



Data Management at SciLifeLab

Swedish Bioinformatics Workshop

2023-11-07

Erik Hedman

**National Bioinformatics
Infrastructure Sweden**

Angela Fuentes Pardo

SciLifeLab Data Centre

Elisabeth Sundström

SciLifeLab Data Centre

data-management@scilifelab.se

What is SciLifeLab Data Management?



- A collaborative activity between **SciLifeLab Data Centre** and the **Data Management team** at **NBIS**.
- Right now, 17 Project Leaders and Data Stewards working with different aspects of RDM.



What is SciLifeLab Data Management?



- A collaborative activity between **SciLifeLab Data Centre** and the **Data Management team** at **NBIS**.
- Right now, 17 Project Leaders and Data Stewards working with different aspects of RDM.
- Promote **Open Science**, **FAIR**, and **good RDM** practices.
- Provide services and resources for data management, IT and **data sharing**.
- Make **RDM** and **bioinformatics** support and training **easily accessible**.



RDM Guidelines



- **Knowledge hub for management of life science research data in Sweden**
-

- **Web portal** with **good examples** on how to adhere to the **FAIR principles**.


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



RDM Guidelines



- Knowledge hub for management of life science research data in Sweden

- **Web portal** with **good examples** on how to adhere to the **FAIR principles**.
 - <https://data-guidelines.scilifelab.se> 
- How to **maximise** the **value** of the **research data** throughout the data life cycle.



Research data life cycle



RDM Guidelines



- **Knowledge hub for management of life science research data in Sweden**

- **Web portal** with **good examples** on how to adhere to the **FAIR principles**.
 - <https://data-guidelines.scilifelab.se> 
- How to **maximise** the **value** of the **research data** throughout the data life cycle.
- Swedish life science perspective.
- Developed and maintained by SciLifeLab Data Centre and NBIS.
- **Contact:** data-management@scilifelab.se 

What is RDM?

Data transfer

Human data

FAIR principles

Research data life cycle



Outline of the Workshop



Presentation (~45 min)

1. Open science and FAIR
2. Data Management Plan
3. Organization
4. Versioning
5. Workflow Management Systems
6. Publishing data

Workshop (~60 min)

1. Introduction
2. Group exercises in Google Docs
3. Presentation and Feedback
4. Wrap-up and Summation

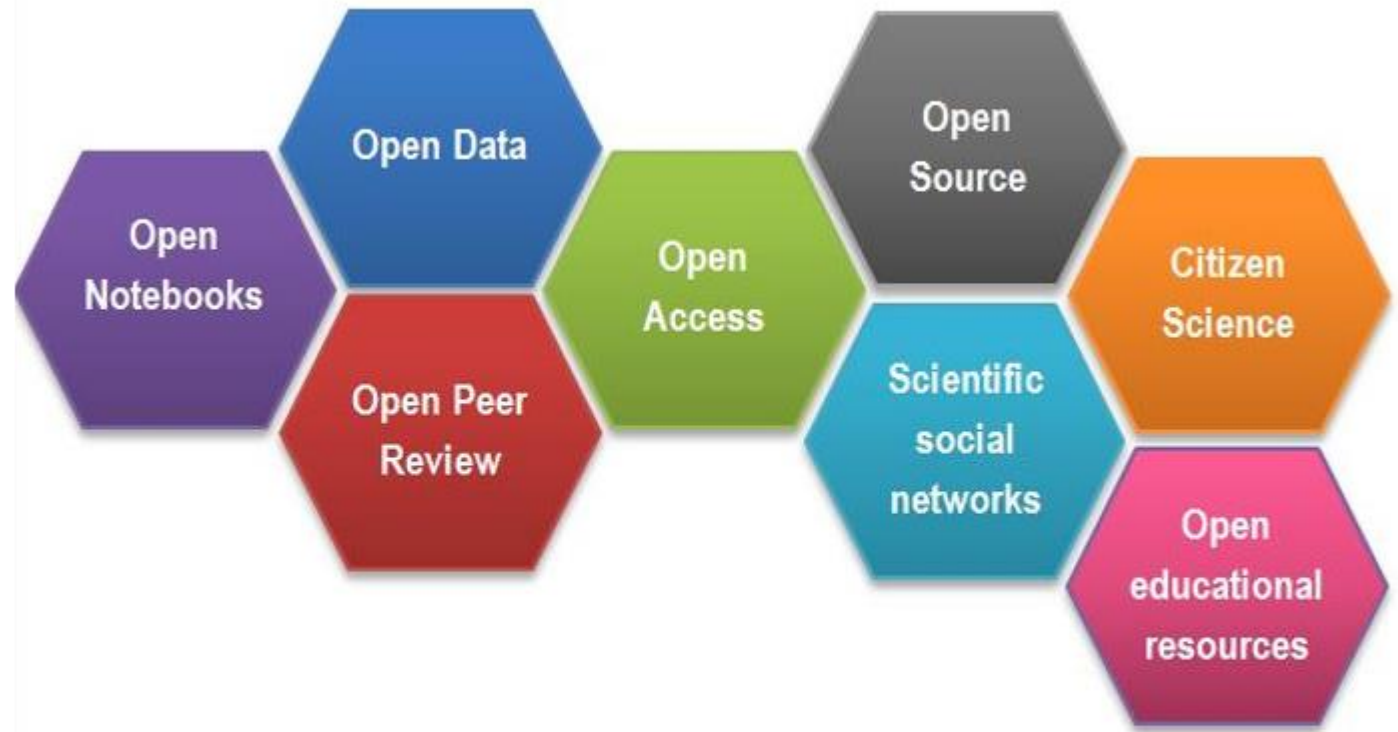
There will be breaks...

Open Science



“Make scientific research and its dissemination accessible to all levels of society”

- Open methodology
- **Open source**
- **Open data**
- Open access
- Open peer review
- Open educational resources



[“Open Science facets as a beehive”](#) by
Gema Bueno de la Fuente licenced under
[CC-BY](#)

Why does Open Science matter?



- **Democracy and transparency**

- Publicly funded research data should be accessible to all
- Published results and conclusions should be possible to check by others



Why does Open Science matter?



- **Democracy and transparency**

- Publicly funded research data should be accessible to all
- Published results and conclusions should be possible to check by others

- **Research**

- Enables others to combine data, address new questions, and develop new analytical methods
- Reduce duplication and waste



Why does Open Science matter?



- **Democracy and transparency**

- Publicly funded research data should be accessible to all
- Published results and conclusions should be possible to check by others

- **Research**

- Enables others to combine data, address new questions, and develop new analytical methods
- Reduce duplication and waste

- **Innovation and utilization outside research**

- Public authorities, companies, and private persons outside research can make use of the data



Why does Open Science matter?



- **Citation**

- Citation of data will be a merit for the researcher that produced it



Why does Open Science matter?



- **Citation**

- Citation of data will be a merit for the researcher that produced it

- **Ethical**

Doing “sloppy” science & not being open and transparent, could result in:

- Waste of resources
- Contributing to the current research credibility crisis
- Contributing to the current reproducibility crisis
- Harming the profession
- Harming public trust in research

My take of material by Rochelle Tractenberg “[Unexpected Ethical Challenges in Bioinformatics and Genomics.](#)”



Picture source: [Karolinska institute library](#)

FAIR principles



- To be useful for others, data should be FAIR

Wilkinson, Mark et al. “*The FAIR Guiding Principles for scientific data management and stewardship*”. Scientific Data 3, Article number: 160018 (2016)

<http://dx.doi.org/10.1038/sdata.2016.18>

www.nature.com/scientificdata

SCIENTIFIC DATA

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

SUBJECT CATEGORIES

- Research data
- Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

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 measurable 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 extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

FAIR principles



- To be useful for others, data should be FAIR

Findable



- Data should be easily found by both, humans and machines
- Data have a globally unique persistent identifier (e.g. a DOI)
- Data are described by metadata (*data about data*)

FAIR principles



- To be useful for others, data should be FAIR

Findable



- Data should be easily found by both, humans and machines
- Data have a globally unique persistent identifier (e.g. a DOI)
- Data are described by metadata (*data about data*)

Accessible



- Data is retrievable through a standardised communication protocol (open, free, allowing authentication & authorisation where necessary) (e.g. *http, sftp, etc.*)
- Metadata are accessible, even if data is no longer available

FAIR principles



- To be useful for others, data should be FAIR

Findable



- Data should be easily found by both, humans and machines
- Data have a globally unique persistent identifier (e.g. a DOI)
- Data are described by metadata (*data about data*)

Accessible



- Data is retrievable through a standardised communication protocol (open, free, allowing authentication & authorisation where necessary) (e.g. *http, sftp, etc.*)
- Metadata are accessible, even if data is no longer available

Interoperable



- Data can be easily integrated with other data, so it can be utilised by other applications (analysis, storage, processing)
- Metadata use vocabularies that follow the FAIR principles (*standardised ways of capturing information about the data*)

FAIR principles



- To be useful for others, data should be FAIR

Findable



- Data should be easily found by both, humans and machines
- Data have a globally unique persistent identifier (e.g. a DOI)
- Data are described by metadata (*data about data*)

Aaccessible



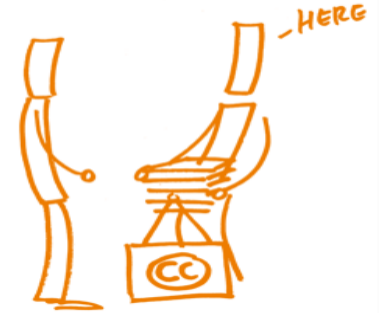
- Data is retrievable through a standardised communication protocol (open, free, allowing authentication & authorisation where necessary) (e.g. *http, sftp, etc.*)
- Metadata are accessible, even if data is no longer available

Interoperable



- Data can be easily integrated with other data, so it can be utilised by other applications (analysis, storage, processing)
- Metadata use vocabularies that follow the FAIR principles (*standardised ways of capturing information about the data*)

Reusable



- Data and related metadata should be well described, so the data can be reused, replicated and/or combined in different settings
- Data have a clear data usage license (*under what conditions it can be reused*)

FAIR principles



- To be useful for others, data should be FAIR

Findable



- Data should be easily found by both, humans and machines
- Data have a globally unique persistent identifier (e.g. a DOI)
- Data are described by metadata (*data about data*)

Accessible



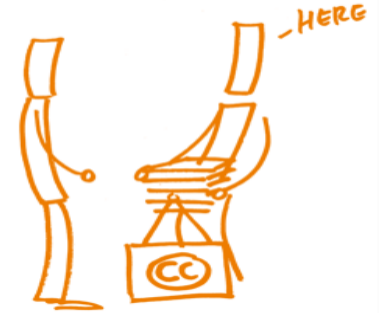
- Data is retrievable through a standardised communication protocol (open, free, allowing authentication & authorisation where necessary) (e.g. *http, sftp, etc.*)
- Metadata are accessible, even if data is no longer available

Interoperable



- Data can be easily integrated with other data, so it can be utilised by other applications (analysis, storage, processing)
- Metadata use vocabularies that follow the FAIR principles (*standardised ways of capturing information about the data*)

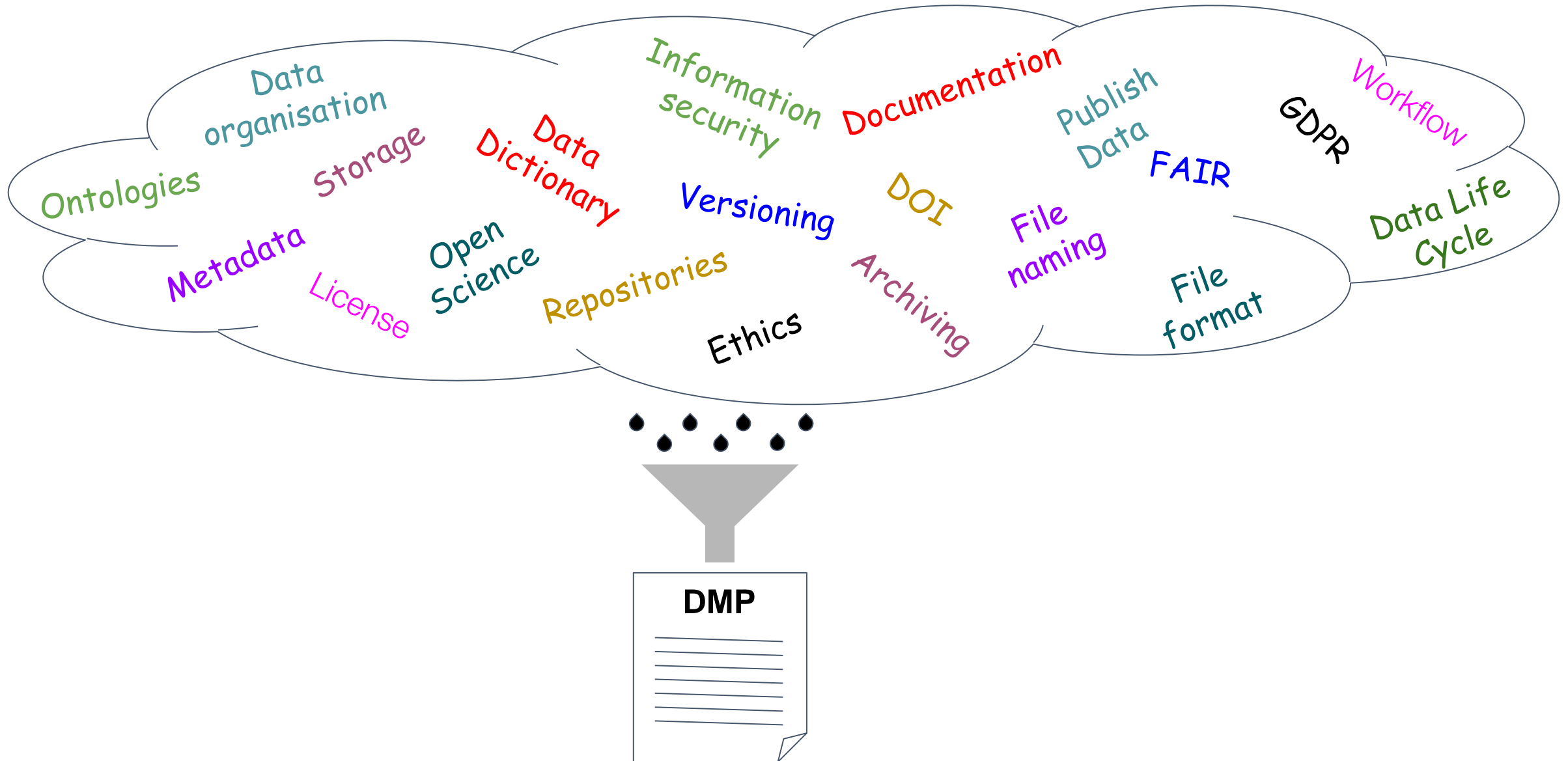
Reusable



- Data and related metadata should be well described, so the data can be reused, replicated and/or combined in different settings
- Data have a clear data usage license (*under what conditions it can be reused*)

Data Management Plan (DMP)

- What is a DMP and why should I write one?



What is a DMP?

- A document addressing requirements and practices for the project's data



The Swedish Research Council: All who are awarded a grant from the Swedish Research Council must have a data management plan if the research generates research data.



What is a DMP?

- A document addressing requirements and practices for the project's data



The Swedish Research Council: All who are awarded a grant from the Swedish Research Council must have a data management plan if the research generates research data.

- **Outlines** the data management **strategies** in a project, and **how** the **data** is:
 - collected
 - documented
 - organized
 - preserved



When to write a DMP?



- A DMP is a living document that will develop throughout the project

- **Project planning**

- Outline the strategies, and estimate the resources needed for funding



When to write a DMP?



- A DMP is a living document that will develop throughout the project

- **Project planning**

- Outline the strategies, and estimate the resources needed for funding

- **Project start**

- Complete with details

Data quality measures

File and folder strategies



When to write a DMP?



- A DMP is a living document that will develop throughout the project

- **Project planning**

- Outline the strategies, and estimate the resources needed for funding

- **Project start**

- Complete with details

Data quality measures

File and folder strategies

- **Project end**

- Update with e.g. links to published data and details about archiving

What data and where

Reusability



Why write a DMP?

- Think of the DMP as a checklist



Why write a DMP?

- Think of the DMP as a checklist

- **Identify** data management **gaps**
- Establish **project-wide** standards



Why write a DMP?

- Think of the DMP as a checklist

- **Identify** data management **gaps**
- Establish **project-wide** standards
- Estimate **costs**
- Define **responsibilities**



Why write a DMP?



- Think of the DMP as a checklist
- **Identify** data management **gaps**
- Establish **project-wide** standards
- Estimate **costs**
- Define **responsibilities**
- Ensure **well-managed** research data
- First step towards **FAIR**ness
- **Meet** funder and stakeholder **demands**



Why write a DMP?



- Think of the DMP as a checklist

- **Identify** data management **gaps**
- Establish **project-wide** standards
- Estimate **costs**
- Define **responsibilities**
- Ensure **well-managed** research data
- First step towards **FAIR**ness
- **Meet** funder and stakeholder **demands**

Reduce time spent later on

Openness

Reproducibility

Facilitating collaboration

Return of investment

Transparency



The main parts of a DMP



- Important chapters to help your future self
-

1. Description of data

- What types of data will be created and/or collected?

Formats

Amount/volume of data

Instrument

Equipment

The main parts of a DMP



- Important chapters to help your future self
-

1. Description of data

- What types of data will be created and/or collected?

Formats

Amount/volume of data

Instrument

Equipment

2. Documentation

- How will the material be documented and described?

ELN & LIMS

Collection method

Metadata standards

Versioning

The main parts of a DMP



- Important chapters to help your future self

1. Description of data

- What types of data will be created and/or collected?

Formats

Amount/volume of data

Instrument

Equipment

2. Documentation

- How will the material be documented and described?

ELN & LIMS

Collection method

Metadata standards

Versioning

3. Storage and backup

- How is data security, storage and backup handled?

Organization

Naming convention

Backup strategy

Access

Security

The main parts of a DMP



- Important chapters to help your future self
-

4. Legal and ethical aspects

- How is data handled? Any legal requirements?

Sensitive data

Confidentiality

Intellectual property rights

The main parts of a DMP



- Important chapters to help your future self
-

4. Legal and ethical aspects

- How is data handled? Any legal requirements?

Sensitive data

Confidentiality

Intellectual property rights

5. Accessibility and long-term storage

- How, when, and where will research data and metadata be made accessible?

Repositories

Raw data

Code & Software

Type of storage

The main parts of a DMP



- Important chapters to help your future self

4. Legal and ethical aspects

- How is data handled? Any legal requirements?

Sensitive data

Confidentiality

Intellectual property rights

5. Accessibility and long-term storage

- How, when, and where will research data and metadata be made accessible?

Repositories

Raw data

Code & Software

Type of storage

6. Responsibility and resources

- Who are the responsible persons for data management?

Organization

Naming convention

Backup strategy

Access

Security

Data Stewardship Wizard (DSW)

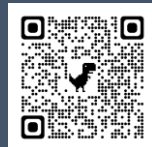


- Tool for creating Data Management Plans (DMPs)

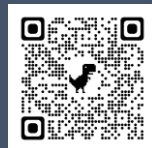
- **Interactive questionnaire** specific for life sciences.
- **Adapted templates** to national funders e.g. **Swedish Research Council** and **H2020**.
- Login by using your university account (SWAMID).
- Do you **need help**? We provide support and guidance!


Learn more:


<https://dsw.scilifelab.se>





youtu.be/HY2DVnNGkAs
(short DSW introduction)





 SciLifeLab DSW


 Projects

 Guide - Write a DMP


 Guide - NBIS Support Checklist


 Short intro DSW

 DSW workshop

 About project templates

Log In


 Email

 Password

[Forgot your password?](#)

[Log In](#)

Or connect with

 Life Science RI (university)



Take a leg-stretcher (5 minutes)

Organization - collecting and process phases



- Why do we need to keep good quality records?
- Ensures data, analysis and results to be **transparent, reproducible and traceable** – Accountability!
- Keeping good records **prevents misunderstandings**. Quality of subsequent research.
- In cumulative science **mistakes can result in cascade effects**.
- **Reduces the risk** of data mistakes, data manipulation and research fraud.
- **Promotes open science**
- Promotes data and documentation being **FAIR**.



Metadata

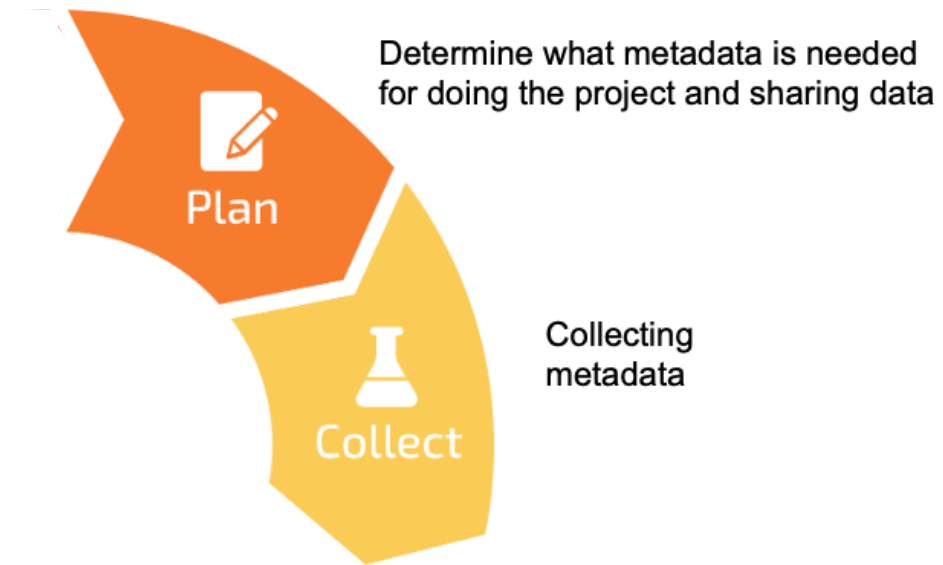
- Used throughout the research data life cycle
-



Determine what metadata is needed
for doing the project and sharing data

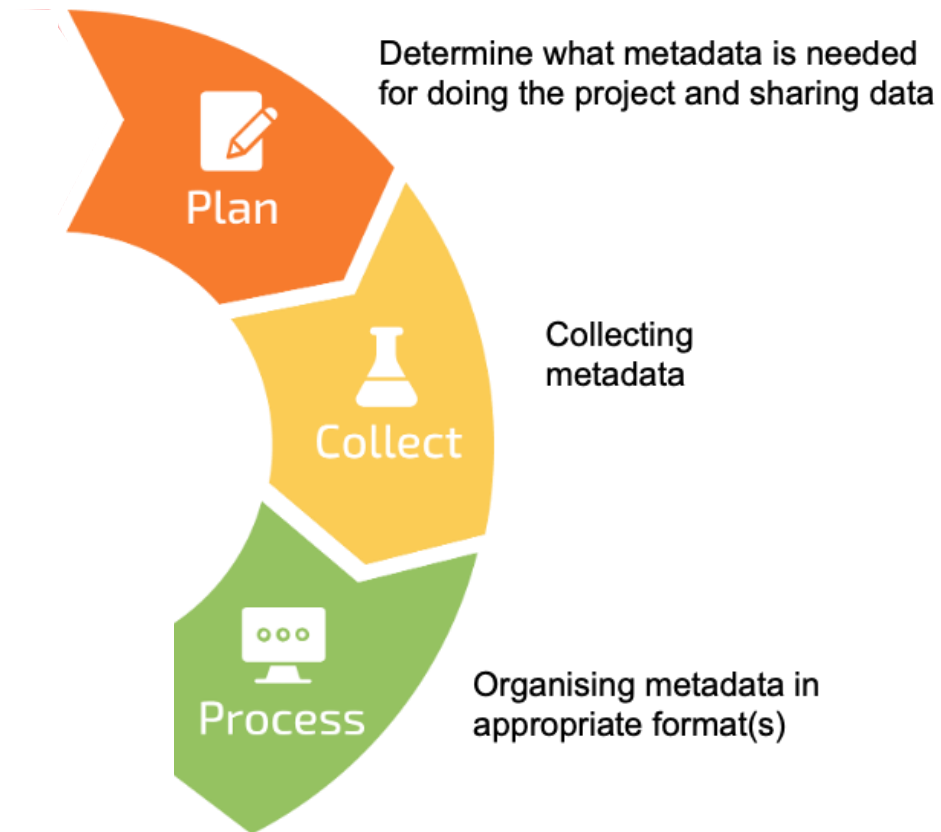
Metadata

- Used throughout the research data life cycle



Metadata

- Used throughout the research data life cycle



Metadata

- Used throughout the research data life cycle



Metadata

- Used throughout the research data life cycle



Metadata

- Used throughout the research data life cycle



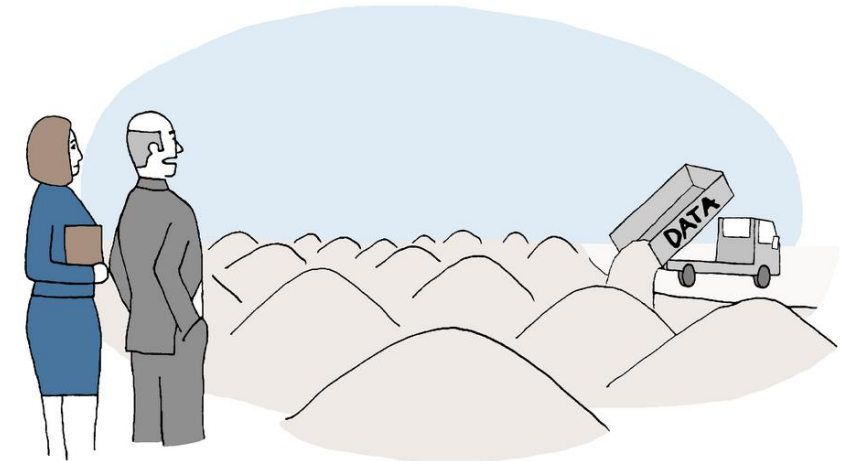
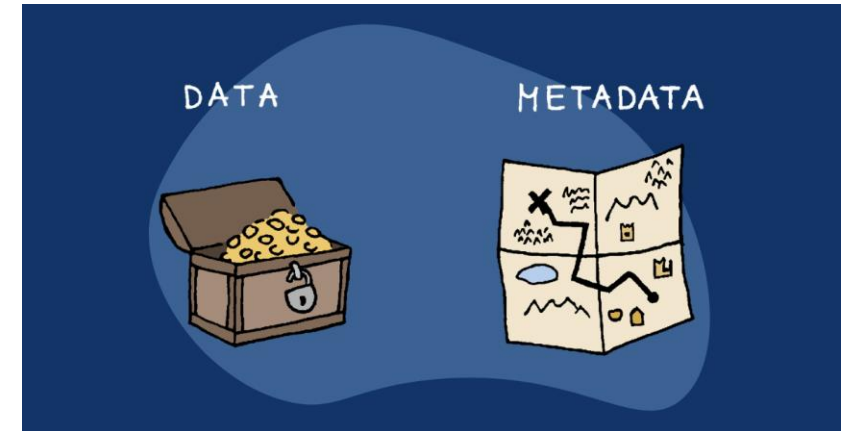
Metadata

- It makes life so much easier!



Examples of metadata:

- File types and formats of the data
- Methodology for data collection
- Analytical and procedural information
- Sources of samples
- Sample treatment
- Geolocation(s) of samples
- Licence for re-use of the data



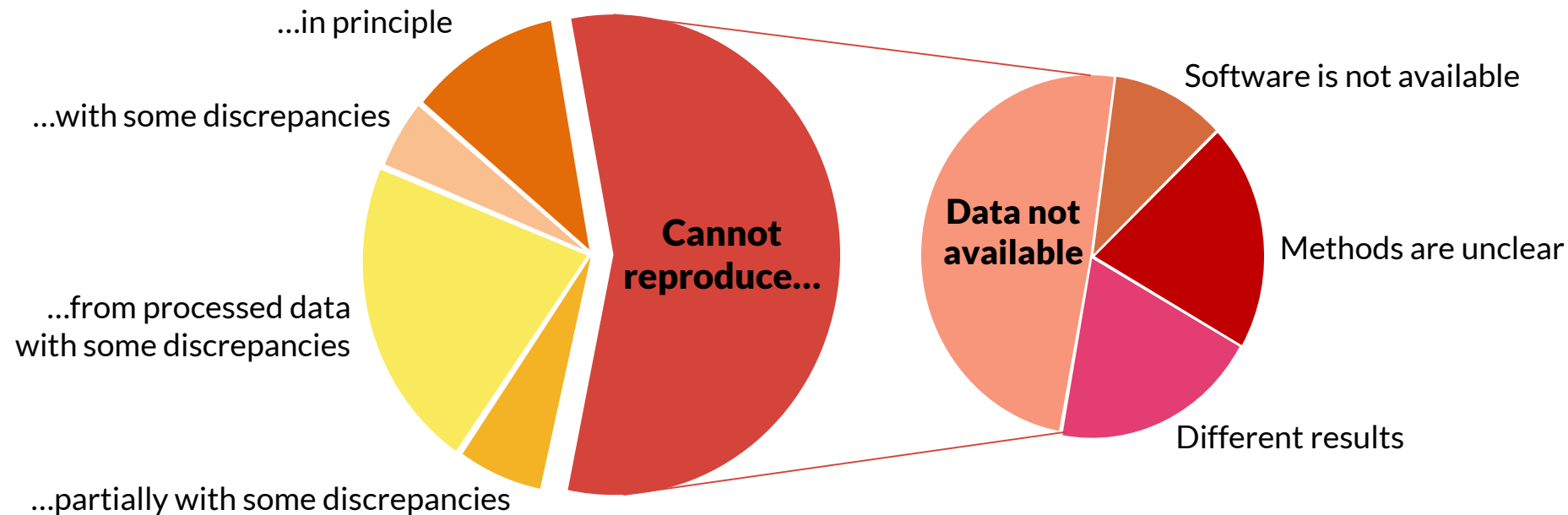
SO YOU'RE TELLING ME THIS WILL TURN INTO VALUE?

Reproducibility crisis



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

Can reproduce...



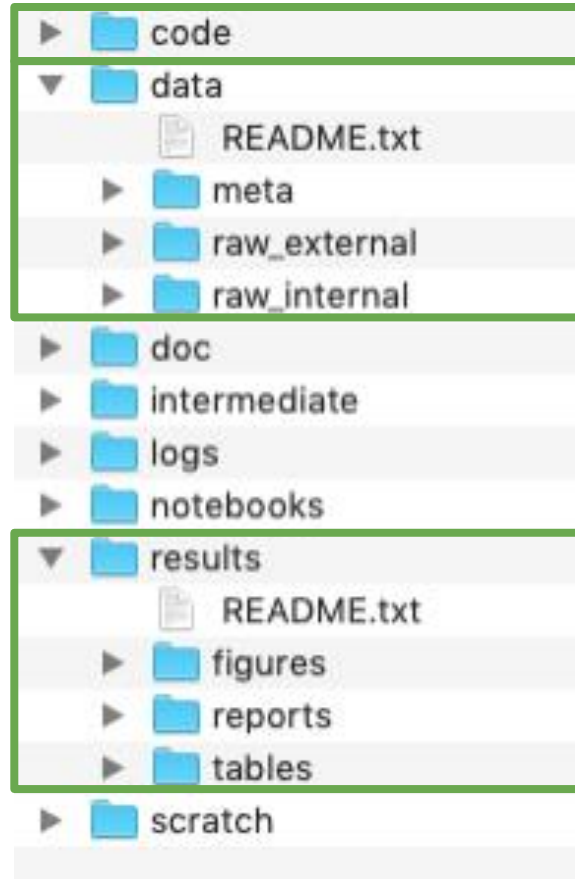
Summary of the efforts to replicate the published analyses.

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

How to organize your data



- Keep a structured directory environment



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*

How to organize your data



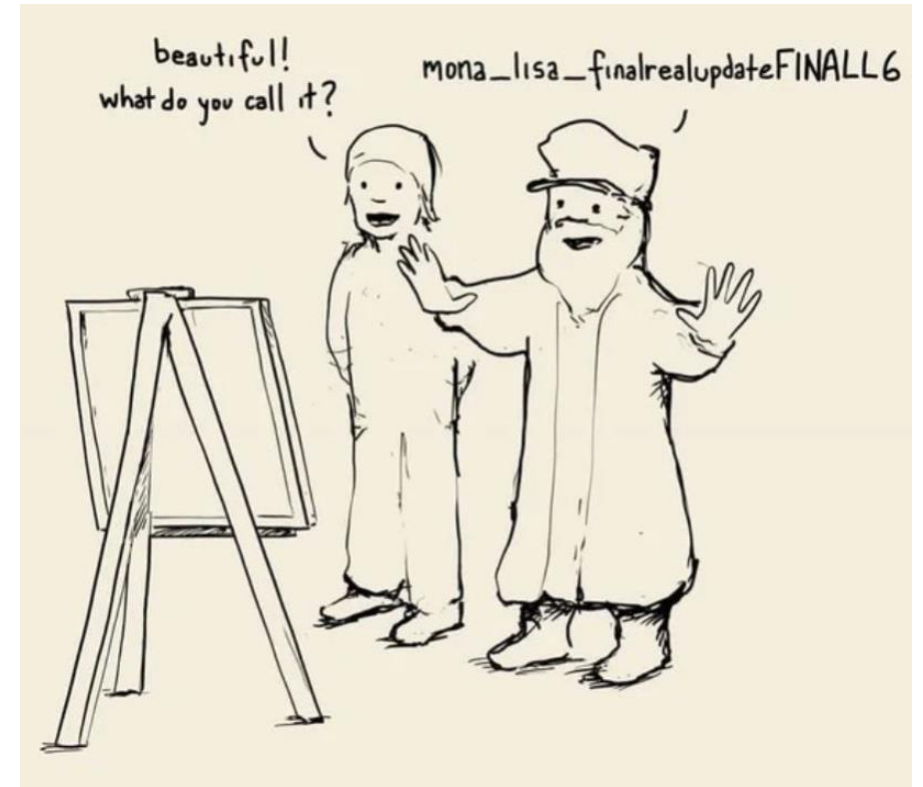
- Name your files wisely

- **Consistent** and **meaningful** to yourself and collaborators
- Allow for **easy tracking/searching** and be somewhat descriptive of content
- Make names **human readable** – name describes content of file
- Make names **machine readable** – Avoid dots, commas and special characters (e.g. ~ ! @ # \$ % ^ & * () ` ; < > ? , [] { } ' ")

[LD_phyA_off_t04_2020-08-12_norm.xlsx](#)

LD
the
phyA
genotype, in a
off
sucrose, at
t04
2020-08-12
norm

- Long day sampling, of
- Phytochrome A
- Medium without
- Time point 4,
- Sampled on Aug 12th, 2020, with
- Normalised data



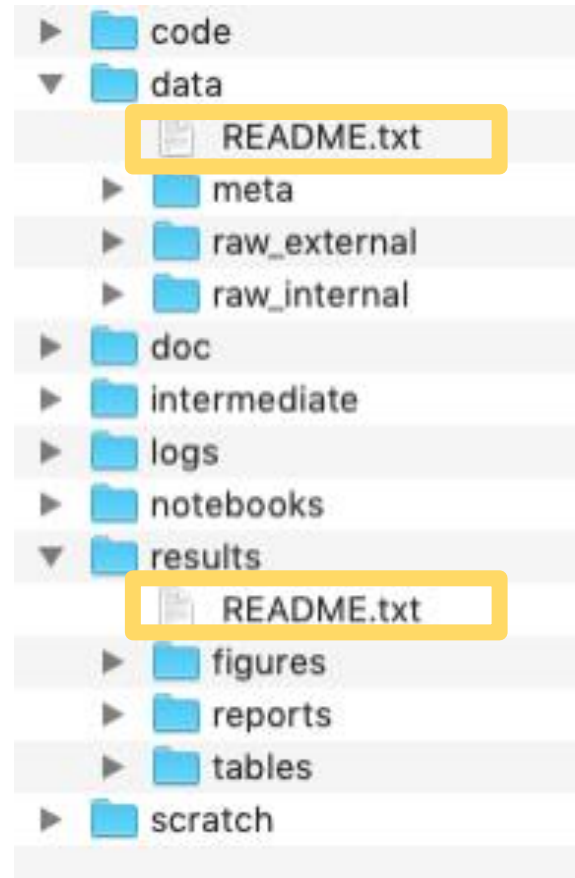
Credit: <https://twitter.com/nathanwpyle/status/1108902487203958784>

- Explain in **README** file!

How to organize your data



- Keep a structured directory environment



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*

How to organize your data



- Use README files

- Always include a README file:
 - **General information** - title, description etc
 - **Folder level** - explaining folder contents, naming, file history, organisation/structure etc
 - **Data info** - 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

This README file was generated on [YYYY-MM-DD] by [NAME]

GENERAL INFORMATION

- Dataset title:
- Description: <provide description of the dataset origin, steps used in its generation, content and its purpose>

ORGANIZATION

- Folder structure: similar to folder structure example above (below)
- File naming conventions: <provide explanation of the elements used, allowed values and examples>
- File formats: <Provide a list of all file formats present in this dataset>

DATA COLLECTION

- Institutional catalog ID (if applicable):
- Date of data collection: <provide single date, range, or approximate date; suggested format YYYY-MM-DD>
- Link to electronic lab book (codebook) where the following is described (if it does not exist, include it here):
 - Methods used for data collection (including references, documentation (e.g. consent form template), links):
 - Geographic location of collection (if applicable):
 - Experimental & environmental conditions of collection (if applicable):
 - Standards and calibration for data collection (if applicable):
 - Uncertainty, precision and accuracy of measurements (if applicable):
 - Known problems & caveats (sampling, blanks, etc.):
 - Codes or symbols used to record missing data with description (if applicable):
- Link to data dictionary:

DATA RE-USE

- DOI/accession number (if dataset is published):
- License (if any):
- Use restrictions (if any):
- Recommended citation for the data (if any):

How to organize your data



- Don't forget to backup

- Good to know for **backup planning** purposes:
 - Overall data volumes
 - Data life cycle in project
 - Ease of access
 - Data sensitivity



Credit: Photo by [Tobias Fischer](#) on [Unsplash](#)

How to organize your data



- Don't forget to backup

- Good to know for **backup planning** purposes:
 - Overall data volumes
 - Data life cycle in project
 - Ease of access
 - Data sensitivity
- On at least **two different kinds of media** (server, portable hard drive, cloud)



Credit: Photo by [Tobias Fischer](#) on [Unsplash](#)

How to organize your data



- **Don't forget to backup**

- Good to know for **backup planning** purposes:
 - Overall data volumes
 - Data life cycle in project
 - Ease of access
 - Data sensitivity
- On at least **two different kinds of media** (server, portable hard drive, cloud)
- Keep backup in **three separate locations**
 - Consider off-site backups



Credit: Photo by [Tobias Fischer](#) on [Unsplash](#)

How to organize your data



- Don't forget to backup

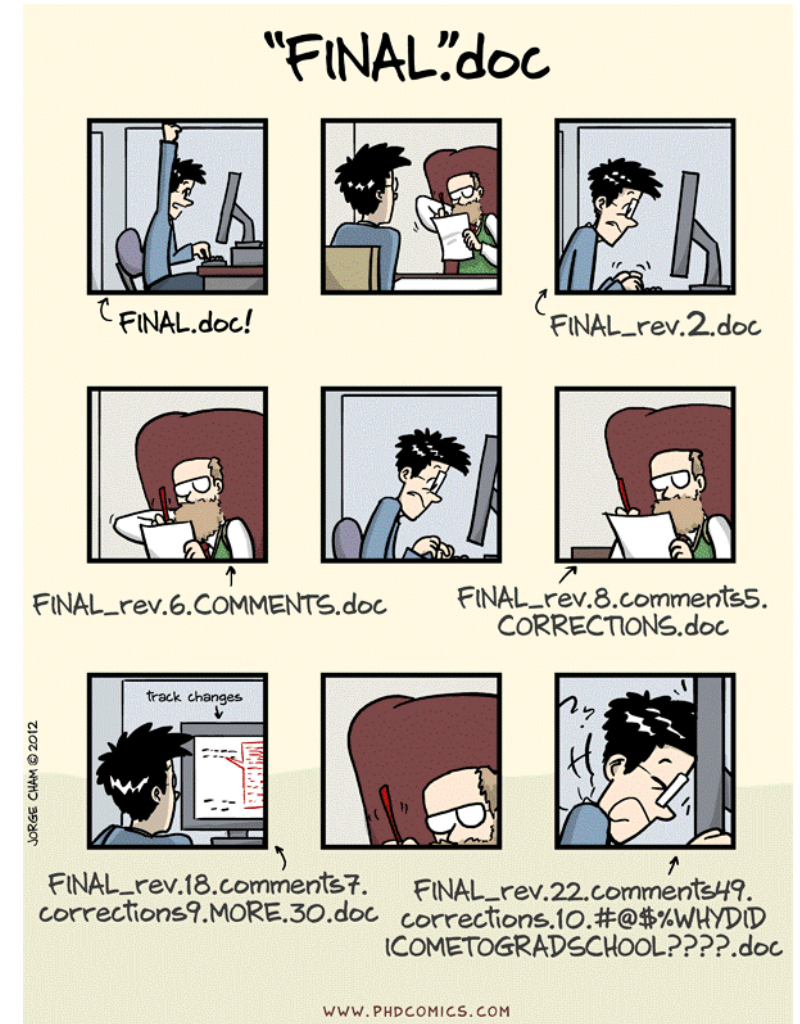
- Good to know for **backup planning** purposes:
 - Overall data volumes
 - Data life cycle in project
 - Ease of access
 - Data sensitivity
- On at least **two different kinds of media** (server, portable hard drive, cloud)
- Keep backup in **three separate locations**
 - Consider off-site backups
- Robust backups need to be **automated**



Credit: Photo by [Tobias Fischer](#) on [Unsplash](#)

Versioning of data and code

- Make changes to files and sleep well at night



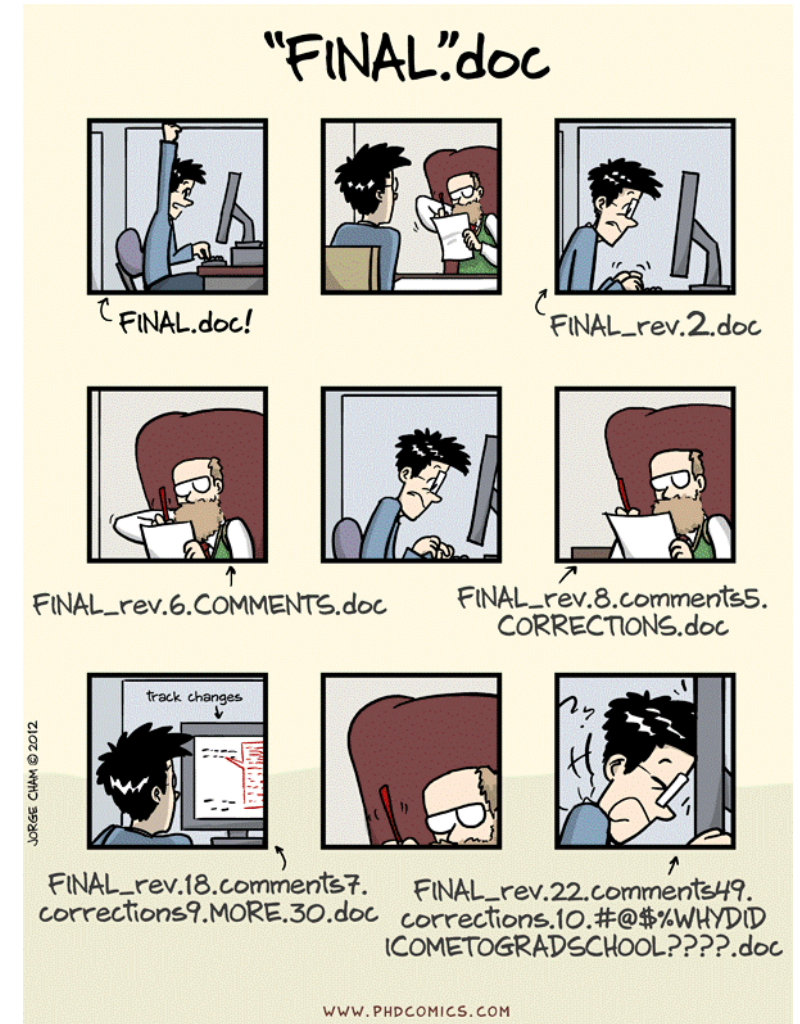
"Piled Higher and Deeper" by Jorge Cham,
<https://phdcomics.com>

Versioning of data and code

- Make changes to files and sleep well at night



- Making changes to files
 - We risk losing content.
 - Unintended side effects.
 - Breaking analysis pipelines.



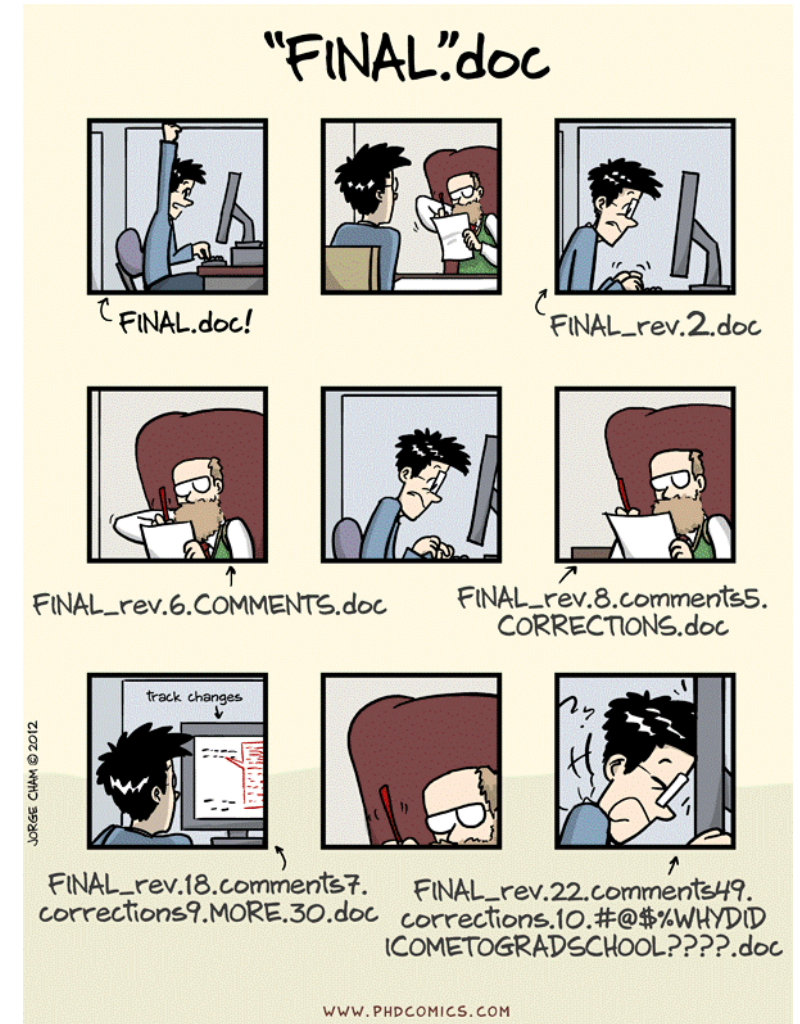
"Piled Higher and Deeper" by Jorge Cham,
<https://phdcomics.com>

Versioning of data and code

- Make changes to files and sleep well at night



- **Making changes to files**
 - We risk losing content.
 - Unintended side effects.
 - Breaking analysis pipelines.
- **Collaborating with others**
 - Coordination between multiple devices.
 - Resolve conflicts.



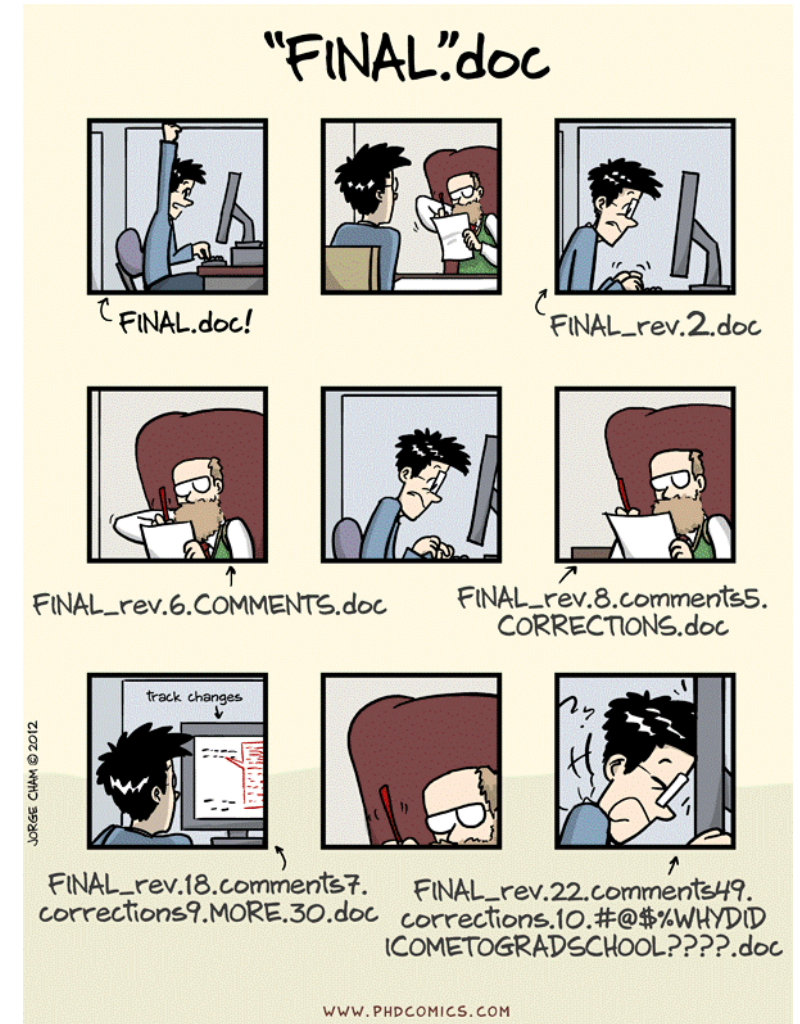
"Piled Higher and Deeper" by Jorge Cham,
<https://phdcomics.com>

Versioning of data and code



- Make changes to files and sleep well at night

- **Making changes to files**
 - We risk losing content.
 - Unintended side effects.
 - Breaking analysis pipelines.
- **Collaborating with others**
 - Coordination between multiple devices
 - Resolve conflicts.
- Addressing these issues by tracking changes is called **version control**.



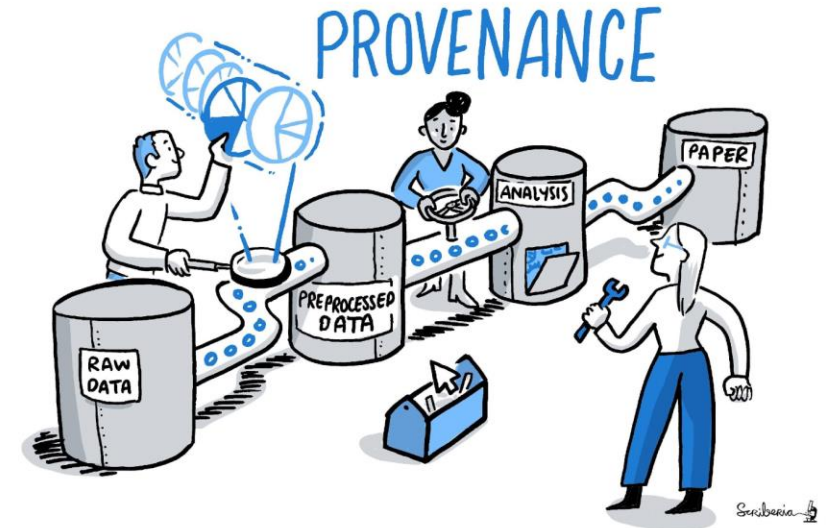
"Piled Higher and Deeper" by Jorge Cham,
<https://phdcomics.com>

Version Control Systems (VCS)



- A systematic and organized approach to preserve the history of changes

- Tracking changes
 - 100% accurate record of what was *actually* changed.



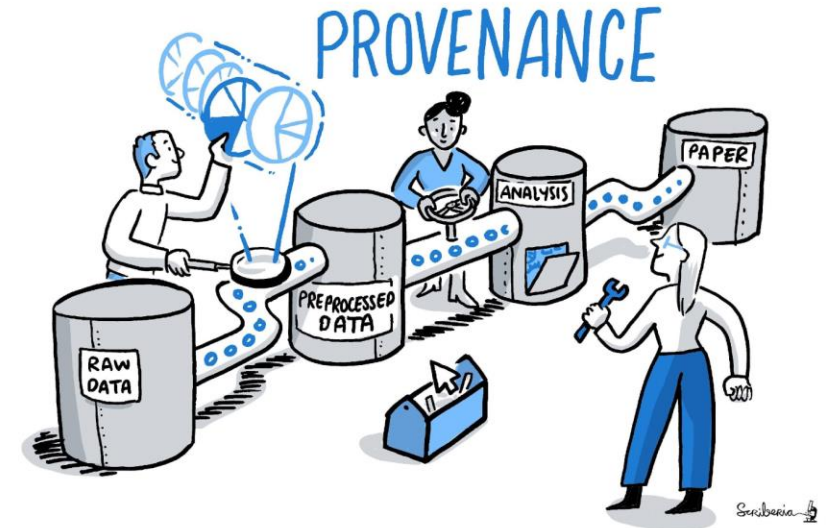
Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Version Control Systems (VCS)



- A systematic and organized approach to preserve the history of changes

- **Tracking changes**
 - 100% accurate record of what was *actually* changed.
- **Collaborative workflow**
 - Multiple users simultaneously, mechanism for resolving conflicts.



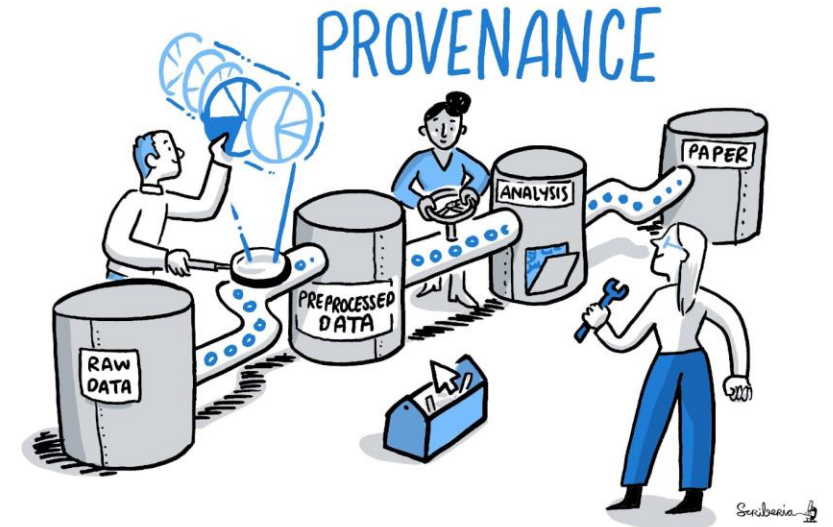
Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Version Control Systems (VCS)



- A systematic and organized approach to preserve the history of changes

- **Tracking changes**
 - 100% accurate record of what was *actually* changed.
- **Collaborative workflow**
 - Multiple users simultaneously, mechanism for resolving conflicts.
- **Backup and Recovery**
 - Stores only the necessary information to recreate previous versions of files on demand.



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

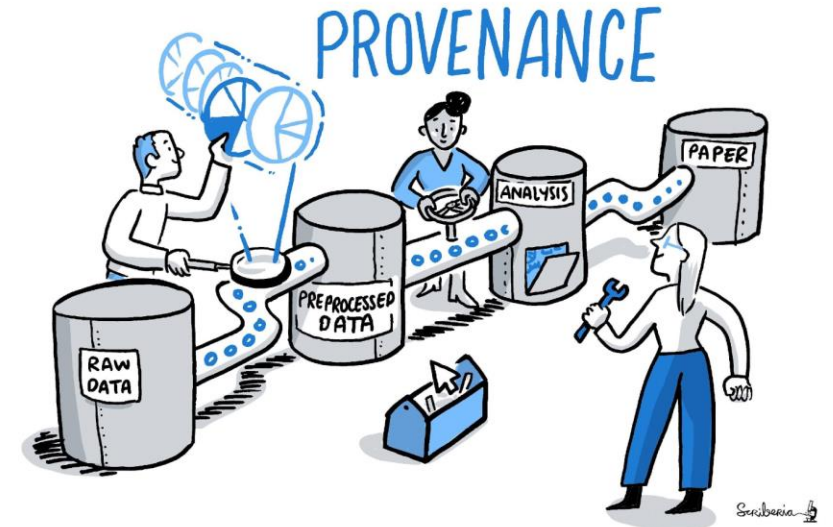
Version Control Systems (VCS)



- A systematic and organized approach to preserve the history of changes

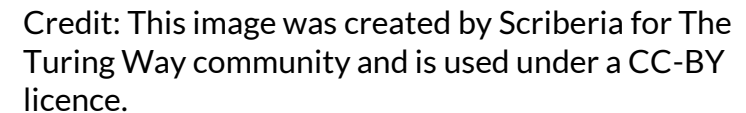
- **Historical insight**

- Users can review and revert to previous versions of files.



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

- **Historical insight**
 - Users can review and revert to previous versions of files.
- **Branching and Merging**
 - New branches can be created for separate work, which can later be merged back.



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

Version Control Systems (VCS)

- **A systematic and organized approach to preserve the history of changes**

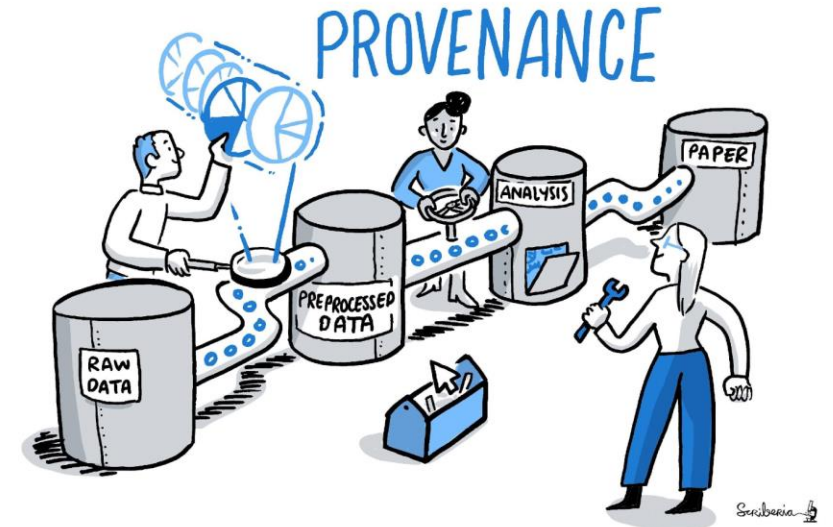
- **Historical insight**

- Users can review and revert to previous versions of files.

- **Branching and Merging**

- New branches can be created for separate work, which can later be merged back.

- Manage all versions of files along with useful metadata (author, timestamp, unique identifier).



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

Git and GitHub

- **Simplifying code management and collaboration**
-



Git and GitHub

- Simplifying code management and collaboration
-



- **What is Git?**
 - Version control system software.
 - Can be installed locally on your computer.



Jason Long, CC BY 3.0, via Wikimedia Commons

Git and GitHub

- Simplifying code management and collaboration



- **What is Git?**

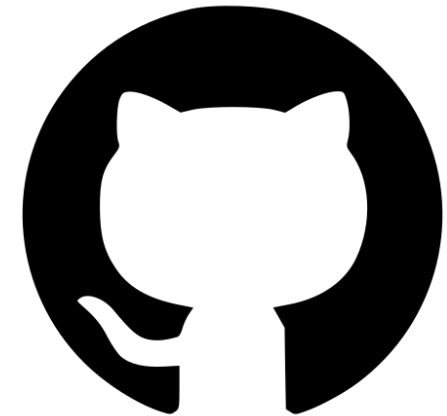
- Version control system software.
- Can be installed locally on your computer.



Jason Long, CC BY 3.0, via Wikimedia Commons

- **What is GitHub?**

- Web-based hosting service for Git repositories.
- Provides a platform for collaborative software development.



GitHub, CC BY 4.0, via Wikimedia Commons

Git and GitHub



- Simplifying code management and collaboration

- **What is Git?**

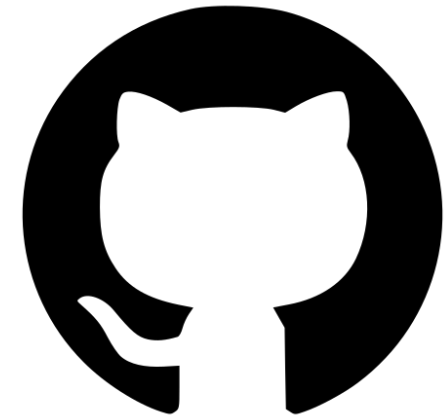
- Version control system software.
- Can be installed locally on your computer.



Jason Long, CC BY 3.0, via Wikimedia Commons

- **What is GitHub?**

- Web-based hosting service for Git repositories.
 - Provides a platform for collaborative software development.
- Many use Git and GitHub for **more than just code**
 - Manuscripts.
 - Course material (teaching).
 - Websites.



GitHub, CC BY 4.0, via Wikimedia Commons

Take a leg-stretcher (5 minutes)

Workflow management systems

- A way to streamline your bioinformatics data analysis



Workflow management systems



- **A way to streamline your bioinformatics data analysis**
- Software that sets up, performs, and monitors a **defined sequence of computational tasks** (i.e. “a workflow”)
 - to minimize redundancies and automate repetitive tasks in data analysis



Workflow management systems

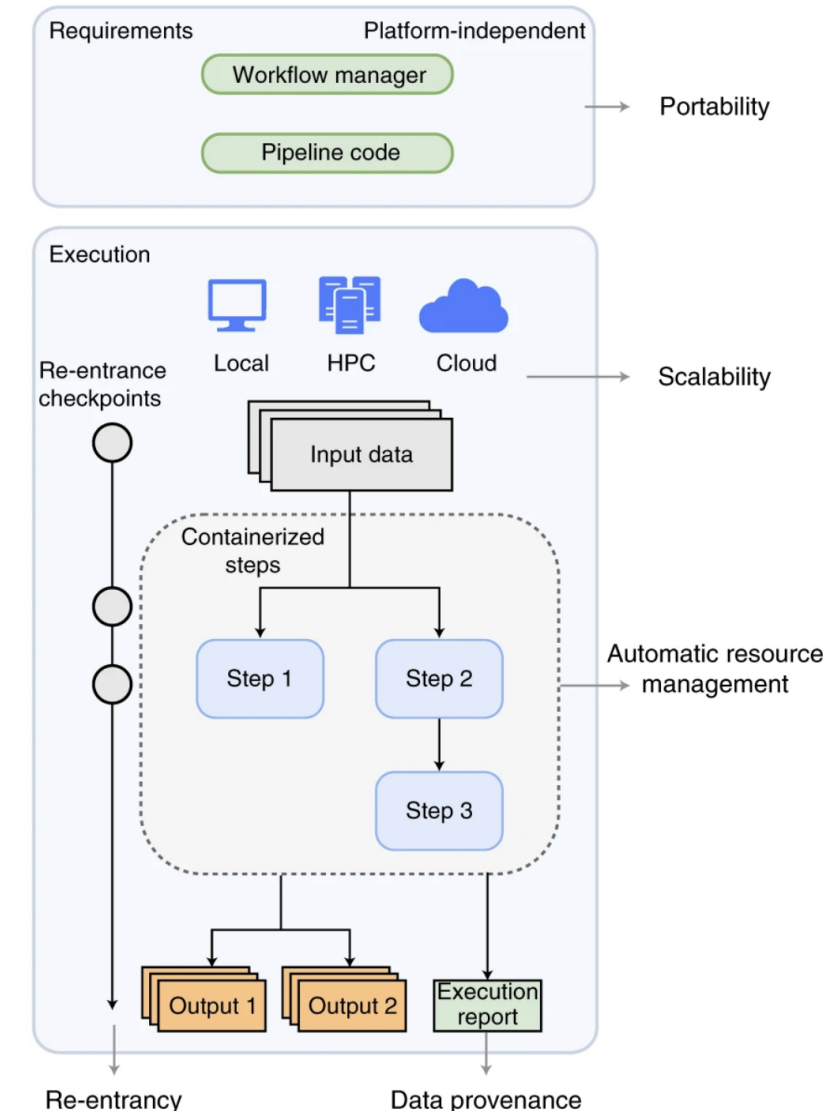


- A way to streamline your bioinformatics data analysis

- Software that sets up, performs, and monitors a **defined sequence of computational tasks** (i.e. “a workflow”)
 - to minimize redundancies and automate repetitive tasks in data analysis
- It makes **data analyses reproducible and scalable**
 - facilitates keeping track of which files have been processed in what way throughout an entire project, and which software was used



Image from: Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* 18, 1161–1168 (2021). <https://doi.org/10.1038/s41592-021-01254-9>



Workflow management systems



- **A way to streamline your bioinformatics data analysis**

- Some of the most common WMSs within the bioinformatic and academic communities are:

- **nextflow** (<https://www.nextflow.io/index.html>)

-  **snake**make (<https://snakemake.readthedocs.io/en/stable/>)

Workflow management systems



- A way to streamline your bioinformatics data analysis
- If you want to learn more about WMS and other tools for reproducible research, check out the NBIS future courses webpage: <https://nbis.se/training/future>

The screenshot shows the NBIS SciLifeLab website with the following content:

NBIS SciLifeLab | About us | Services | Training | Contact | Search...

Future courses at NBIS

See our [course catalogue](#) or [ELIXIR TeSS](#) for more training events across Europe.

Course Title	Dates	Deadline	Location	More info	Homepage	Apply
DDLs Population genomics in practice	2023-11-06 to 2023-11-10 (5 days)	2023-09-30	Uppsala	► More info	Homepage	Apply
Introduction to bioinformatics using NGS data	2023-11-13 to 2023-11-17 (5 days)	2023-10-04	Uppsala	► More info	Homepage	Apply
NBIS/ELIXIR-SE Tools for Reproducible research	2023-11-20 to 2023-11-24 (5 days)	2023-10-20		► More info	Homepage	Apply
Snakemake BYOC (bring-your-own-code) workshop	2023-12-04 to 2023-12-06 (3 days)	2023-10-29	Online	► More info	Homepage	Apply

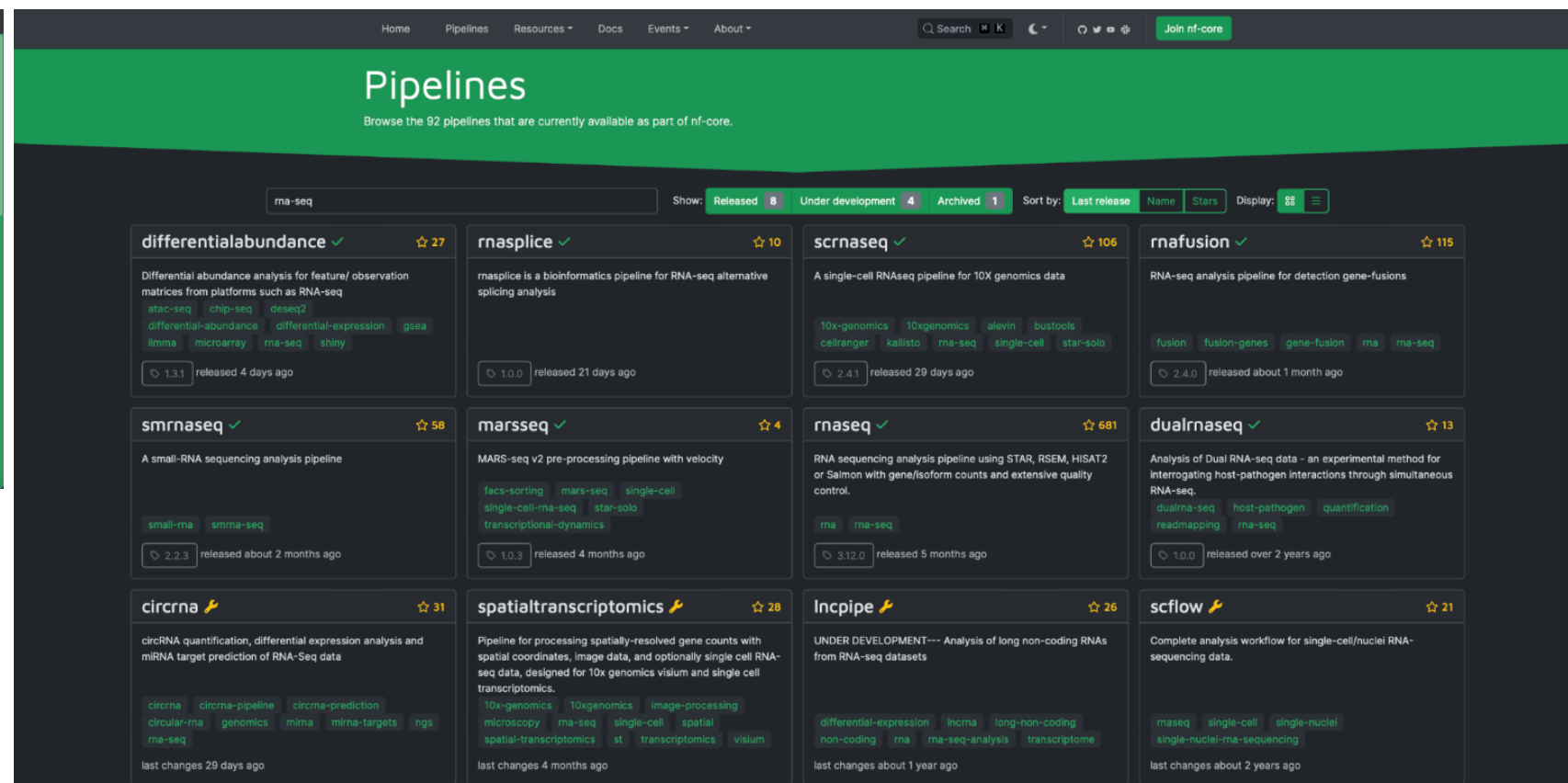
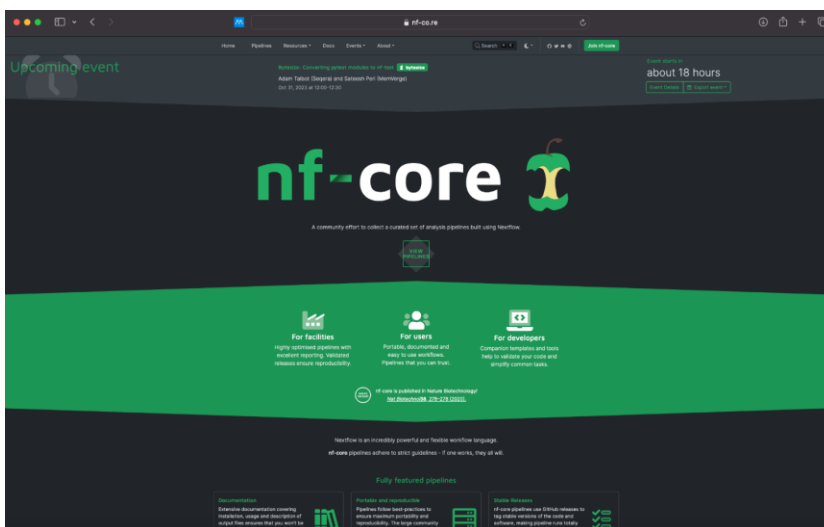
[Privacy](#) - [Terms](#)

Workflow management systems



- A way to streamline your bioinformatics data analysis

- **nf-core:** A collection of community-curated analysis pipelines built using Nextflow (<https://nf-co.re/>)

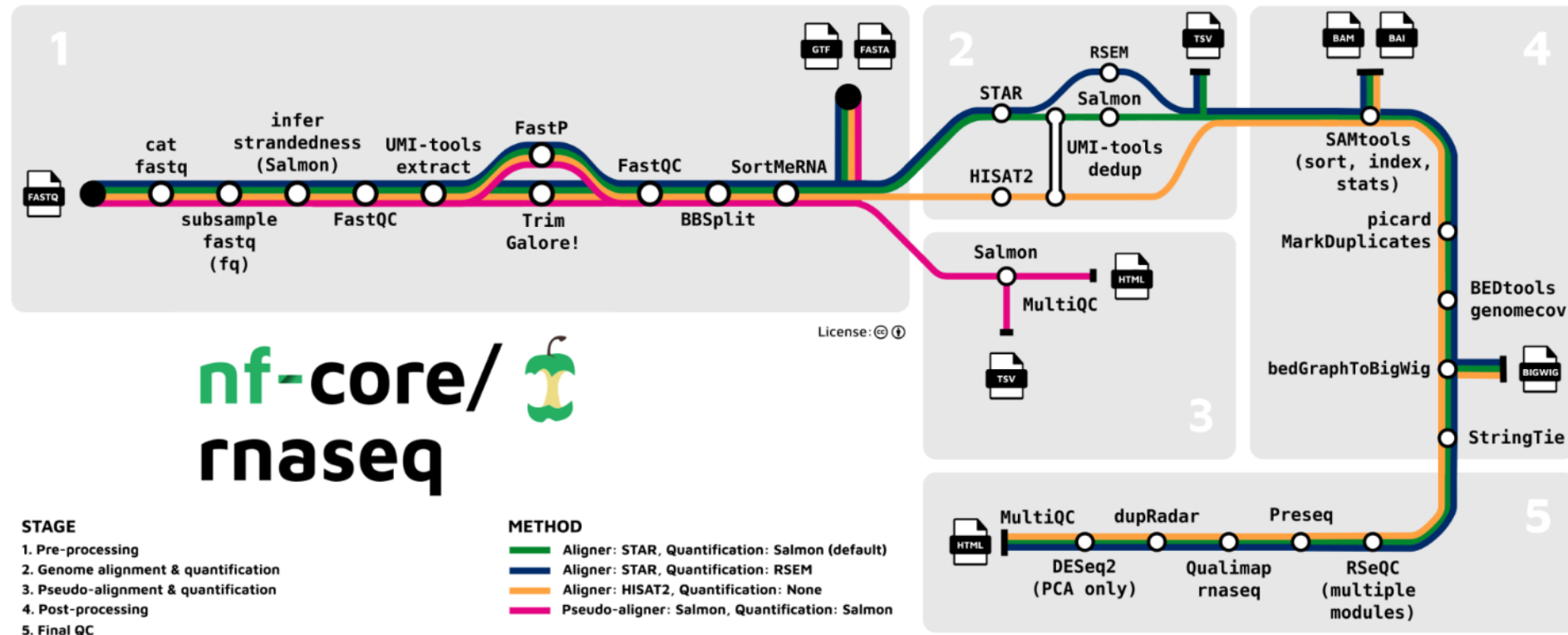


Workflow management systems



- A way to streamline your bioinformatics data analysis

- Example: RNA-seq pipeline



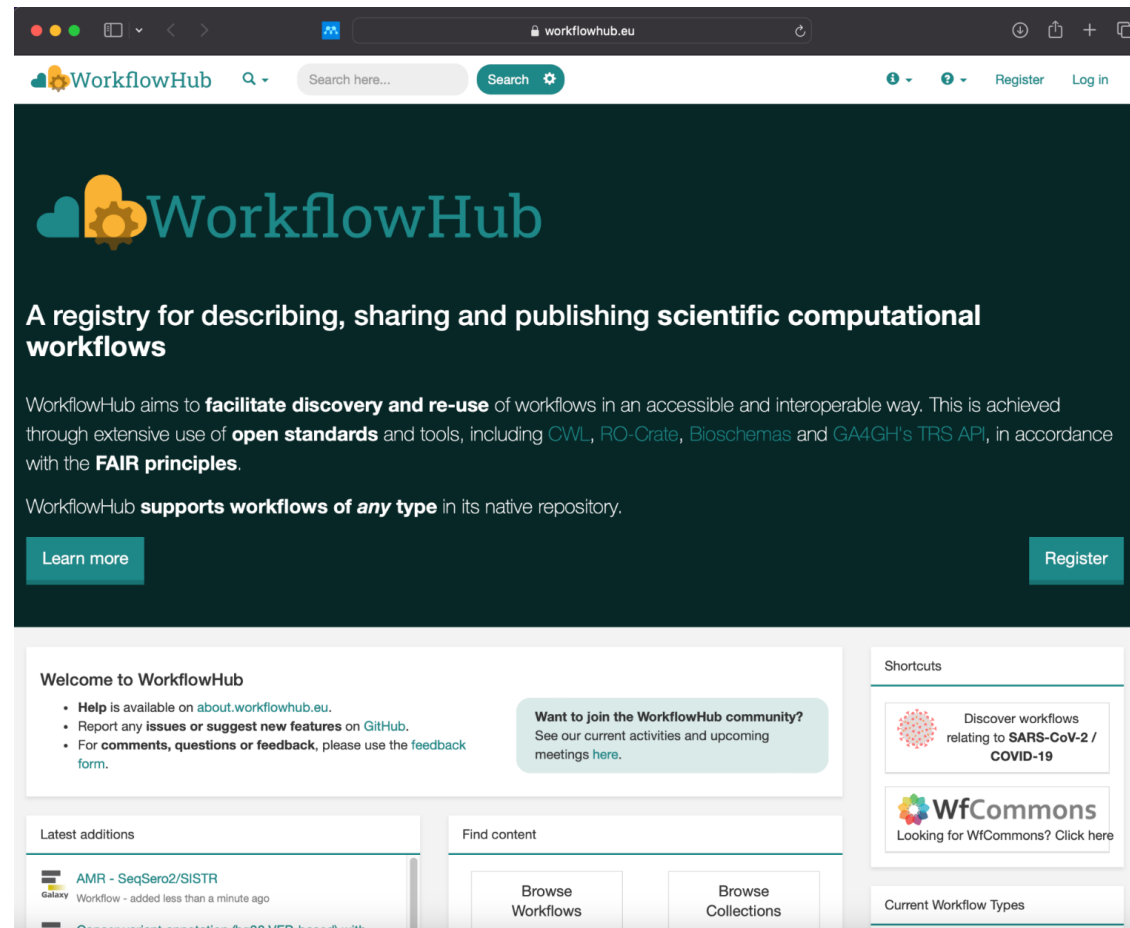
- Generates interactive html reports: <https://nf-co.re/rnaseq/3.12.0/docs/output>

Workflow management systems



- A way to streamline your bioinformatics data analysis

- Publish and obtain a DOI for your data analysis workflow using **Workflow Hub** (<https://workflowhub.eu/>)

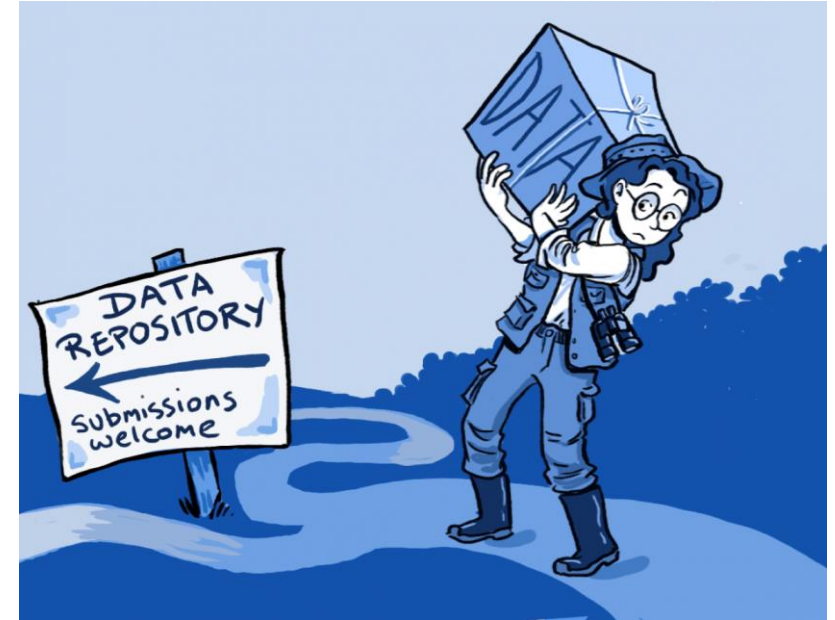


Publishing data - share and reuse phase



Why submit to a repository?

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

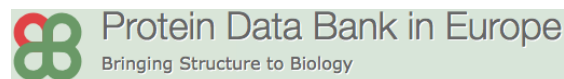
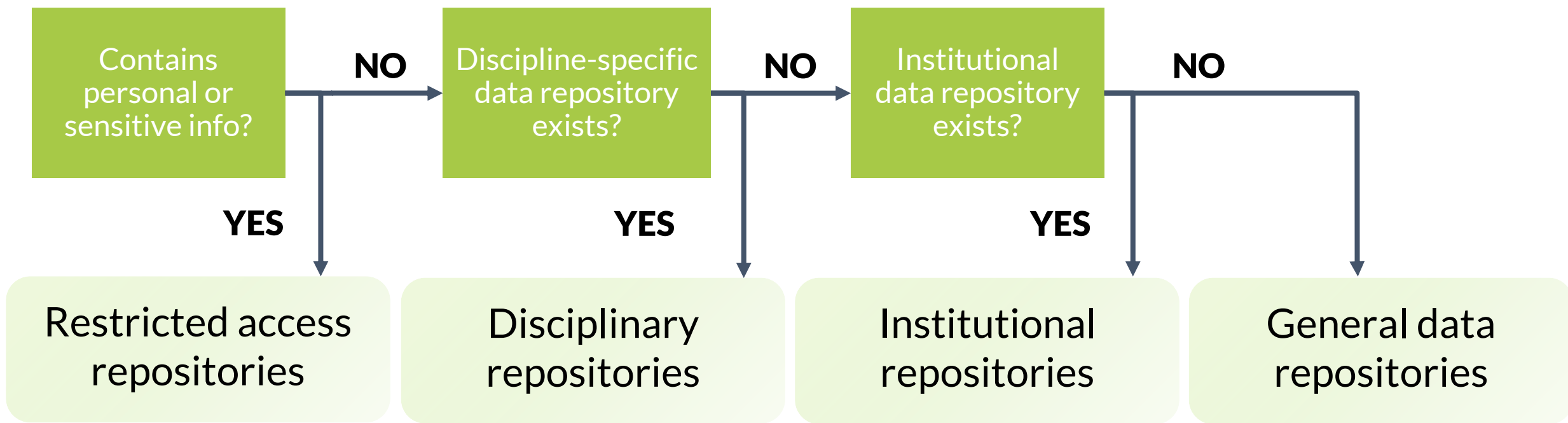


Credit: <http://www.openaire.eu/blogs/research-data-management-rdm-support-at-the-university-of-vienna> License: CC ATTRIBUTION 4.0 INTERNATIONAL





Selecting a data repository



How do you find a domain specific repository?



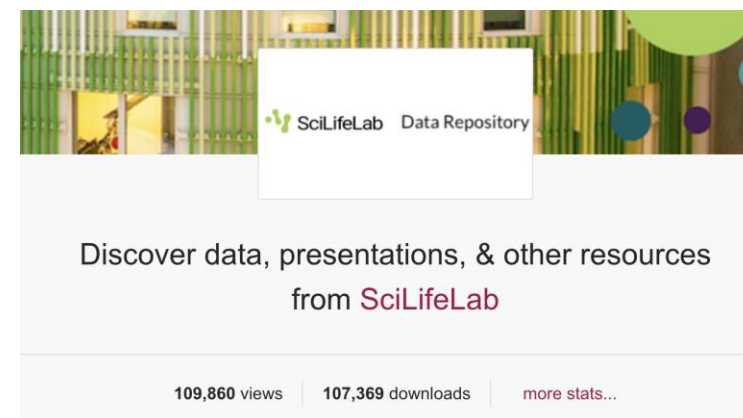
- [EBI wizard](#) - guide depending on data type
- [ELIXIR Deposition Databases for Biomolecular Data](#) - deposition database list
- [FAIRsharing.org/databases](#) - catalogue of many repositories, with possibility to filter on e.g. domain

SciLifeLab Data Repository

- Supporting data publishing: figshare.scilifelab.se



- SciLifeLab instance of [Figshare](https://figshare.com)
- **Available for researchers** affiliated to all Swedish universities and institutes working **within SciLifeLab areas of activity**
- Examples of item types that can be uploaded: **dataset, DMP, software, figures, poster, presentation, educational resource**
 - Data is made citable through its DOI
- **Publish** your data **as early as possible**, and as late as necessary
 - **Embargo** and **restricted access** can be used

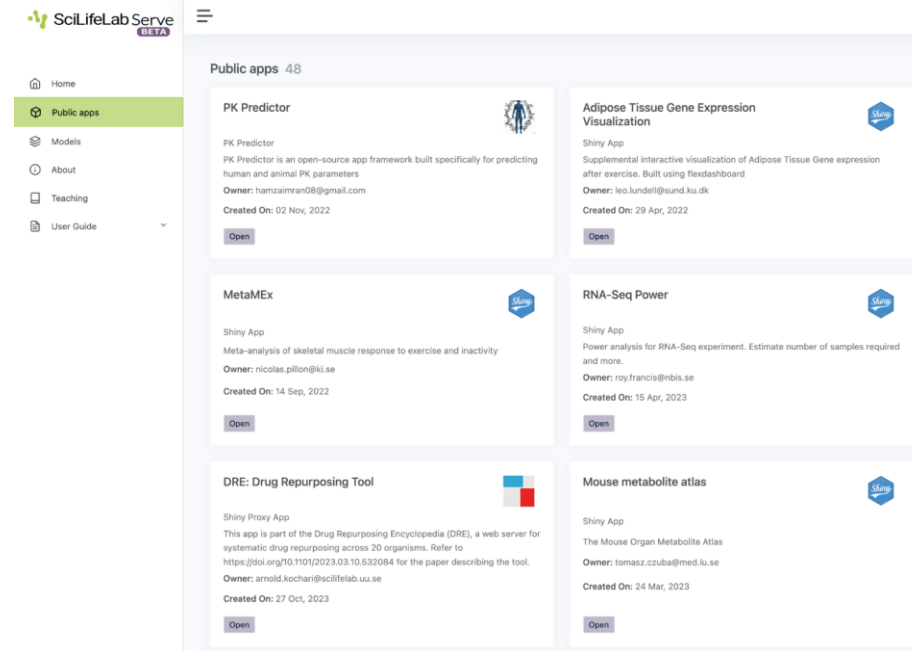


SciLifeLab Serve

- Hosting ML models and applications



- **Machine learning model serving**
 - Serving of trained machine learning models using dedicated tools – e.g. **PyTorch Serve**, **Tensorflow Serving**, **MLFlow Serve**, etc.
- **Application hosting**
 - Host applications such as **Streamlit**, **Gradio**, **Flask**, **Django**, etc. with user interfaces and APIs for ML models built on various frameworks to allow others **to explore your data or results**, or **to provide tools** based on your deployed machine learning models.
- **Interactive development environments (IDEs)**
 - Use a web-browser based **Jupyter Lab** or **RStudio** instance to run analyses or to **collaborate with your team**. Serve also accept requests for **use in teaching**.
- Use case example: **RNASeq Power**
<https://r8844d6ec.serve.scilifelab.se/>
- Dedicated team offers consultations, support, and training.
Contact: serve@scilifelab.se



Thank you!



- Contact us throughout the entire data life cycle!
<https://data-guidelines.scilifelab.se/> or data-management@scilifelab.se
- Monthly event (hybrid Uppsala, Solna & Zoom):
“Meet a Data Steward and get Data Management support”
- Upcoming virtual event together with UU and UMU Libraries:
“Data Management Plans in practise – research funder perspectives and practical demos”

When: November 22, 2023 10-12

Who: Open to researchers, staff and interested in all scientific disciplines

What: Listen to the **Funders Perspective** (VR, Riksbankens Jubileumsfond, Forte and KAW) on Data Management Plans (DMPs) during the first hour and get a **demo of two DMP tools DMPonline and Data Stewardship Wizard (DSW)** during the 2nd hour

Where: Online, Zoom

Deadline to sign up: November 22, at 10.

<https://www.scilifelab.se/event/data-management-plans-in-practise/>

