)
- [Contact Data Management Support](#)
- [FAIR data](#)
- [Publishing data via Swedish National Data Service \(SND\)](#)

<https://www.slu.se/en/subweb/library/publish-and-analyse/archiving-and-publishing-research-data/data-management-guide/>

Where are you today?

	Ad Hoc	One-Time	Active and Informative	Optimized for Re-Use
Planning your project	When it comes to my data, I have a "way of doing things" but no standard or documented plans.	I create some formal plans about how I will manage my data, but I generally don't refer back to them.	I develop detailed plans about how I will manage my data that I actively revisit and revise over the course of a project.	I design my plans for managing data to streamline future use by myself or others.
Organizing your data	I don't follow a consistent approach for keeping my data organized, so it often takes time to find things.	I have an approach for organizing my data, but I only put it into action after my project is complete.	I have an approach for organizing my data that I implement prospectively, but it not necessarily standardized.	I organize my data to the so that others can navigate, understand, and use it without me being present.
Saving and backing up your data	I decide what data is important while I am working on it and typically save it in a single location.	I know what data needs to be saved and I back it up after I'm done working on it to reduce the risk of loss.	I have a system for regularly saving important data while I am working on it. I have multiple backups.	I save my data in a manner and location designed maximize opportunities for re-use by myself and others.
Getting your data ready for analysis	I don't have a standardized or well documented process for preparing my data for analysis.	I have thought about how I will need to prepare my data, but I handle each case in a different manner.	My process for preparing data is standardized and well documented.	I prepare my data in such a way as to facilitate use by both myself and others in the future.
Analyzing your data and handling the outputs	I often have to redo my analyses or examine their products to determine what procedures or parameters were applied.	After I finish my analysis, I document the specific parameters, procedures, and protocols applied.	I regularly report the specifics of both my analysis workflow and decision making process while I am analyzing my data.	I have ensured that the specifics of my analysis workflow and decision making process can be put into action by others.
Sharing and publishing your data	I share the results of my research, but generally I do not share the underlying data.	I share my my data only when I'm required to do so or in response to direct requests from other researchers.	I regularly share the data that underlies my results and conclusions in a form that enables use by others.	Because of my excellent data management practices, I am able to efficiently share my data whenever I need to with whomever I need to.

Borghi, J. et al (2018). Support your Data.
<https://doi.org/10.3897/rio.4.e26439>