Data Management in Practice

Stephan Nylinder

NBIS / SciLifeLab - ELIXIR Sweden

data-management@scilifelab.se

https://doi.org/ADD_DOI_HERE

Policy landscape

UNESCO's Recommendation on Open Science

EU

- Open Science Policy
- Directives
- European Research Council
- Horizon Europe
- EOSC

Swedish research bills 2016 & 2020

- Transition to open research data implemented by 2026
- Government assignments to KB & VR

SUHF national roadmap for open science and University policies

Lund Declaration on Maximising the Benefits of Research Data

Open Science

FAIR

"FAIR […] open data sharing should become the default […]"

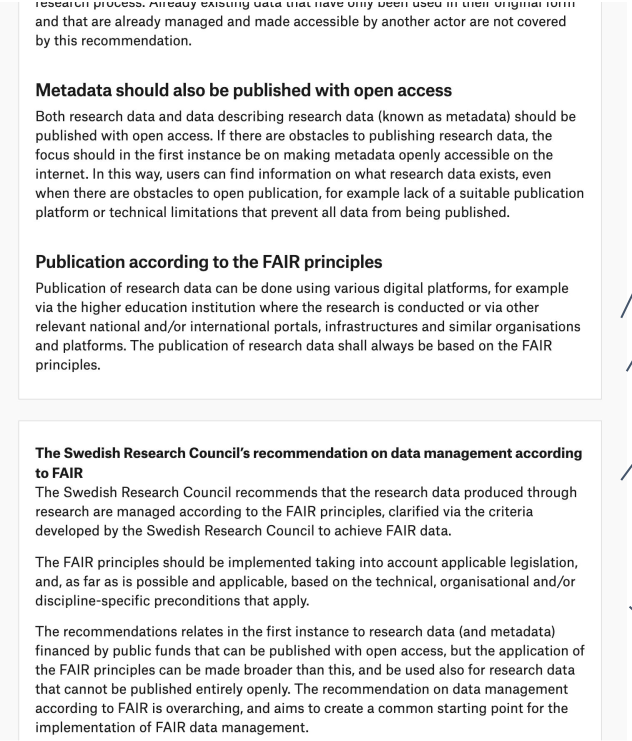
"As open as possible, as closed as necessary"

National guidelines for open science (KB)

Nationella riktlinjer för öppen vetenskap

VR - Swedish Research Council

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**. [...]*

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

Sharing - Stay FAIR

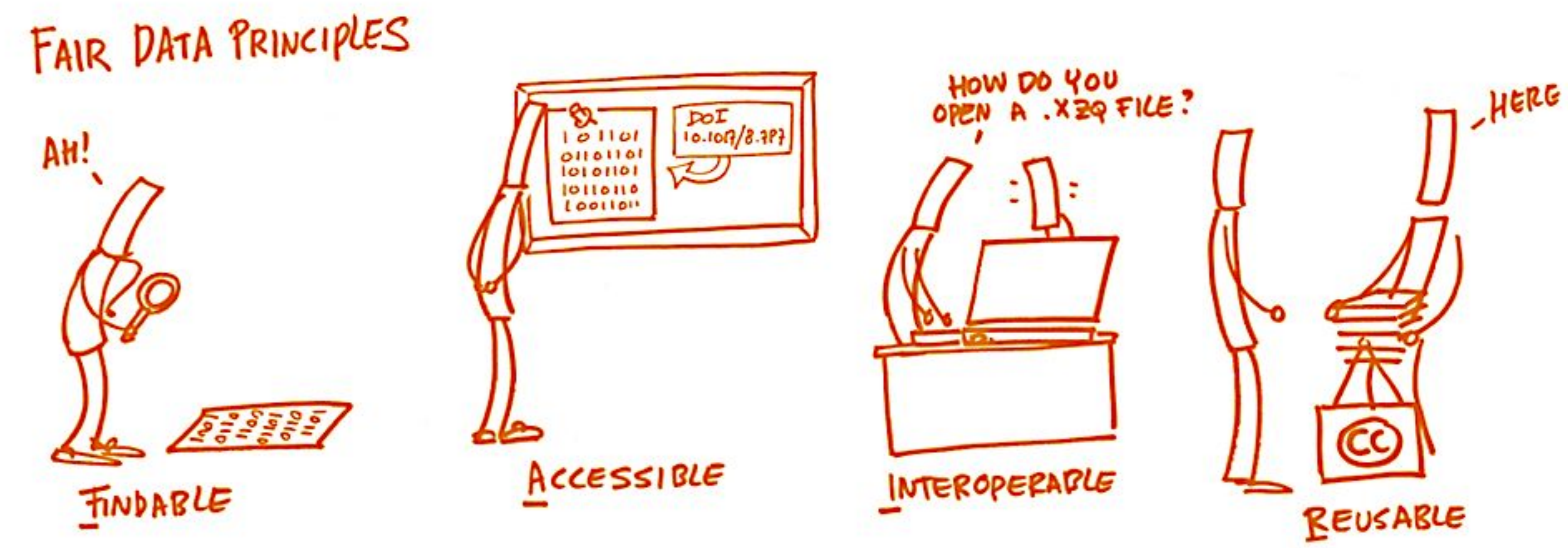


Image: https://book.fosteropenscience.eu/

Keeping data in a good shape

Organise your files in a structured way

- Use a file naming-convention
- Separate raw data from processed data
- Use simple README text files to describe content of folders

Use standard, non-proprietary, file formats

Have a plan for storing your data

Have a strategy for backing up your data

Stick to available standards for metadata (i.e. data about the data)

- Learn from repositories where the type of data can be deposited
- Make sure to collect relevant metadata as soon as the information is available
- Store the metadata where it is easy to find

Document changes to your project

- Consider using a version control system

Data Pipeline - The Scientist Perspective

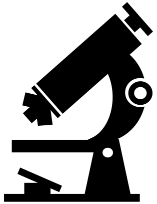
Year 0

The ideas!

Year 1

Data collection!

Year 2



Results!

Year 3

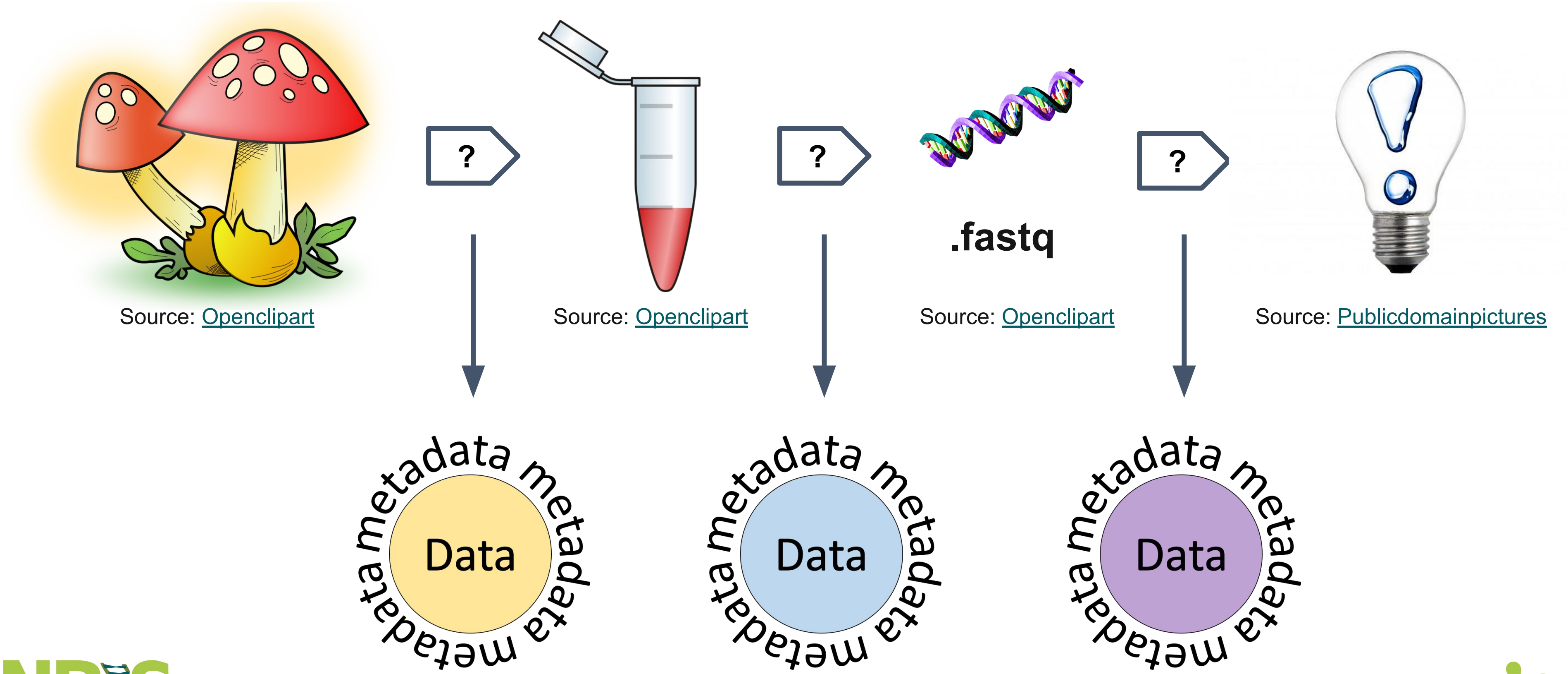
Paper out!

Finished!

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

I can do that!

What is metadata?


Source: Openclipart

Source: Openclipart

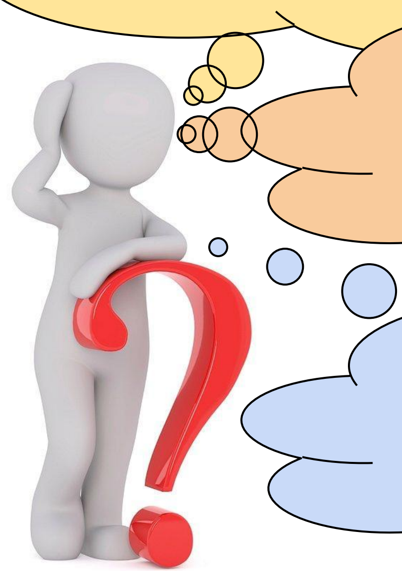
Source: Openclipart

Source: Publicdomainpictures

?
?
?
.fastq
Data
Data
Data
metadata metadata metadata metadata
metadata metadata metadata metadata
metadata metadata metadata metadata

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)

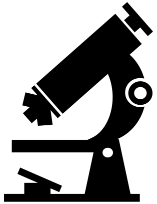
Which files to publish, and
where are they?

The Data Pipeline - Scientist perspective

Year 0

Year 1

Year 2



Year 3

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

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

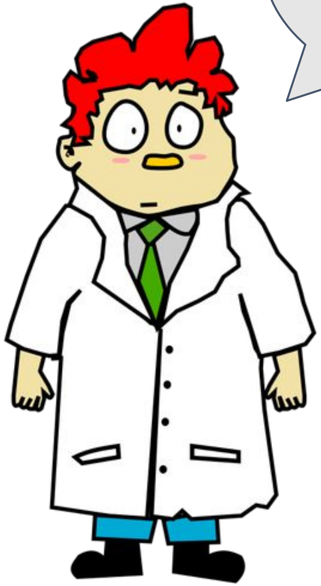
The data files? They are ALL here! At least I think so…
(1-2 years ago)

The Nice Reality

What do I have?

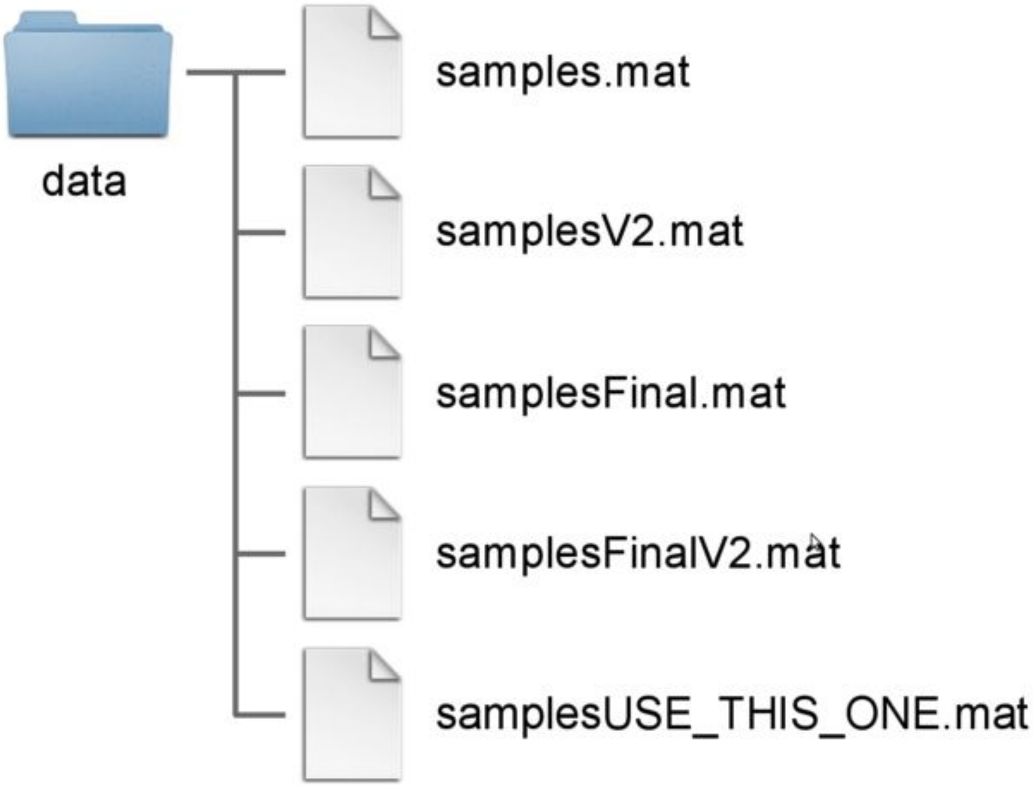
- Spreadsheet with sample information
- Electronic notebook with lab protocol
- Delivery report from sequencing facility
- A bioinformatic analysis report
- A bunch of data files somewhere

How do I describe so that others can understand?



Source: Openclipart

The Grim Reality



data

samples.mat

samplesV2.mat

samplesFinal.mat

samplesFinalV2.mat

samplesUSE_THIS_ONE.mat

Why submit to a repository?

“The data is available upon request”

Many reasons:

- Open Science & FAIR
- Reproducibility
- Trail of evidence
- 3rd party access
- Archival purposes
- Publication of paper requires it

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

What research outputs should be submitted?

- **Raw data:** straight from the instrument eg fastq, bam, cram
- **Processed data:** normalization, removal of outliers, expression measurements, statistics
- **Metadata:** minimum information to reproduce the data, sample information, precise protocols
- **Code:** software code that is needed to re-run analyses

Types of repositories

Domain-specific:

- Best choice - long-term plan, typically free, maximum reach
- E.g. European Nucleotide Archive, European Genome Phenome Archive, ArrayExpress, PRIDE

General purpose:

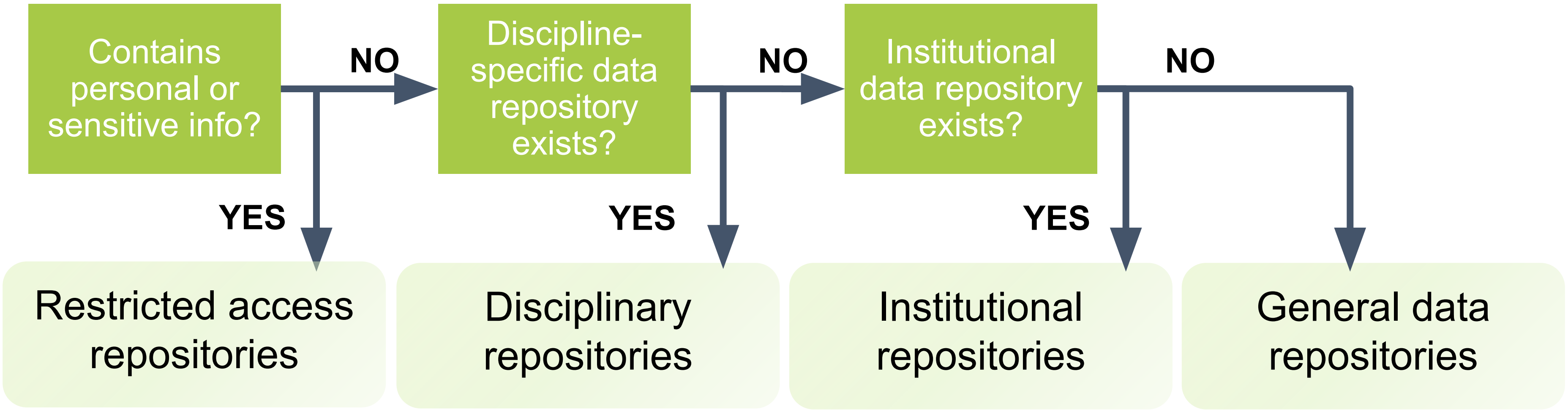
- Second best – long-term plan, might cost (now or in future), good reach but less specific in metadata → more difficult for future users to judge if a dataset will be useful
- E.g. Zenodo, SciLifeLab Data Repository (Figshare), Dryad

In-house/institutional

- For archive/backup purpose mainly, might cost, limited reach unless also published in a data catalogue

Domain-specific
General purpose
In-house

Selecting a Data Repository



Contains personal or sensitive info?

NO

Discipline-specific data repository exists?

NO

Institutional data repository exists?

NO

YES

YES

YES

Restricted access repositories

Disciplinary repositories

Institutional repositories

General data repositories

EUROPEAN GENOME-PHENOME ARCHIVE

Federated European Genome-phenome Archive Sweden

ENA European Nucleotide Archive

PRIDE

BioSamples

Protein Data Bank in Europe
Bringing Structure to Biology

Stockholm University

SND Svensk nationell datatjänst

SciLifeLab Data Repository

zenodo

figshare

EUDAT

NBIS

https://www.re3data.org/

https://scilifelab-data-guidelines.readthedocs.io/

What about sensitive data?

Data regarded as special category data under GDPR may be possible to share under **controlled access**.

Controlled access means that researchers only will be granted access after a formal application procedure.

- The European Genome-phenome Archive (EGA) is a repository for archiving and sharing sensitive personal data from biomedical research projects
- FEGA Sweden – the Swedish node of the Federated EGA which is working to become operational.

EUROPEAN
GENOME-PHENOME
ARCHIVE

Federated
European
Genome-phenome
Archive
Sweden

If you cannot deposit the data in a repository: create at least a record describing the data (a ”metadata-record”) in e.g. SciLifeLab Data Repository.