ETH zürich

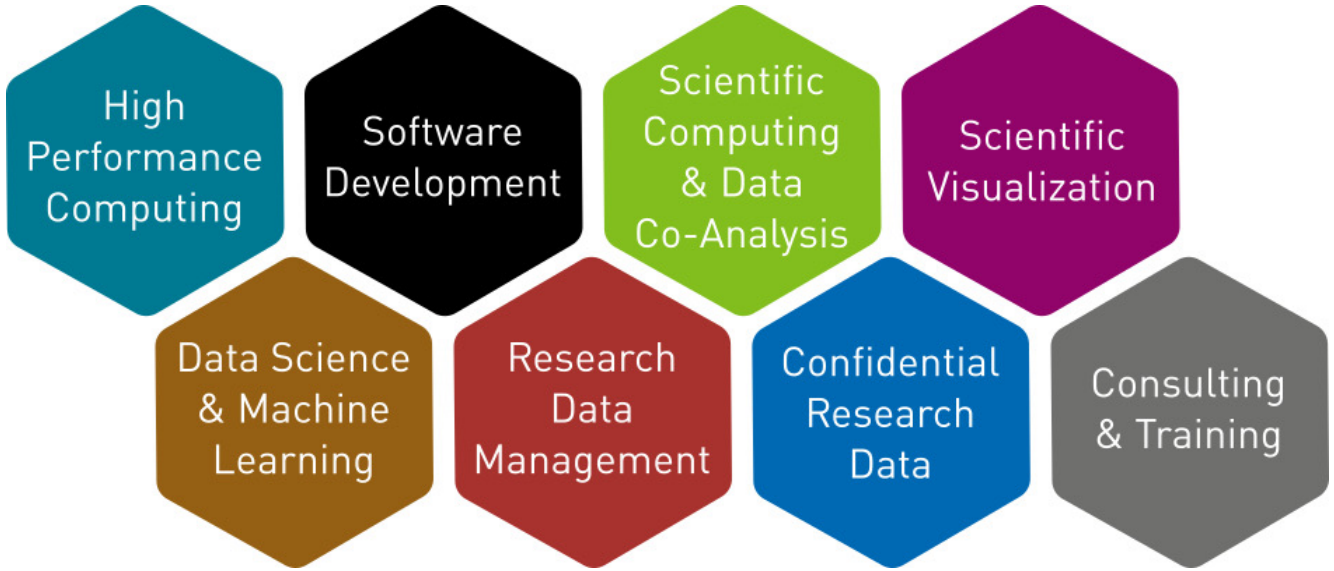
IT Services

Introduction to Active Research Data Management

Caterina Barillari, Henry Lütcke
Scientific IT Services, ETH Zurich

October 26, 2022

Who is Scientific IT Services?



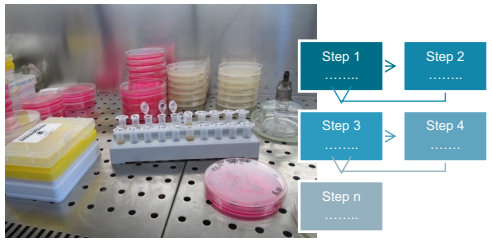
- ❑ A section of ETHZ IT Services
- ❑ Over 40 experts in various areas of scientific computing
- ❑ With a background in different areas of science

Tell us a bit about yourselves

- ❑ Your affiliation
- ❑ Your research topic

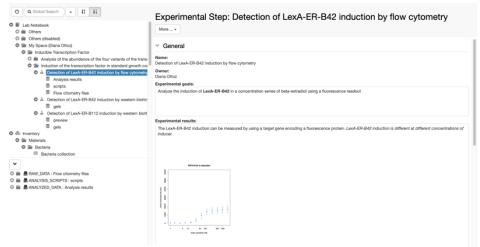
Overview of today's workshop

ARDM intro



Sample & protocol management

Management of Data & Metadata

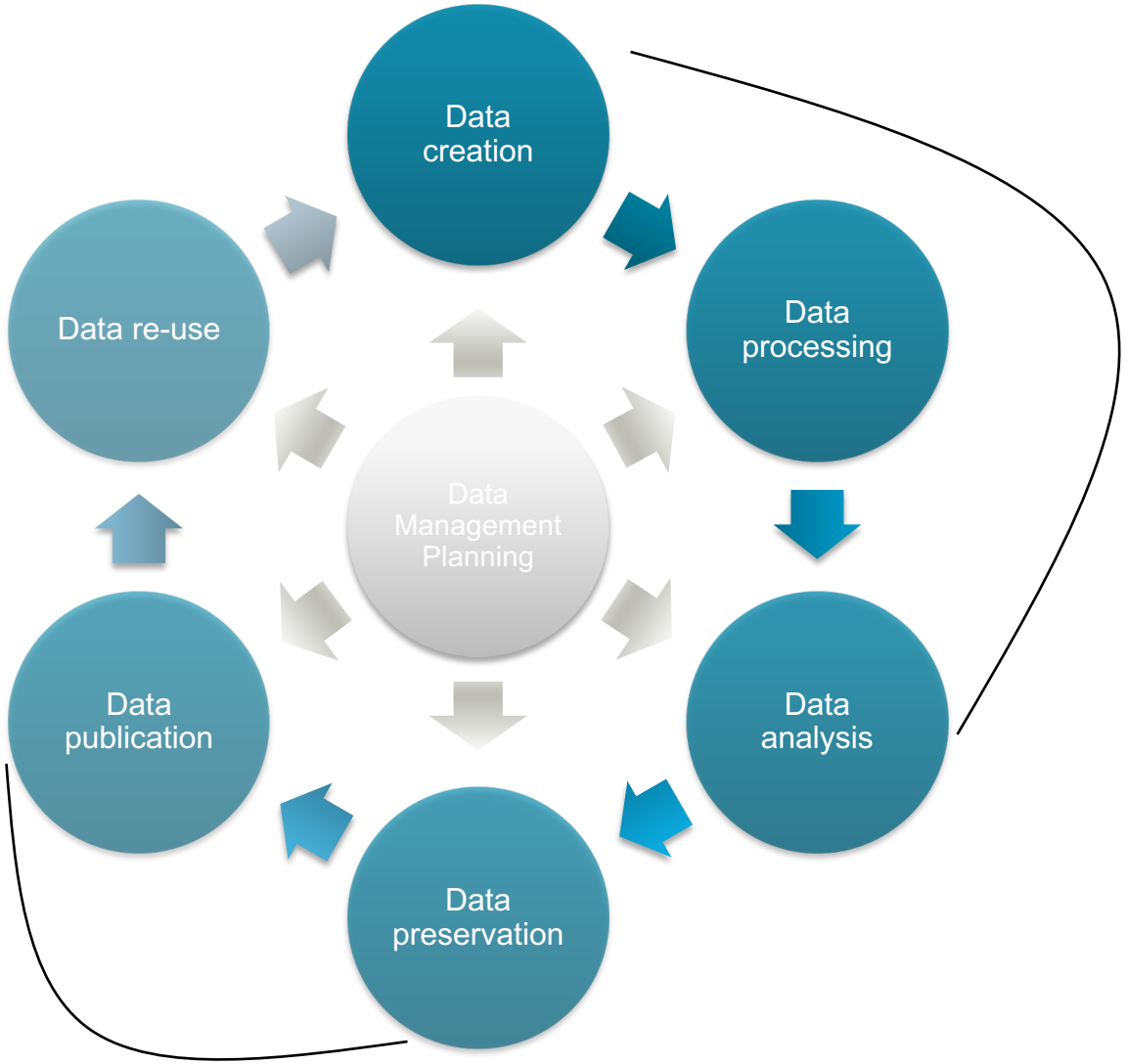


ELNs

ETHZ SIS's RDM solution

Overview of active research data management

The data life cycle



Active data management:
Annotate, store, backup data while it is produced

Long term preservation:
Annotate, store, backup data at the end of a project or after publication

Open Science and Open Research Data

Open Research Data

Requirement from funding agencies, journals, academic institutions

FAIR data

Persistent Identifier; metadata

Open protocol

Data format

Metadata

How can we share data in a FAIR way?

A few generic [data repositories recommended by SNSF]

ETH zürich Research Collection

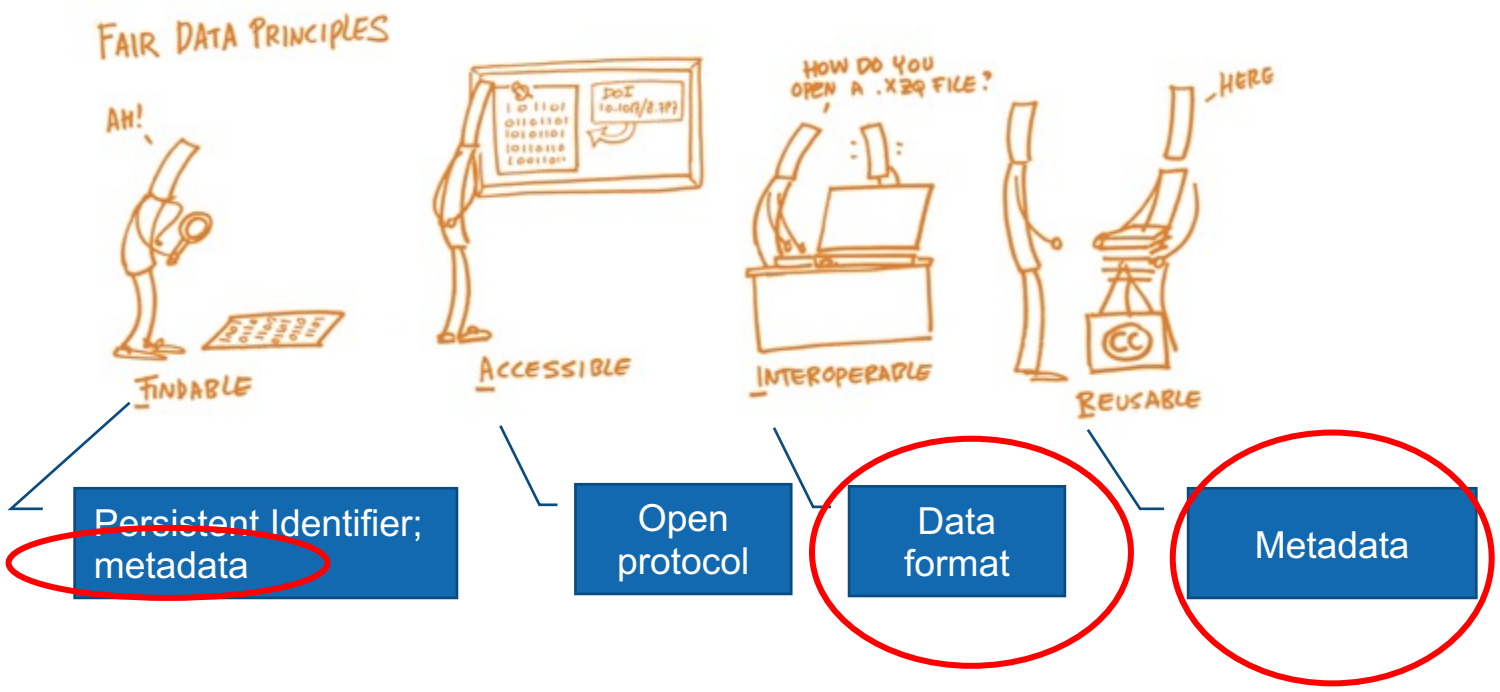
OSF

zenodo

DRYAD

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

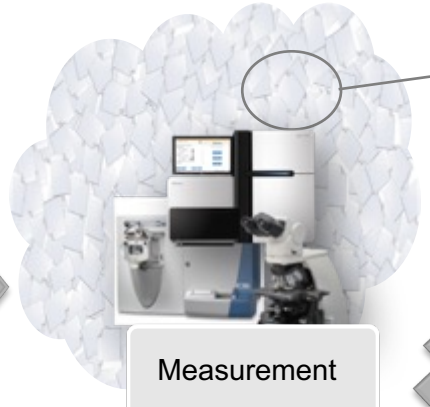
Prepare to meet FAIR requirements when data are generated



Data and information generation during a research project

Sample preparation

- Sample info
- Sample preparation info (protocol)

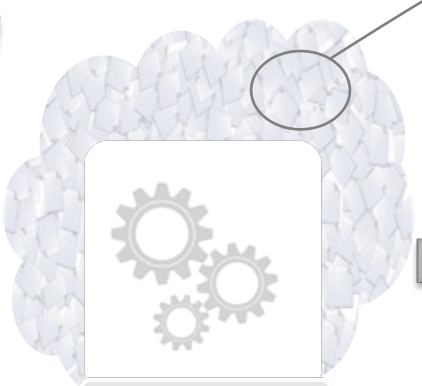


Measurement

- Machine info
- Machine settings

Data collection

- Data collection info

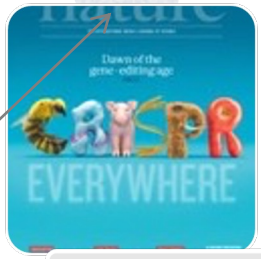


Processing

- Code/programs
- Settings/Procedures

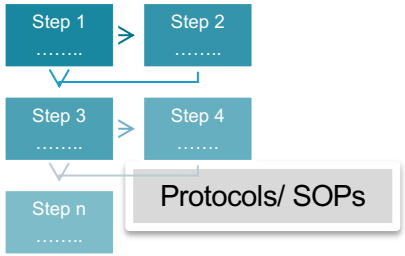
Analysis

- Code/programs
- Settings/Procedures



Publication

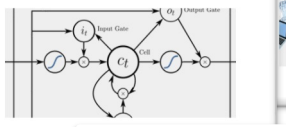
The “*data spread*”: a common scenario in academic institutions



Protocols/ SOPs

Materials/samples

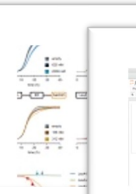
Code

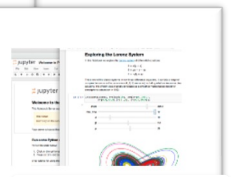


Model

Raw data

Processed data



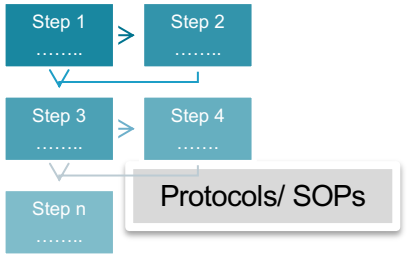


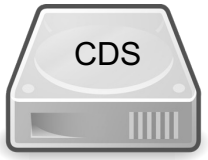
Results

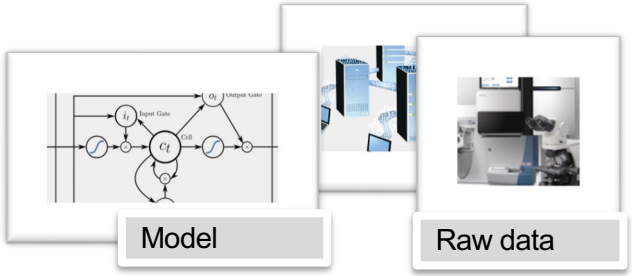
Analysis notebooks

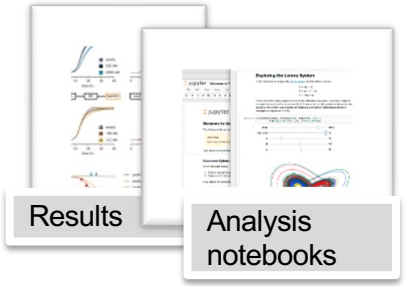
Experimental description/notes

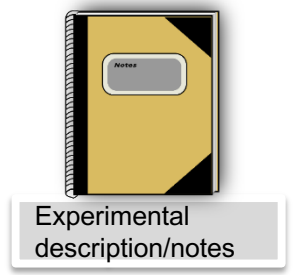
The “*data spread*”: a common scenario in academic institutions



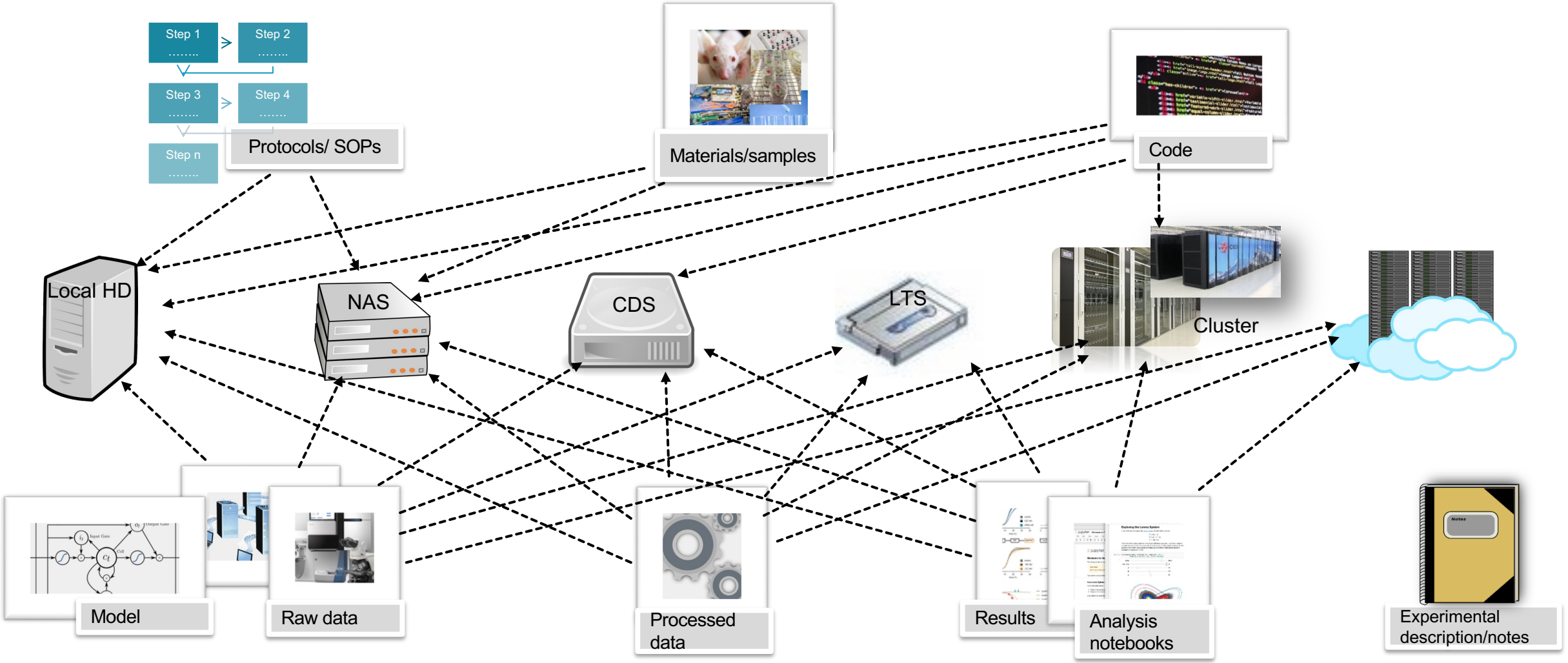




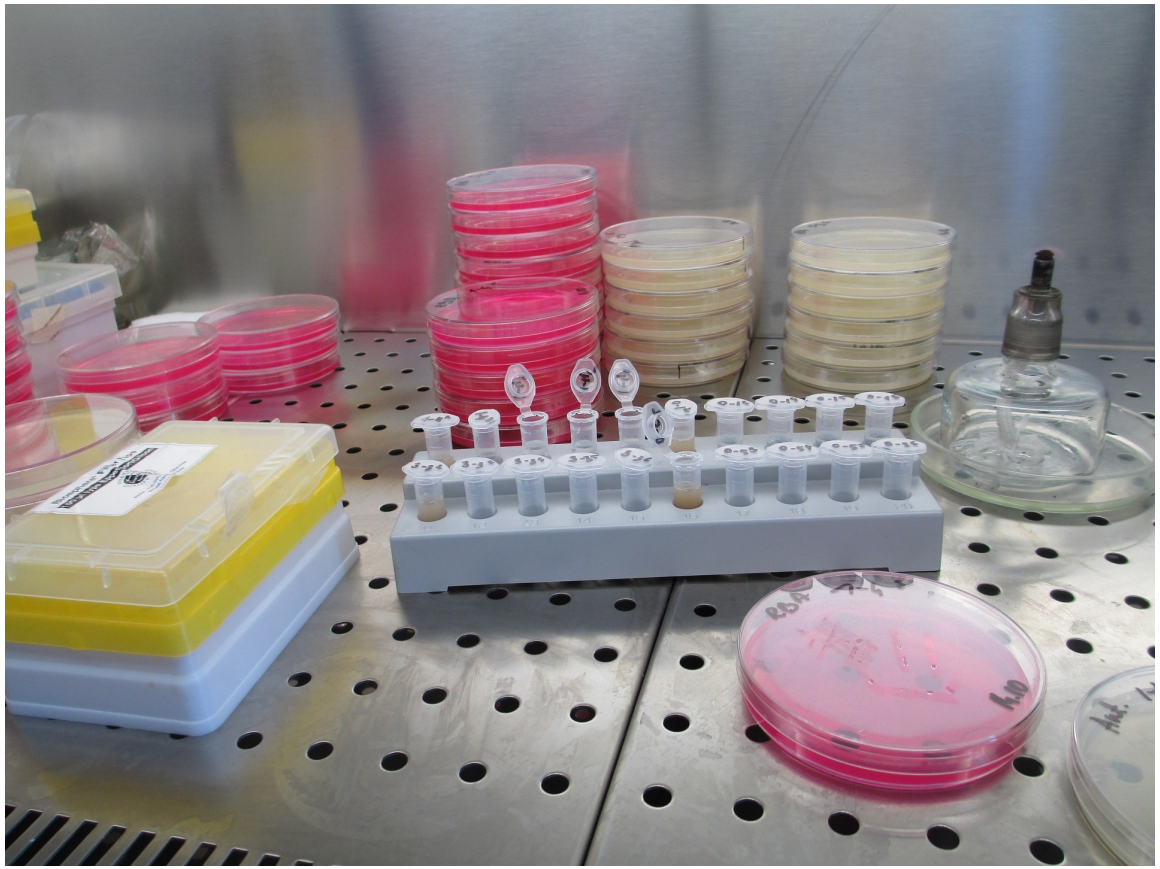


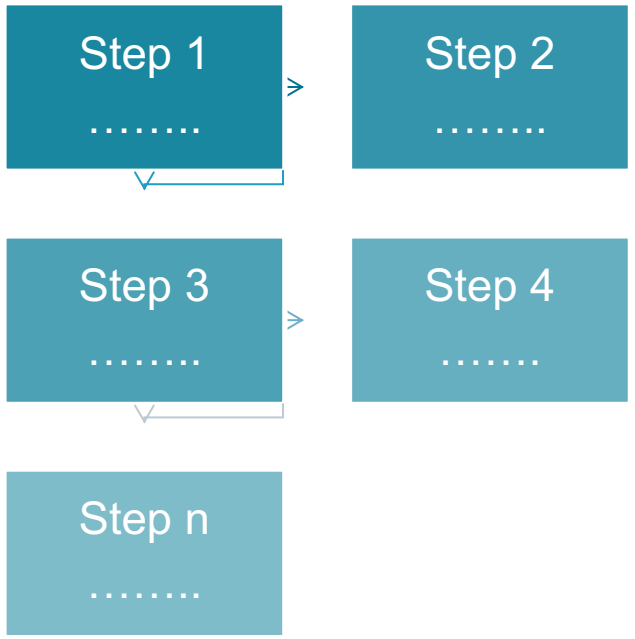


The “*data spread*”: a common scenario in academic institutions



How can we take care of the individual components and how can we bring things together?





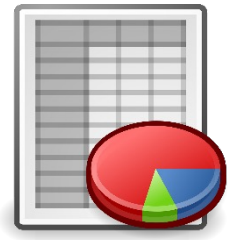
Management of samples and protocols

Management of materials and samples

What?

- ❑ Biological samples
- ❑ Chemical samples
- ❑ Materials
- ❑ Devices
- ❑ ….

How?



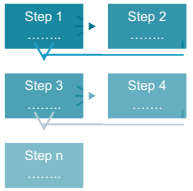
Spreadsheets

- *Not scalable*
- *No sharing*
- *No efficient search*
- *Easy to use*

Database/ LIMS

- *Scalable*
- *Sharing*
- *Search functionality*
- *Require time for set up and maintenance*

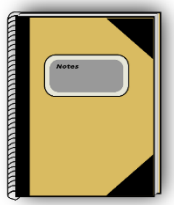
Management of protocols



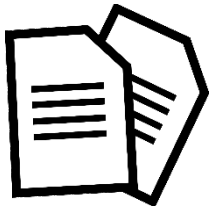
What?

- ❑ Step by step description of procedure
- ❑ Experimental/computational parameters (e.g. temperature, time, etc.)
- ❑ Machine used (experimental)
- ❑ OS, program, version, etc. (calculation)

How?



Paper notebook



Text files

- *Not scalable*
- *No sharing*
- *No efficient search*
- *Easy to use*

WIKI

- *Scalable*
- *Sharing*
- *Search functionality*
- *Versioning*

Database/ LIMS

- *Scalable*
- *Sharing*
- *Search functionality*
- *Require time for set up and maintenance*

Laboratory Information Management System (LIMS)

- LIMS are software for managing laboratory operations:
  - **sample tracking** (*info about samples and about their storage*)
  - **sample data tracking** (*upload of data measured from samples, e.g. sequencing data, NMR, MS, etc.*)
  - **protocol management** (*info about sample preparation/handling*)

- LIMS were first used in companies for tracking the growing number of samples.

- LIMS were originally stand-alone solutions, which had to be integrated with other solutions such as ELNs. Nowadays **ELNs and LIMS are often combined** in one platform.

Wikis at ETH

- ETH ITS provides wiki services:

Confluence

https://ethz.ch/services/en/it-services/catalogue/web-application-hosting/wiki.html

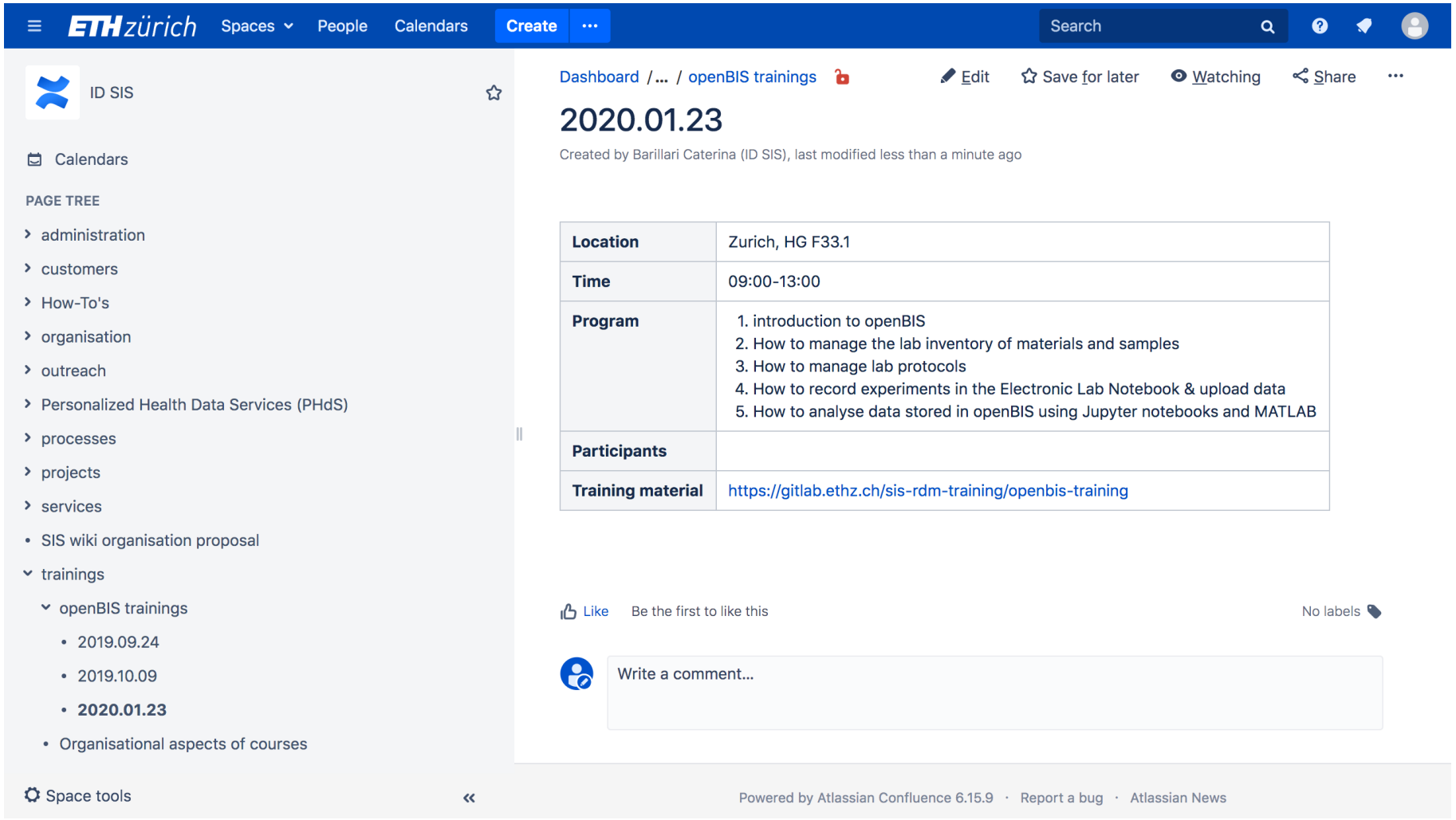
GitLab

https://gitlab.ethz.ch

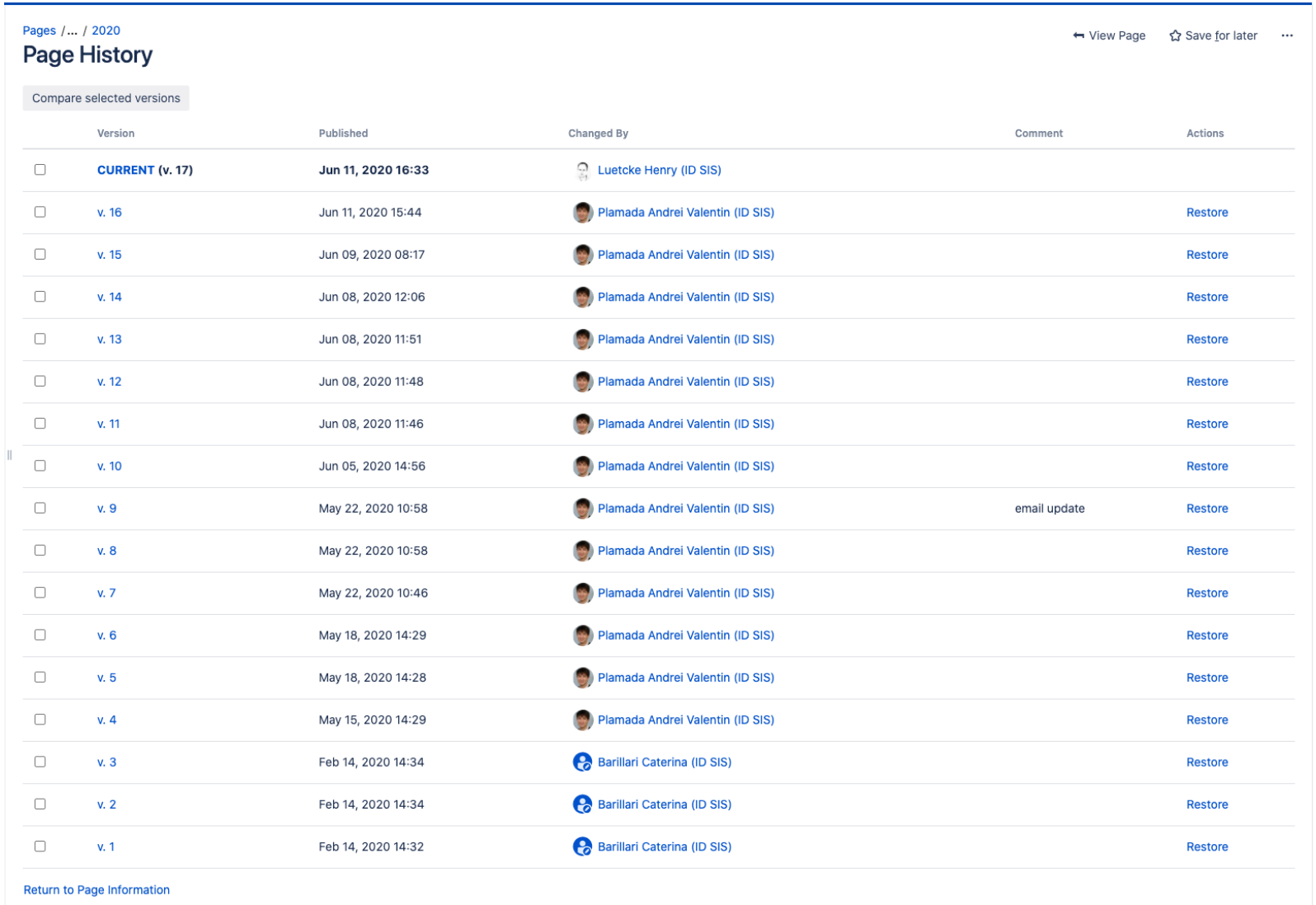
Size of Confluence Space/ git repository	Yearly price
<2 GB	free
<10 GB	300 CHF
<50 GB	1000 CHF

- Some departments and/or institutes host their own wiki. Always contact your ISG for info.

Example of confluence wiki use in SIS



Example of versioning in wiki



Pages /... / 2020

View Page Save for later ...

Page History

Compare selected versions

	Version	Published	Changed By	Comment	Actions
☐	**CURRENT (v. 17)**	**Jun 11, 2020 16:33**	Luetcke Henry (ID SIS)		
☐	v. 16	Jun 11, 2020 15:44	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 15	Jun 09, 2020 08:17	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 14	Jun 08, 2020 12:06	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 13	Jun 08, 2020 11:51	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 12	Jun 08, 2020 11:48	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 11	Jun 08, 2020 11:46	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 10	Jun 05, 2020 14:56	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 9	May 22, 2020 10:58	Plamada Andrei Valentin (ID SIS)	email update	Restore
☐	v. 8	May 22, 2020 10:58	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 7	May 22, 2020 10:46	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 6	May 18, 2020 14:29	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 5	May 18, 2020 14:28	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 4	May 15, 2020 14:29	Plamada Andrei Valentin (ID SIS)		Restore
☐	v. 3	Feb 14, 2020 14:34	Barillari Caterina (ID SIS)		Restore
☐	v. 2	Feb 14, 2020 14:34	Barillari Caterina (ID SIS)		Restore
☐	v. 1	Feb 14, 2020 14:32	Barillari Caterina (ID SIS)		Restore

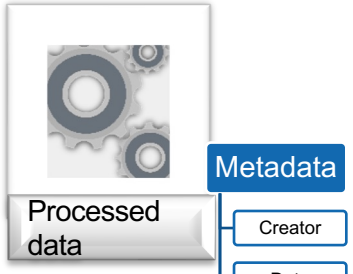
Return to Page Information

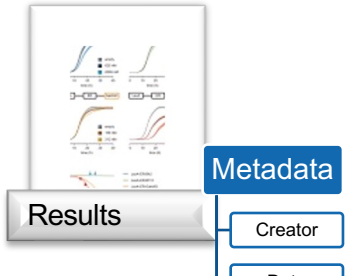
Questions on Management of Samples and Protocols?

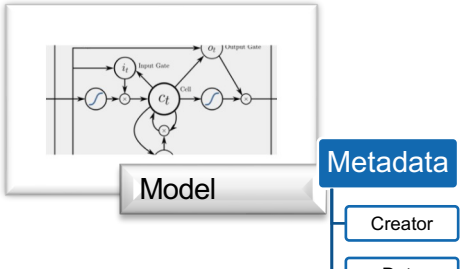
Management of Data & Metadata

Management of research data files

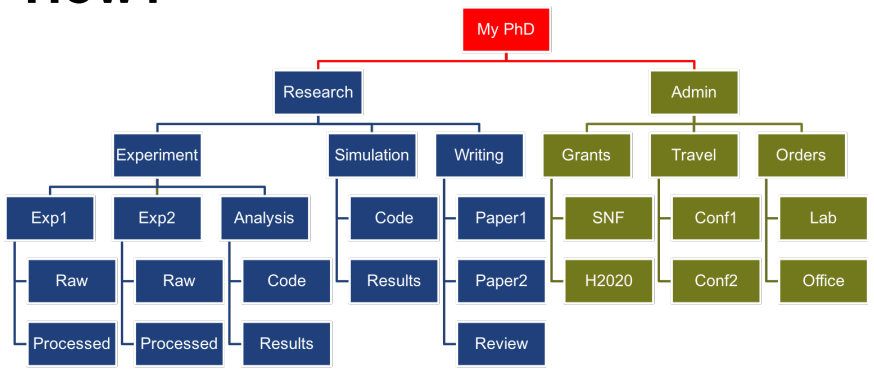
What?



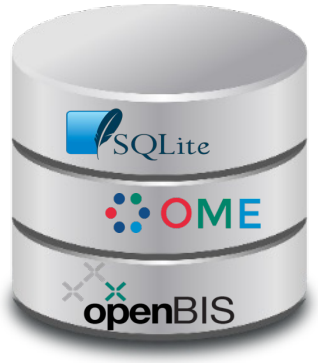




How?

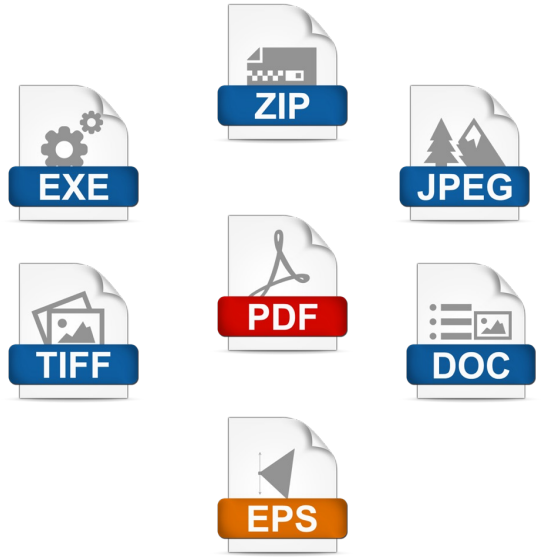


Files / folders hierarchy



Data management platform

File types and formats



- File format: a **convention for encoding information** in a computer file
- Extensions typically *indicate* a specific file format
- Some file formats are preferable to others
- Stick to **non-proprietary** and **widely used** formats!
- Several general-purpose **scientific data formats** exist (e.g. HDF5, netCDF, FITS)

-	+
Binary	Text-based
Proprietary	Open
New kid on the block	Old as the hills
Compressed/encrypted	Uncompressed/unencrypted
Platform dependent	Interoperable
Complex	Simple

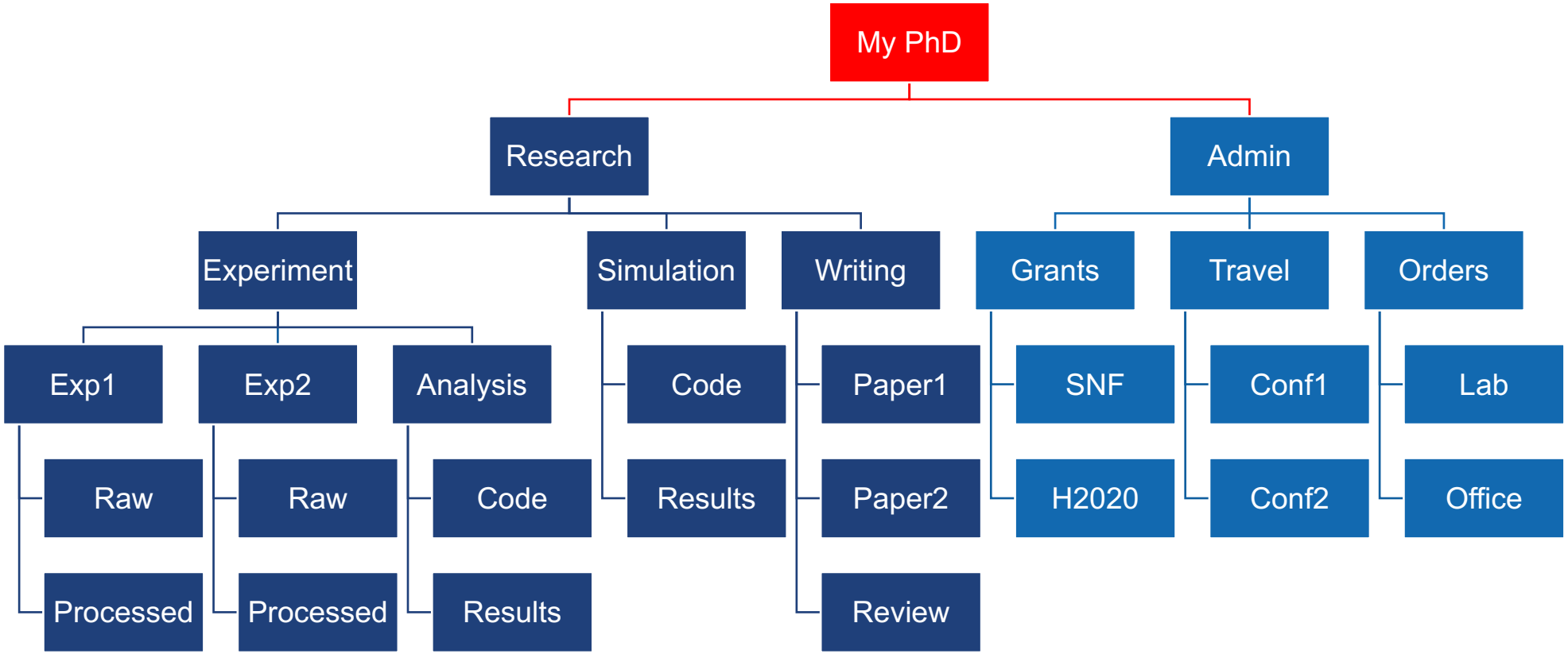
File & folder organisation

- **Goals** of efficient file / folder organization:
  - Easy to **find** something in the future (you, others)
  - Easy to **file** something
  - Save **disk space** (avoid multiple copies of files)
  - **Reusable** components
  - Avoid problems on different **operating systems**

- Planning a good **folder hierarchy**
  - **Who?** Individual, research group
  - **Where?** Local disk, shared network drive
  - Which operating system?
  - What information are you going to **search** for?
  - Avoid non-descriptive file and folder names (*figure_02_summary_stats.png* and not *stats.png*)
  - Add **descriptive text files** to folders (→ Metadata)
  - **Document** your hierarchy and file naming convention

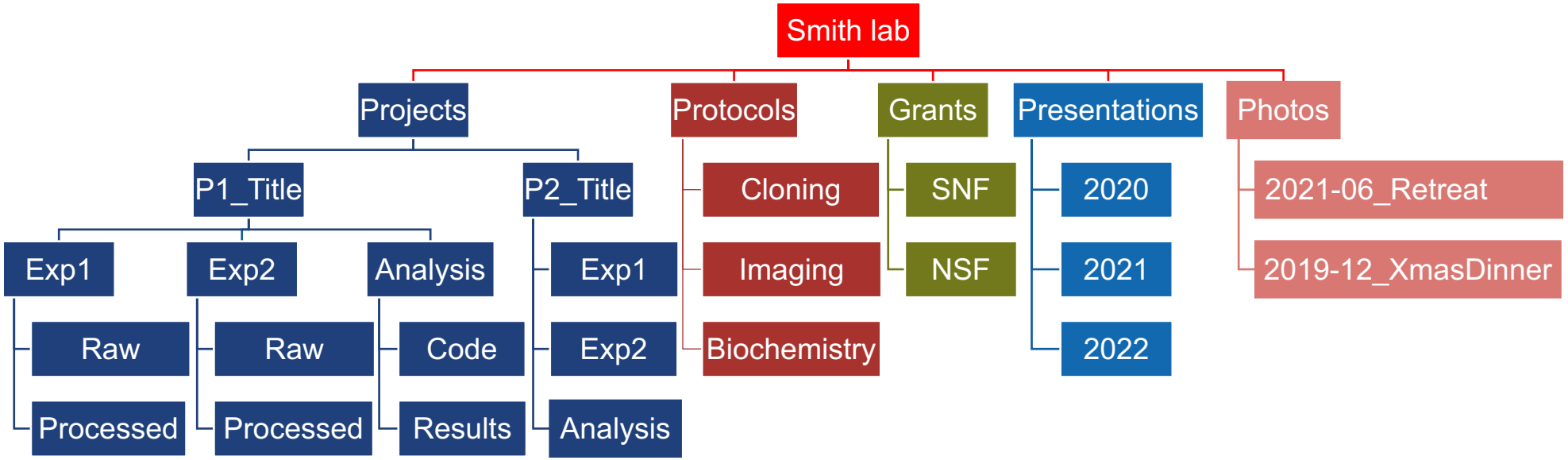
File & folder organisation

Example hierarchy for a PhD project



File & folder organisation

Example hierarchy for a research group



File & folder organisation

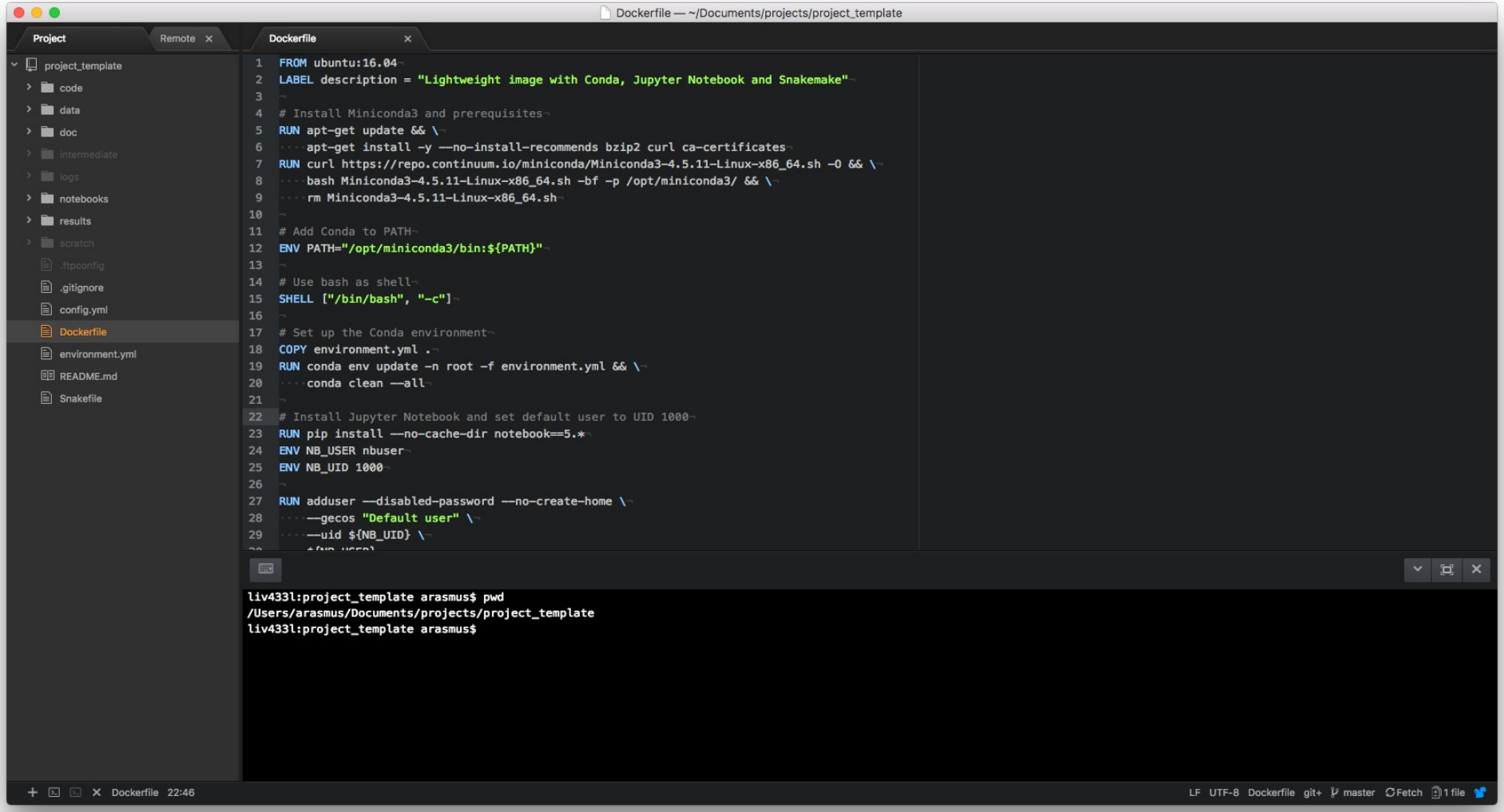
The project directory (for a computational project)

```
project
|- doc/                 documentation for the study
|
|- data/                raw and primary data, essentially all input files, never edit!
|   |- raw_external/
|   |- raw_internal/
|   |- meta/
|
|- code/                all code needed to go from input files to final results
|- notebooks/           notebooks that document your day-to-day work
|
|- intermediate/        output files from different analysis steps, can be deleted
|- scratch/             temporary files that can be safely deleted or lost
|- logs/                logs from the different analysis steps
|
|- results/             output from workflows and analyses
|   |- figures/
|   |- tables/
|   |- reports/
|
|- Snakefile            project workflow, carries out analysis contained in code/
|- config.yml           configuration of the project workflow
|- environment.yml      software dependencies list, used to create a project environment
|- Dockerfile           recipe to create a project container
```

Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424

File & folder organisation

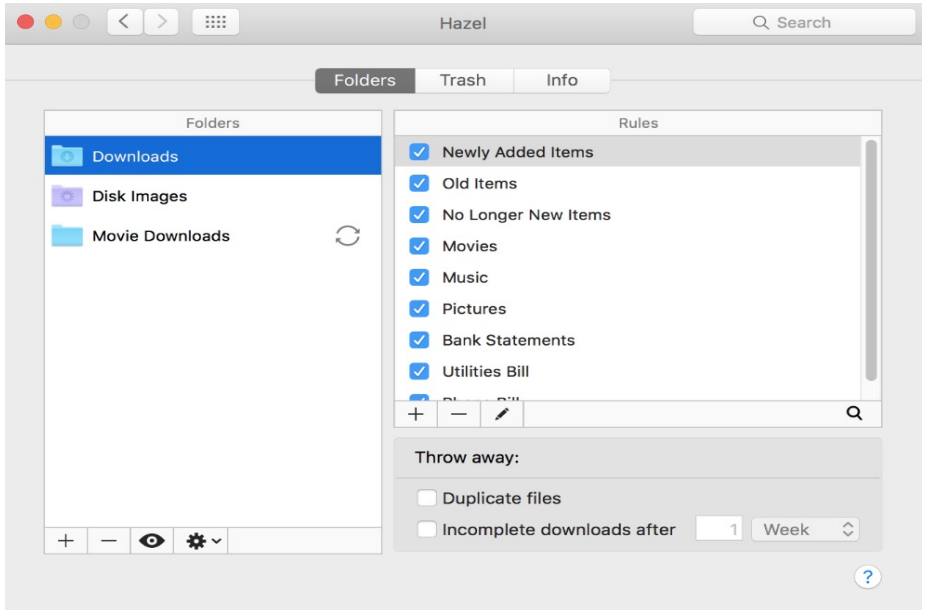
A project in ATOM



File & folder organisation

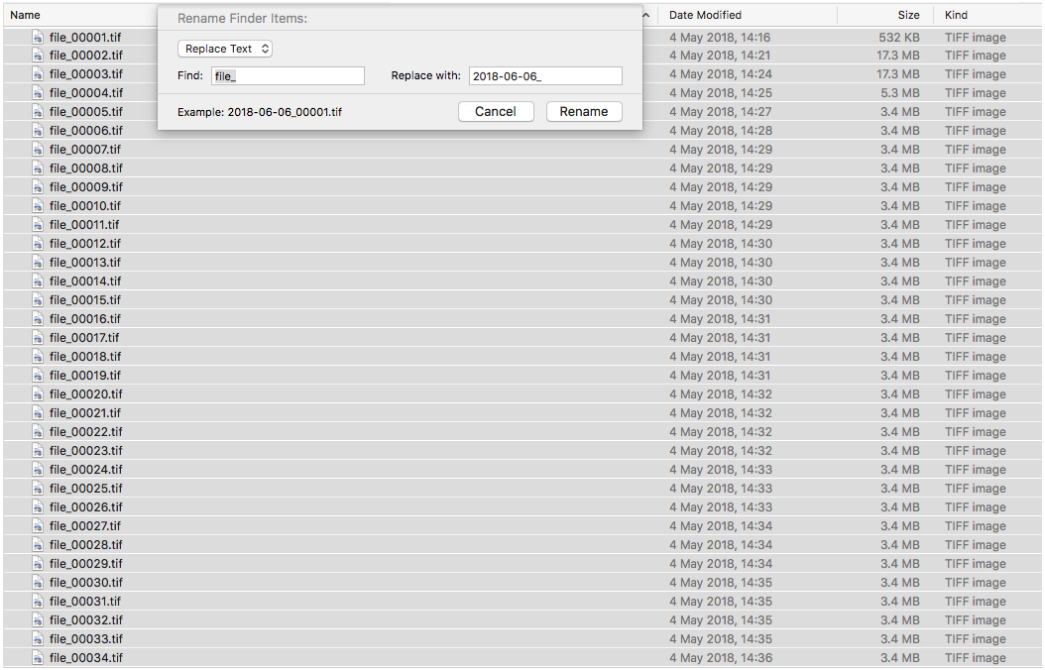
- ❑ Keep path names short (< 256 characters)
- ❑ Recommendation for file names:
  - Unique, reflect content (if possible)
  - Use only ASCII characters
  - Not include spaces
  - Be aware of case sensitivity
- ❑ Bad examples:
  - data%20management%20plan.docx
  - sup figure 2.png
  - lab meeting 19.10.2021.pptx
- ❑ Good examples:
  - Data_management_plan_SNF.docx
  - sup_figure_02_summary_stats.png
  - lab_meeting_2021-10-19.pptx

- ❑ Use links / shortcuts to avoid duplications
- ❑ Use tags for orthogonal classifications
- ❑ Create template folders
- ❑ Some tools for automated file organization:
  - Mac: Hazel (https://www.noodlesoft.com/)
  - PC: DropIt (http://www.dropitproject.com/)



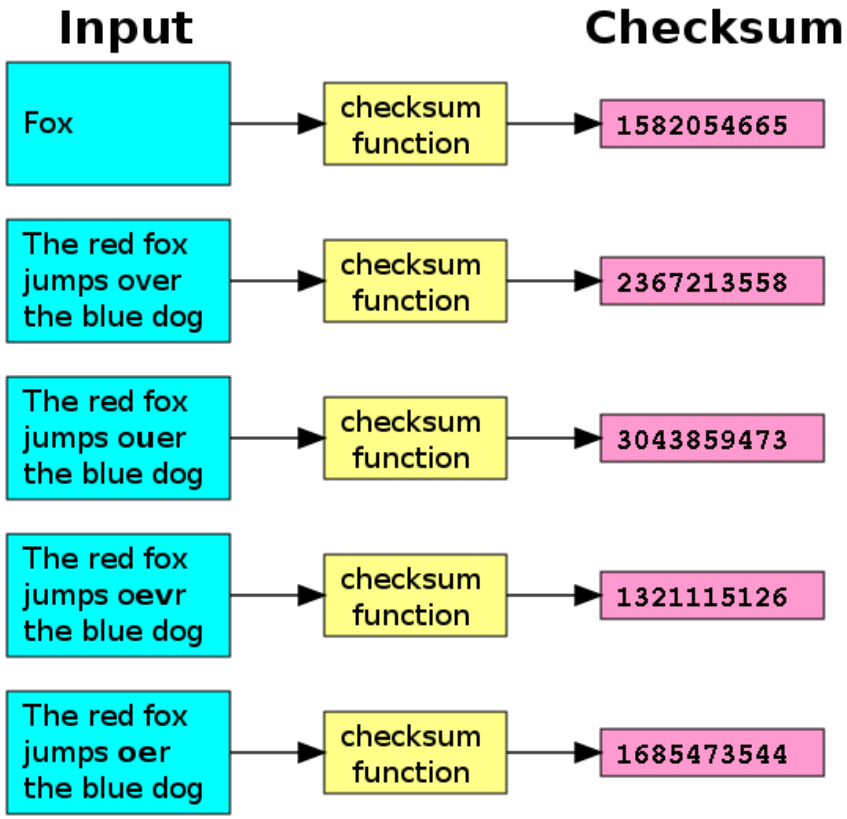
Batch renaming of files

- **Windows**
  - Bulk Rename Utility (www.bulkrenameutility.co.uk)
  - Advanced Renamer (www.advancedrenamer.com)
  - Command prompt / PowerShell scripts
- **macOS**
  - Finder rename functionality
  - Automator
  - Command line / scripts
- **Linux**
  - Command line utility `rename`
  - Métamorphose (http://file-folder-ren.sourceforge.net/)



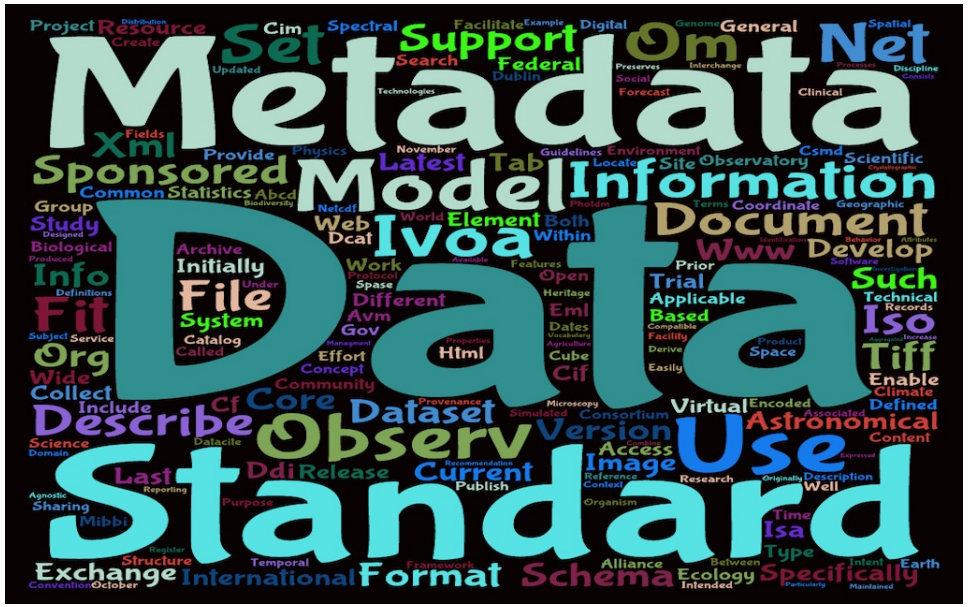
File validation and compression

- **Checksum** algorithms are useful to **verify data integrity**
  - For datasets stored over long time periods
  - When transferring from one storage to another
  - Examples: *cksum*, *md5sum*, *sha1sum*
- **Compression** algorithms encode information in fewer bits than the original (bit-rate reduction)
  - Lossless vs. lossy compression
  - Lossy compression may lead to distortion / loss in quality but compression rates are typically much higher
  - Trade-off between processing time and disk space
  - Examples: zip, gzip, specific formats (TIFF, HDF5)
- **Archiving** combines many (small) files into a single (large) file
  - Nowadays mainly used together with compression
  - Required for long-term storage (LTS) and efficient data transfer
  - Examples: Tape Archive format (tar)



Metadata

- **Metadata** is the *data about your data* (a.k.a. **data model**)
- Use of structured metadata **facilitates data organization** and searches
- (Machine-readable) **Metadata** is a key element of the **FAIR data** principles
- Existing **metadata schemas** are preferred (can be extended, if necessary)



https://rdamsc.bath.ac.uk/

Metadata

Example for general research data: **DataCite Metadata schema**

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

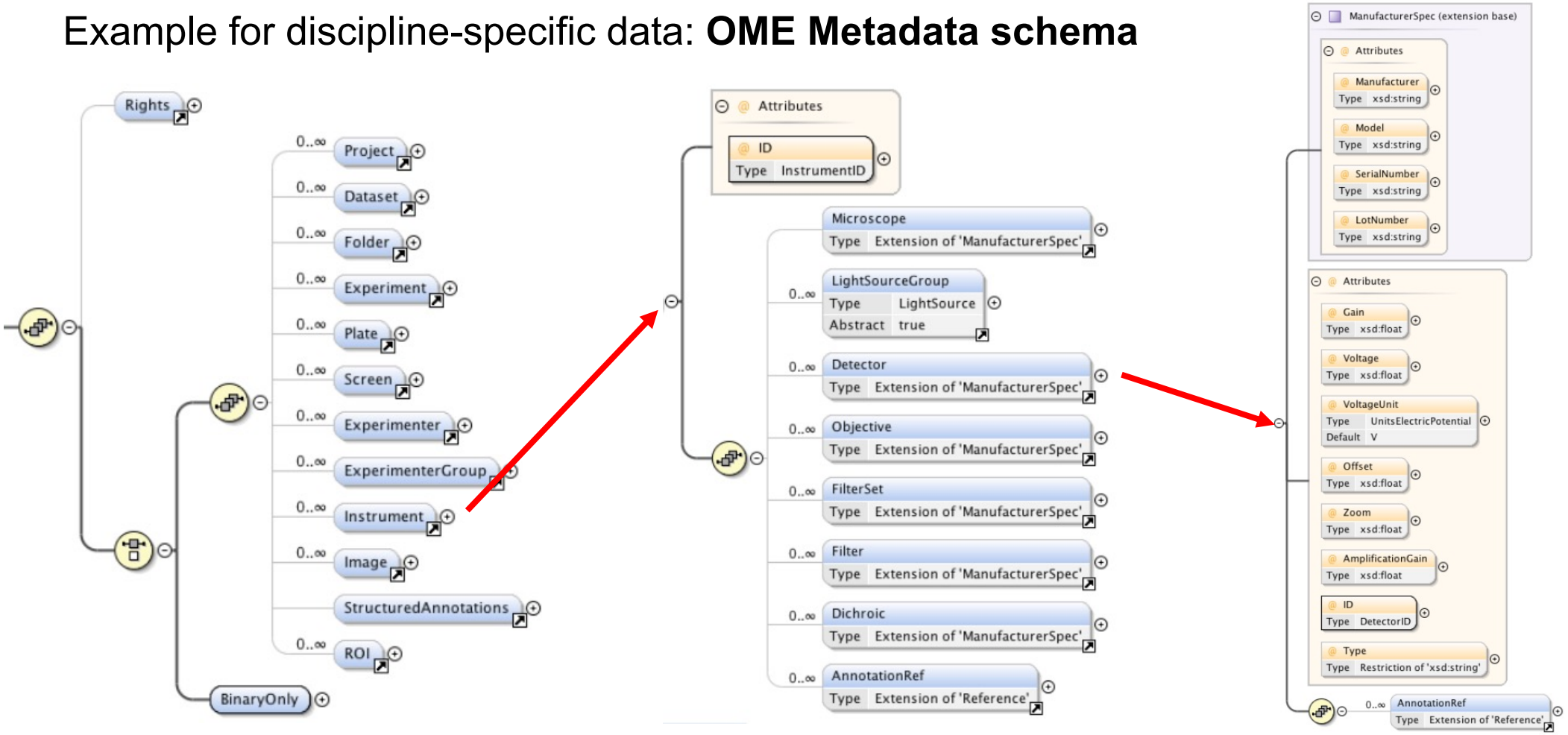
https://schema.datacite.org/

Table 2: DataCite Recommended and Optional Properties

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O

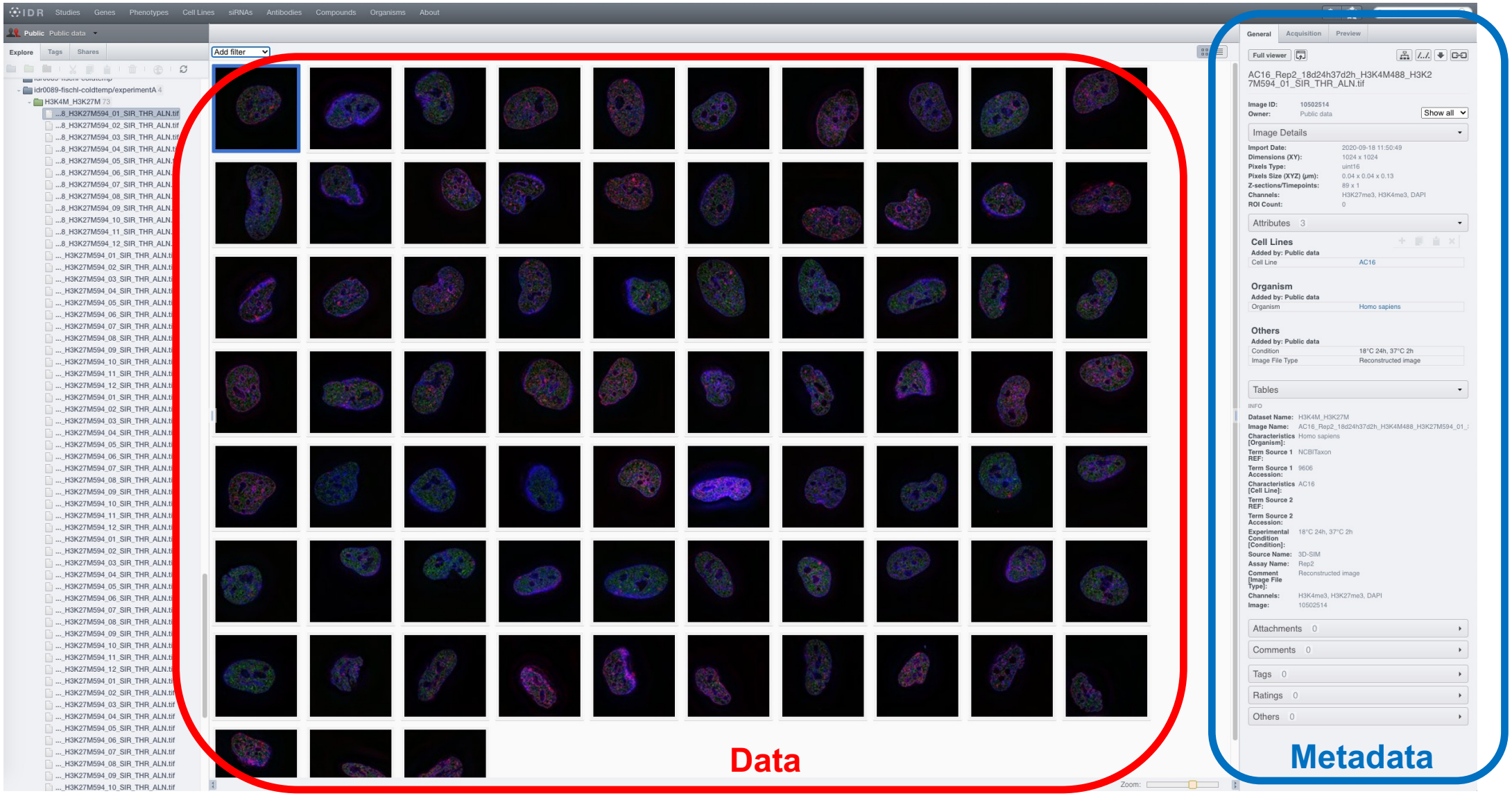
Metadata

Example for discipline-specific data: **OME Metadata schema**



https://www.openmicroscopy.org/Schemas/

Metadata



https://idr.openmicroscopy.org/webclient/?show=project-1304

Metadata Exercise – Discipline-specific Standards

The Research Data Alliance (RDA) has created a community-maintained catalog of metadata standards: https://rdamsc.bath.ac.uk/

The catalog contains metadata standards that have been established in different research fields as well as relevant tools and use cases. Consider the following short descriptions of hypothetical research projects. Based on the description, try to identify one or several suitable standards from the metadata catalog.

Case 1: An online survey / questionnaire about attitudes towards climate change in different age groups. Documentation of the survey responses.

Metadata Exercise – Discipline-specific Standards

The Research Data Alliance (RDA) has created a community-maintained catalog of metadata standards: https://rdamsc.bath.ac.uk/

The catalog contains metadata standards that have been established in different research fields as well as relevant tools and use cases. Consider the following short descriptions of hypothetical research projects. Based on the description, try to identify one or several suitable standards from the metadata catalog.

Case 1: An online survey / questionnaire about attitudes towards climate change in different age groups. Documentation of the survey responses.

- ❑ Data Documentation Initiative (DDI), version 2 or 3
- ❑ An international standard for describing surveys, questionnaires, statistical data files, and social sciences study-level information

Metadata Exercise – Discipline-specific Standards

Case 2: A study on the diversity and geographic distribution of butterflies in the Swiss Alps. Documentation of the recorded occurrences.

Metadata Exercise – Discipline-specific Standards

Case 2: A study on the diversity and geographic distribution of butterflies in the Swiss Alps. Documentation of the recorded occurrences.

- Darwin Core (DwC)
- Ecological Metadata Language (EML)
- **Darwin Core** is a standard to facilitate the sharing of information about biological diversity. DwC includes a glossary of terms, and documentation providing reference definitions, examples, and commentary. It is an extension of Dublin Core.
- **EML** defines a vocabulary and a syntax for documenting research data. It is in widespread use in the earth and environmental sciences. EML includes modules for describing the spatial, temporal, taxonomic, and thematic extent of data, for describing research methods and protocols, for describing the structure and content of data, and for precisely annotating data with semantic vocabularies.

Metadata Exercise – Discipline-specific Standards

Case 3: The crystal structure of a novel viral protein determined by X-Ray crystallography. Documentation of the protein structure.

Metadata Exercise – Discipline-specific Standards

Case 3: The crystal structure of a novel viral protein determined by X-Ray crystallography. Documentation of the protein structure.

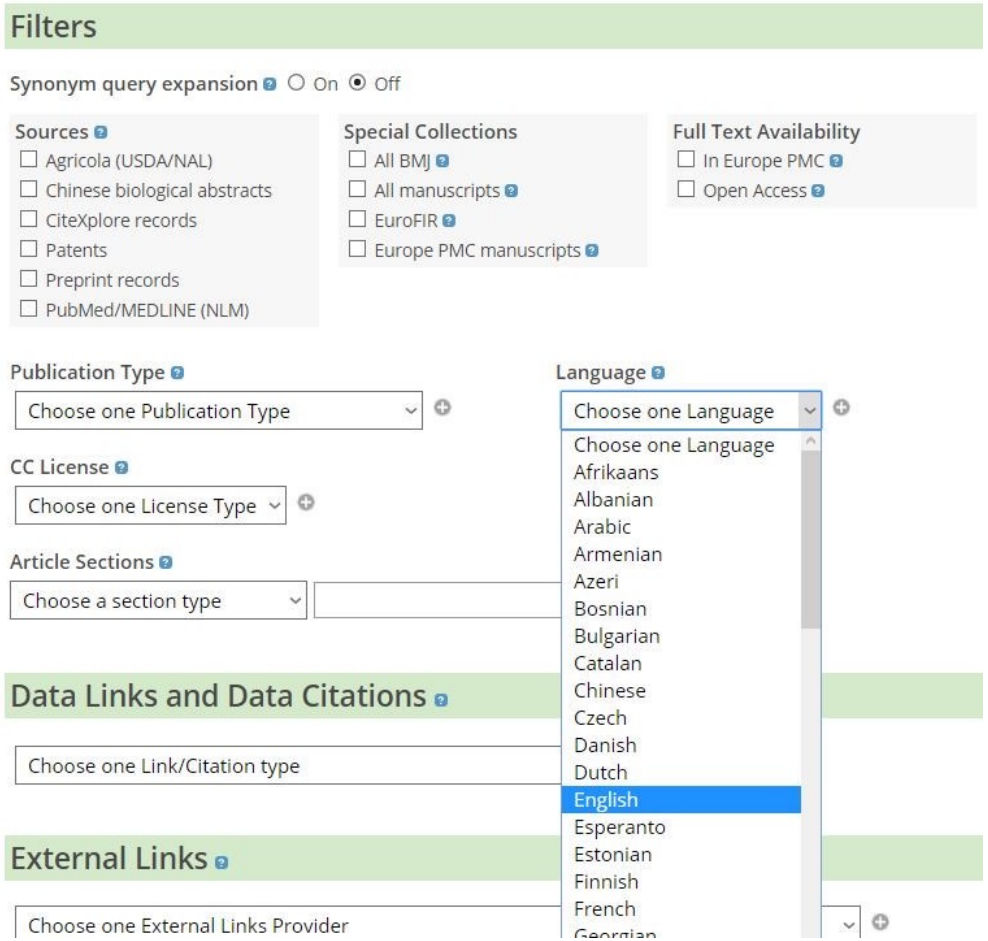
- Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework (PDBx/mmCIF)

Metadata Exercise – Discipline-specific Standards

FAIRsharing.org provides an even more extensive resource of metadata standards, vocabularies, taxonomies etc.

Metadata

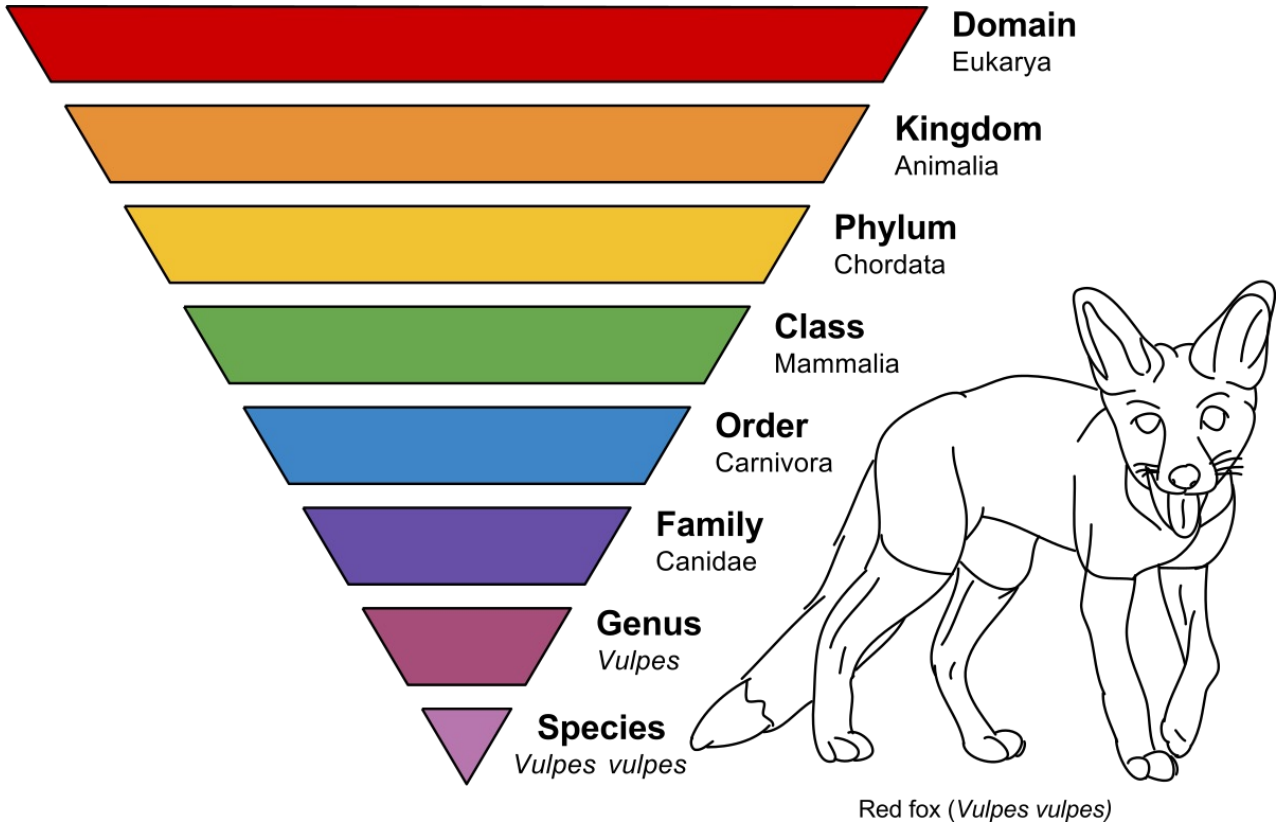
- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies



https://europepmc.org/advancesearch

Metadata

- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies



Red fox (*Vulpes vulpes*)

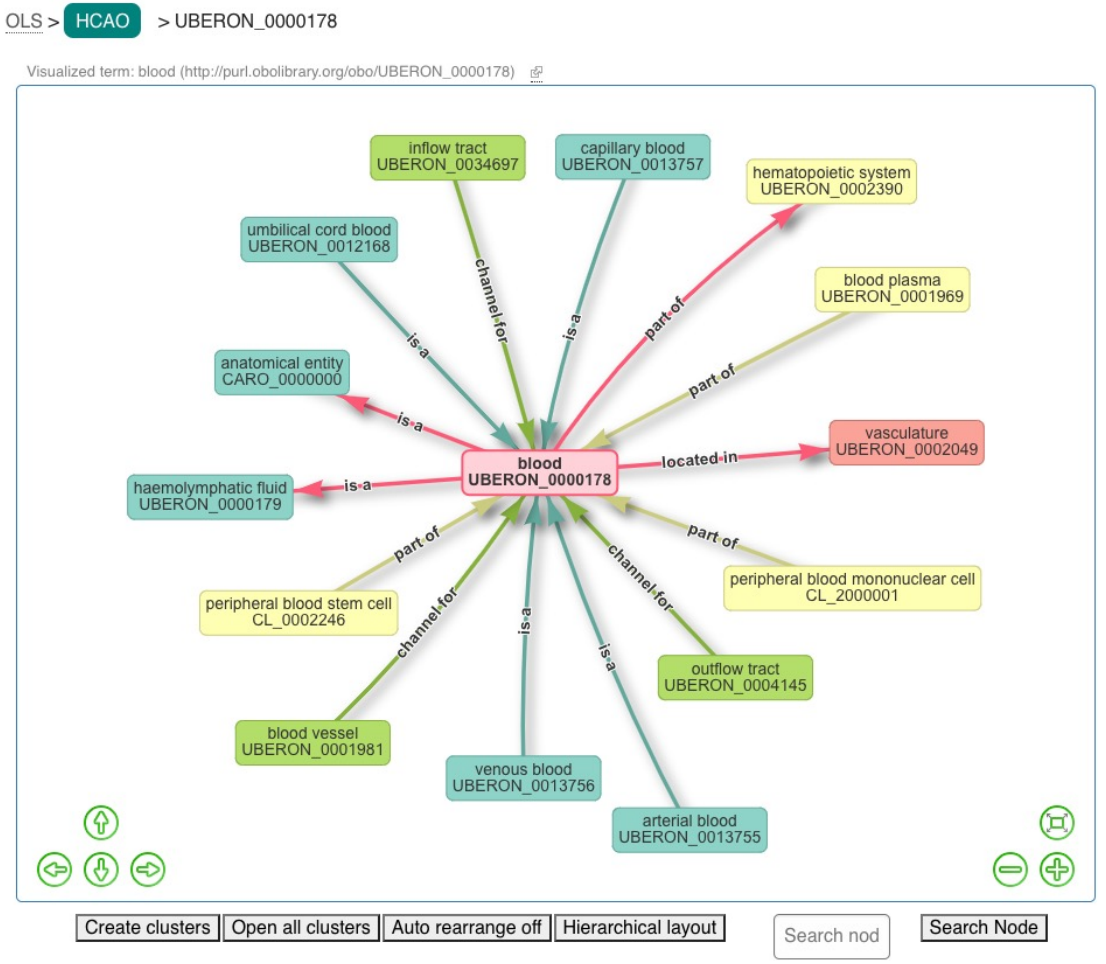
https://en.wikipedia.org/wiki/Domain_%28biology%29

Metadata

- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies
  - Standardized metadata are the basis for ontologies (knowledge representations)

[Basic Register of Thesauri, Ontologies & Classifications]

[EMBL-EBI Ontology Lookup Service]



Knowledge graph for *'blood'* from the [Human Cell Atlas Ontology]

Metadata

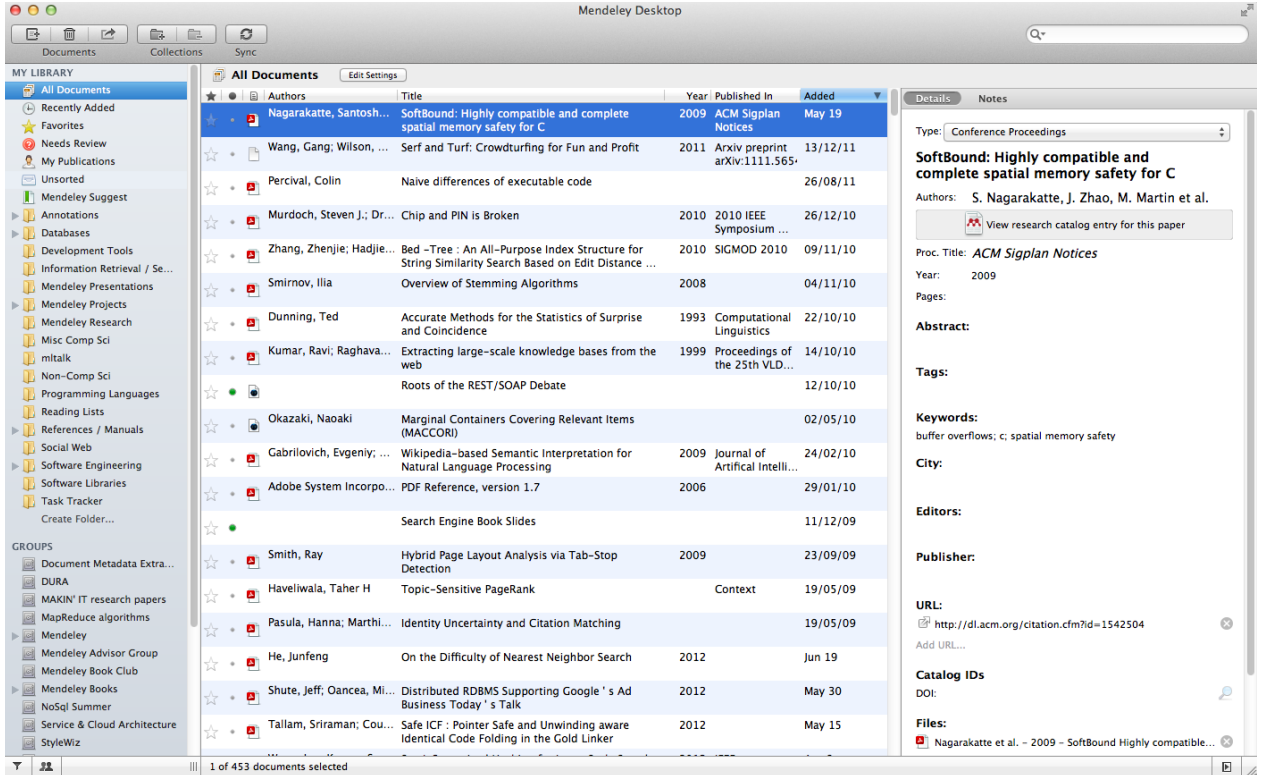
- **Types of metadata**
  - Descriptive (Title, author, identifier)
  - Administrative (License)
  - Technical (File size, checksums)
  - Structural (Relation to other data)
- **Machine-readable metadata**
  - Annotation based on common standards
  - Controlled vocabularies, taxonomies
  - Standardized metadata are the basis for ontologies (knowledge representations)

  Basic Register of Thesauri, Ontologies & Classifications

  EMBL-EBI Ontology Lookup Service
- **How to apply metadata?**
  - File header (e.g. TIFF, netCDF)
  - Sidecar files (e.g. XML, JSON)
  - As columns in a database
  - README text files
- **Apply metadata early** in your workflow (ideally during acquisition)

Data management software

- System that allows **structured organization** of data
- Data is described by **metadata**
- Usually more FAIR-compliant than Files / Folders
- Searchable, scalable, flexible
- Allows user rights management
- Back up procedures are easily implemented



Data management software

- System that allows **structured organization** of data
- Data is described by **metadata**
- Usually more FAIR-compliant than Files / Folders
- Searchable, scalable, flexible
- Allows user rights management
- Back up procedures are easily implemented
- Examples

Generic

PostgreSQL SQLite FileMaker An Apple Subsidiary Microsoft Access

Scientific

openBIS LabKey PostGIS Spatial PostgreSQL

OME rasdaman raster data manager

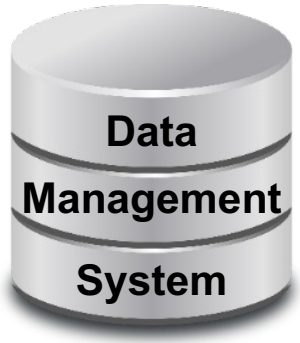
Data management software

Generic database workflow

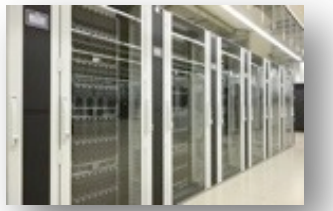
Measure / Simulate

Automated or manual

Import & Annotation



- **Search**
- **List**
- **Curate**
- **Share**



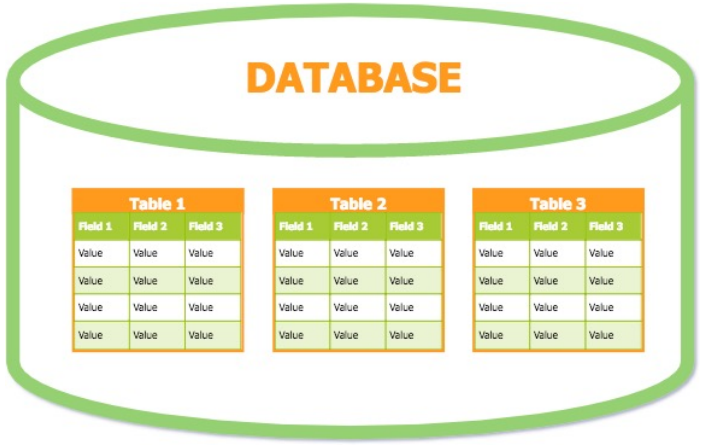
- **Process**
- **Analyze**
- **Visualize**

- **Backup**
- **Archive**

Data management software

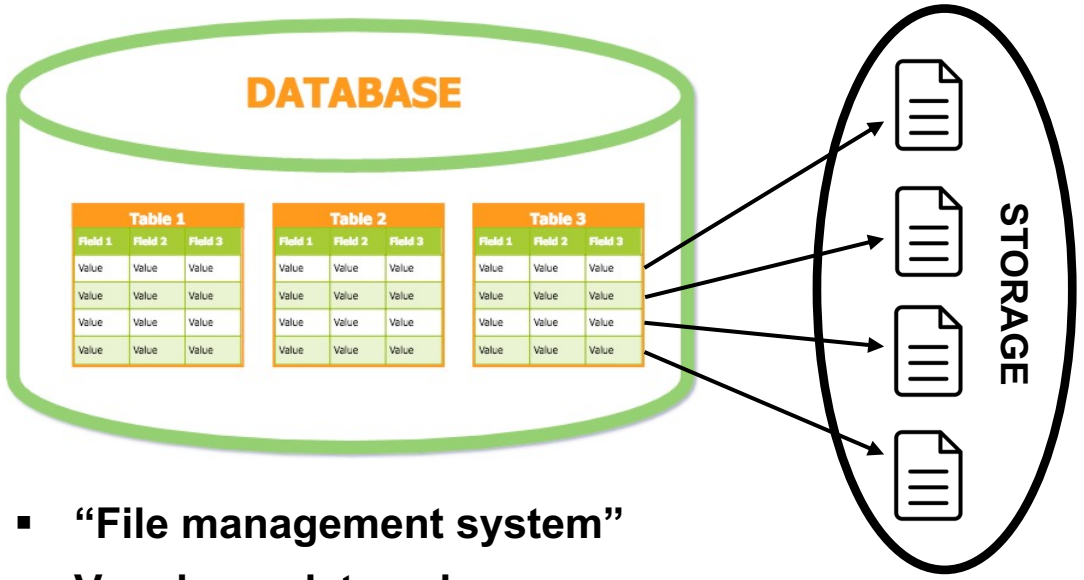
Where is my data???

Metadata + Data



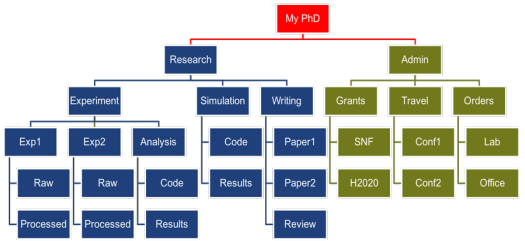
- Structured data required
- DB-specific data (array DBs)
- No conventional file access

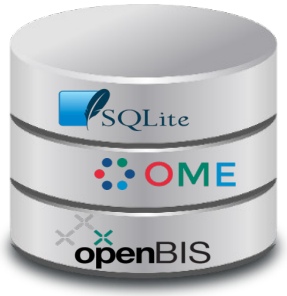
Metadata + Link to Data Files



- "File management system"
- Very large data volumes
- Conventional file access

Comparison: folder hierarchy vs. databases





	File / folder Hierarchy	Database System
Easy to use	Yes	Initial learning curve
Flexibility	High	Limited
Scalability (data volume)	Limited	Yes
Scalability (users)	Limited	Yes
Versioning included	No	Yes
Backup included	No	Yes

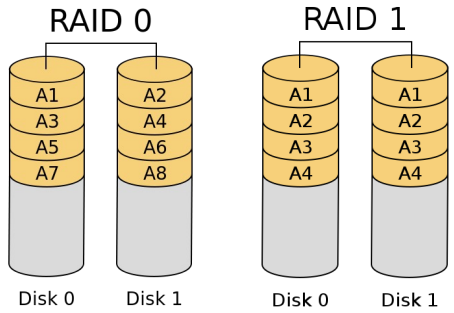
Data storage – Your options …

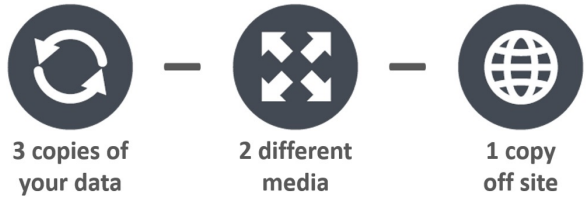
- External hard disks are cheap but unreliable and don't scale!

Data storage – Your options …

- ❑ External hard disks are cheap but unreliable and don't scale!
- ❑ Polybox / SwitchDrive
- ❑ Network Attached Storage (NAS)
  - NAS offering from IT Services – always check with your ISG first
- ❑ Cost Defined Storage (CDS)
  - For very large storage requirements (>100 TB)
- ❑ Long-term Storage (LTS)
  - Tape storage for data worthy of preservation ("Datenendlager")
  - Data should be compressed and sized between 10 - 200 GB
- ❑ Also consider backups (ideally follow 3-2-1 rule)
  - RAID is not a backup!
- ❑ Details: https://ethz.ch/services/en/it-services/catalogue/storage.html
- ❑ ETH news on cloud storage regulations: https://bit.ly/3BoQQD1

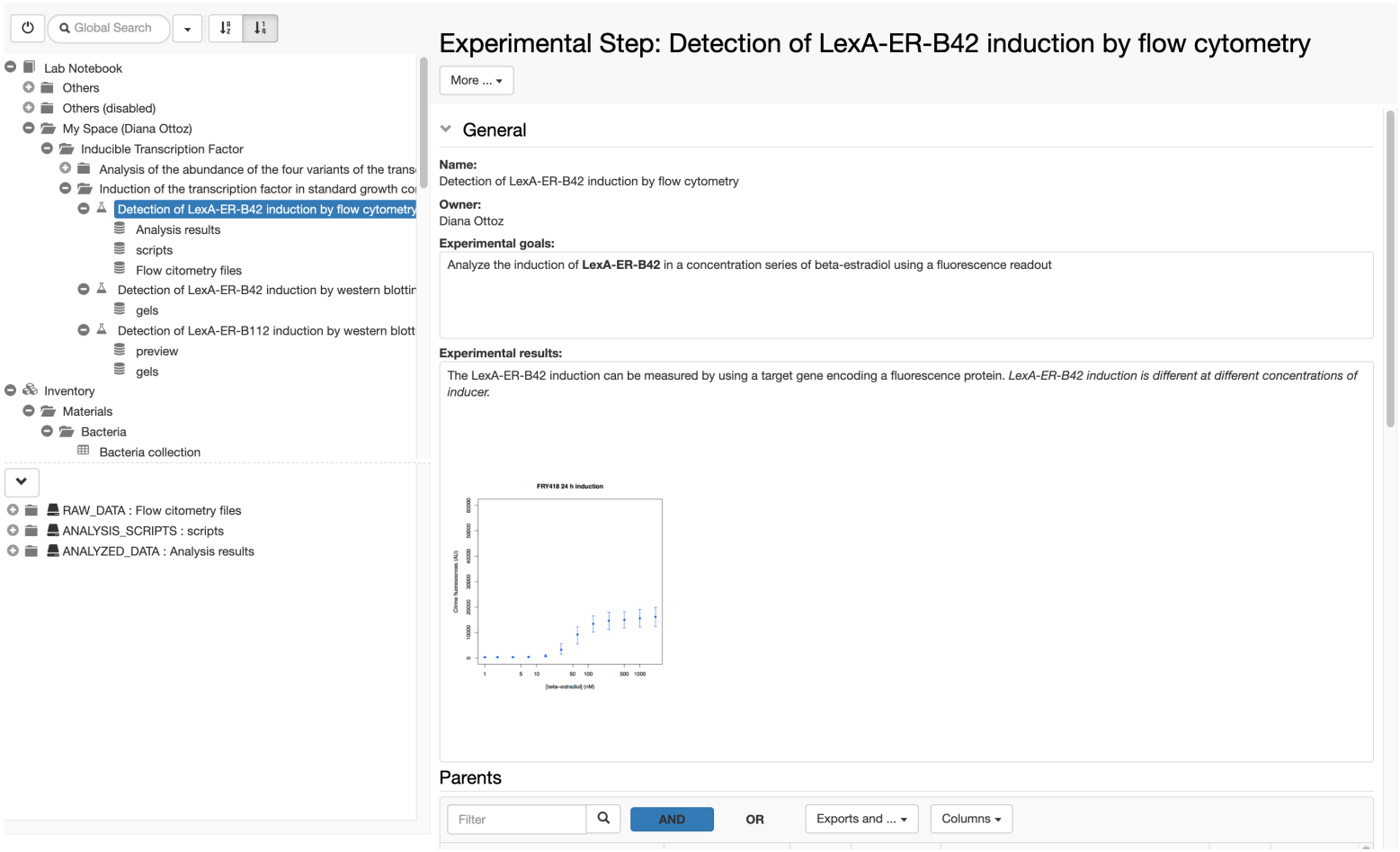
Redundant
Array of
Independent
Disks





Questions on Management of Data Files?

Time for a break



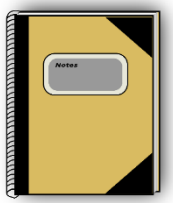
Electronic Laboratory Notebooks

Experimental description / notes

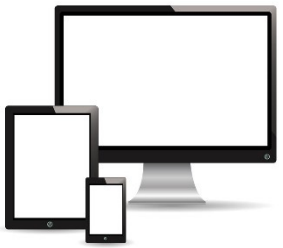
What to document?

- ❑ Goals
- ❑ Materials
- ❑ Methods
  - Experimental/computational procedure
  - Analysis procedures
- ❑ Results
- ❑ Links to data

How?



Paper laboratory notebook



Electronic laboratory notebook (ELN)

Definition of ELN & requirements

*An **electronic laboratory notebook** (also known as **electronic lab notebook** or **ELN**) is a software program or package designed to replace more traditional paper laboratory notebook . Laboratory notebooks in general are used by scientists and technicians to document, store, retrieve, and share fully electronic laboratory records in ways that meet all legal, regulatory, technical and scientific requirements.*

Legal requirements

Electronic lab notebooks used for development or research in regulated industries, such as medical devices or pharmaceuticals, are expected to comply with the **21 CRF Part 11 FDA** regulations: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?cfrpart=11

- To our knowledge, in academia, there is no such requirement, but this can vary from one Institution to another.

Source: https://www.limswiki.org/index.php/Electronic_laboratory_notebook

ETHZ Regulations

- **ETHZ RDM guidelines** came into action on **01.07.2022:**
  https://rechtssammlung.sp.ethz.ch/Dokumente/414.2en.pdf

Art. 5 Data collection and processing

[1] Digital data that are in the process of being collected or processed must be stored such that data can be restored in case of hardware problems or human error. Using storage systems centrally supplied by ETH fulfils this criterion. If such data are too big for being stored redundantly in the available infrastructure, researchers must be able to recreate them using existing code and documentation when required.

[2] Access rights to the storage system must be specified and regularly updated, e.g. upon the start of a new project. The scientific needs of all researchers working with the data must be considered by the access right system. Unauthorised access must be prevented by suitable technical and organisational measures, such as ETH Zurich's identity and access management system.

[3] In order to ensure long-term accessibility and reusability, the use of well-documented, non-proprietary file formats is recommended.

[4] The structure and the processing steps of all *Research Data* must be digitally documented in order to ensure adherence to the *FAIR principles*. Where documentation includes a lab journal, Electronic Laboratory Notebooks (ELN) are recommended.

ELNs vs. paper notebook

❑ **Advantages of ELNs over paper notebooks**:

1. Sharing
2. Most ELNs have rights management
3. Most ELNs keep track of changes
4. Search functionality
5. Easier to link digital data
6. No issues with handwriting
7. Can be backed up

❑ **Disadvantages of ELNs over paper notebooks:**

1. Require change in working mode
2. Some ELNs have a learning curve

Note-keeping applications

A few examples

cloud-based

self-hosted

- Most solutions are moving towards cloud-based services
- Straight replacement of paper notebooks with some added values (e.g. sharing, searching)
- Popular in academia due to ease of use
- Do not provide a solution for data management
- Do not comply with **21 CRF Part 11 FDA** regulations

ELNs with database back end

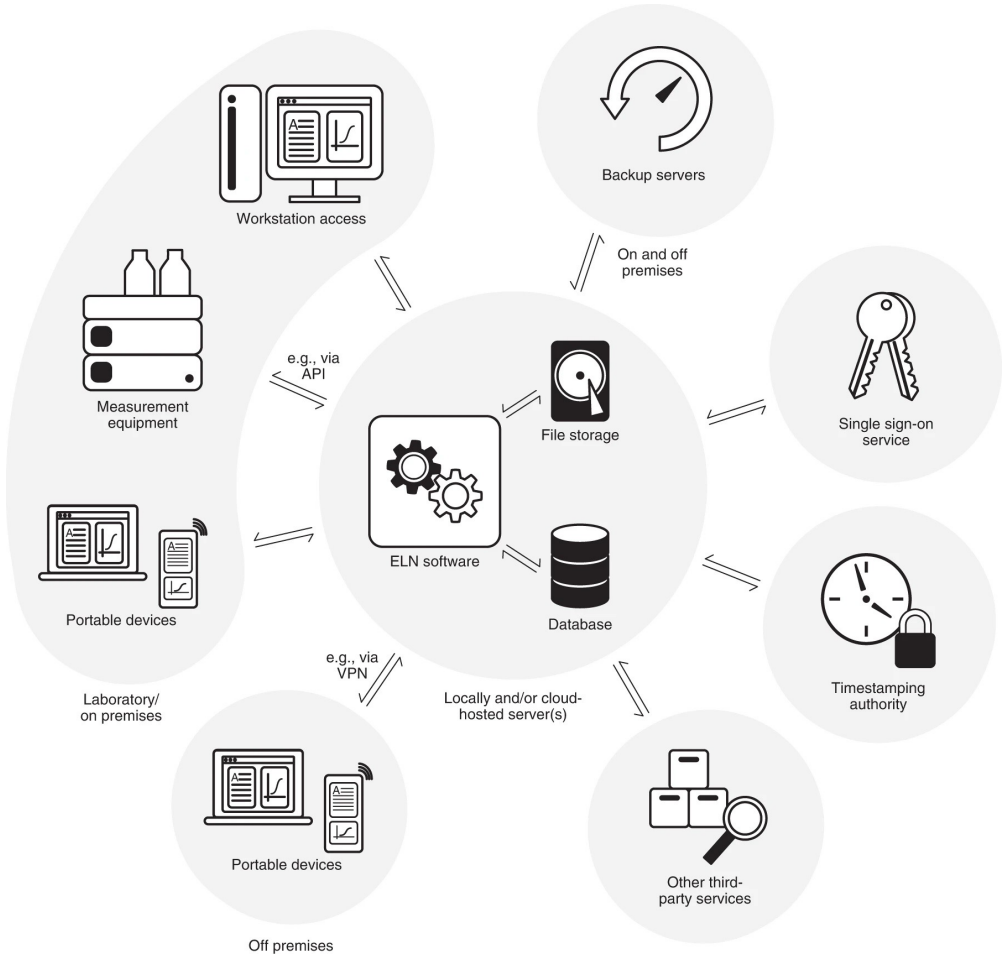
A few examples

cloud-based

self-hosted

- Structured ELNs compared to note-keeping applications
- Additional functionalities compared to note-keeping applications (e.g. workflow management, chemical structures drawing, etc)
- Can be discipline-specific or cross-disciplines
- Some systems offer an all-in-one solution for RDM
- Many systems comply with **21 CRF Part 11 FDA** regulations

Simplified illustration of a commonly used ELN architecture



Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nat Protoc* **17**, 179–189 (2022). https://doi.org/10.1038/s41596-021-00645-8

Which ELN to choose?

- The ELNs offer on the market is very large. Answering some of these questions might help you restricting the choice:

1. Is it for personal use or group use?
2. Can I/we use a cloud-based solution?
3. Do I/we need specific features?
4. What do I/we want to do with the ELN? (e.g. only write experimental descriptions, manage samples, manage data – how big?, etc.)
5. Commercial v. open-source
6. Budget?
7. Can I export my data?

- *How to pick an Electronic Laboratory Notebook: https://www.nature.com/articles/d41586-018-05895-3*
- *Harvard University Comparison Grid:* https://datamanagement.hms.harvard.edu/electronic-lab-notebooks

ELN guidelines for academic research labs

- As part of the national DLCM project, ETH Zurich and EPFL prepared some ELN guidelines, available from: https://www.dlcm.ch/services/dlcm-eln

Stages of introduction of an ELN in an academic lab:

1. Initialization
2. Analysis of laboratory needs
3. Pilot phase
4. Production phase (*migration of existing databases or information; customization, integration with measuring instruments, integration with analysis pipelines and workflows*).

What is needed for a successful introduction of an ELN in an academic lab?

- Based on our experience at ETHZ:

  1. Commitment from the PI
  2. Lead person in the lab who drives the process
  3. RDM expert support

- Examples of ELN implementation in academic institutions:

  - Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nat Protoc* **17**, 179–189 (2022). https://doi.org/10.1038/s41596-021-00645-8
  - Institutional ELN/LIMS deployment. EMBO Rep. 2020 Mar 4; 21(3): e49862. doi: 10.15252/embr.201949862

Slow adoption of ELNs/LIMS in academia

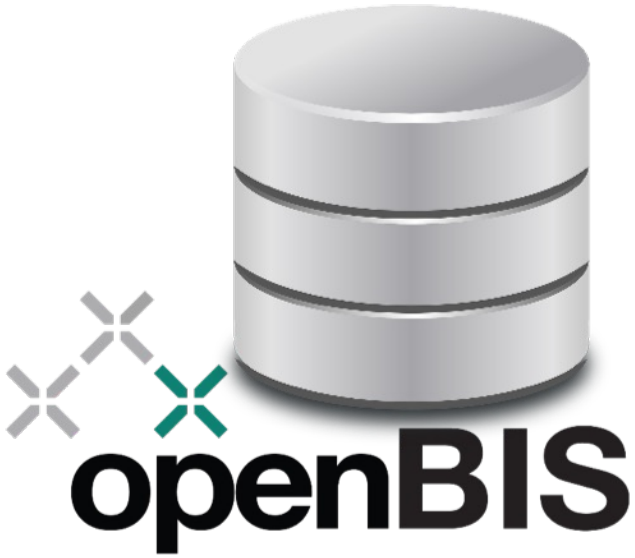
The Barriers	Research lab	Diagnostic lab
Limited budget	352	23
Time needed for implementation	235	21
Changes to existing working habits	266	25
Concerns about ELN being a system in the cloud	215	22
Contains features we do not need	130	7
Software in English only	76	10
No need for improvement	112	7

Source: Potential uses of ELNs in Academia survey (University of Southampton). J. Cheminform (2017) 9:31.

In our experience at ETHZ, the main reasons are:

1. Change in working habits needed.
2. Time needed for introduction in a lab.
3. Concerns about data retrieval when leaving the lab.

Questions on Electronic Lab Notebooks?

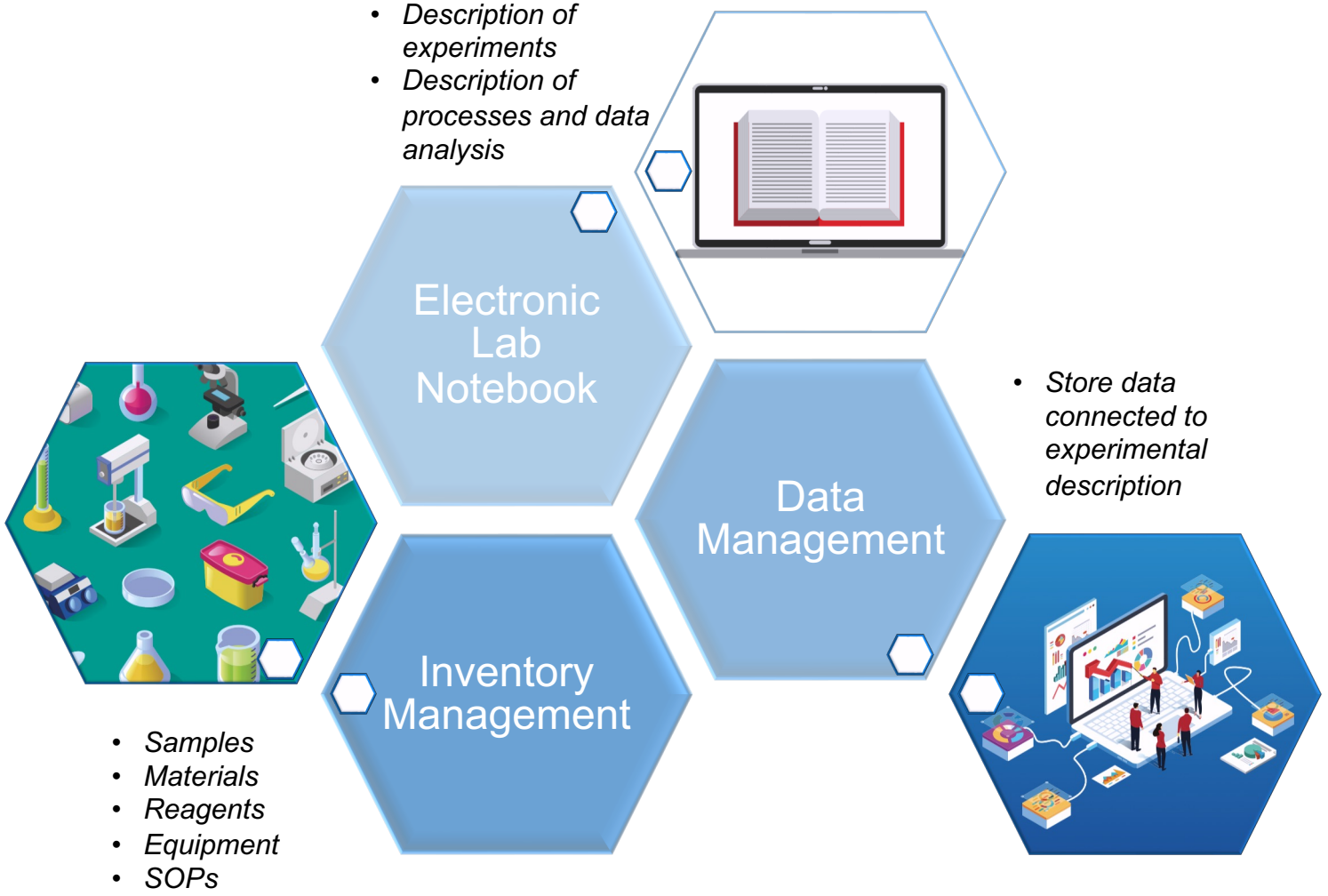


The ETH Scientific IT Services data management solution for research groups

openBIS info

- Developed at **ETHZ** since 2007.
- Used in **different quantitative scientific disciplines** by users inside the **ETH domain** and **beyond**.

openBIS: a complete solution towards FAIR data management

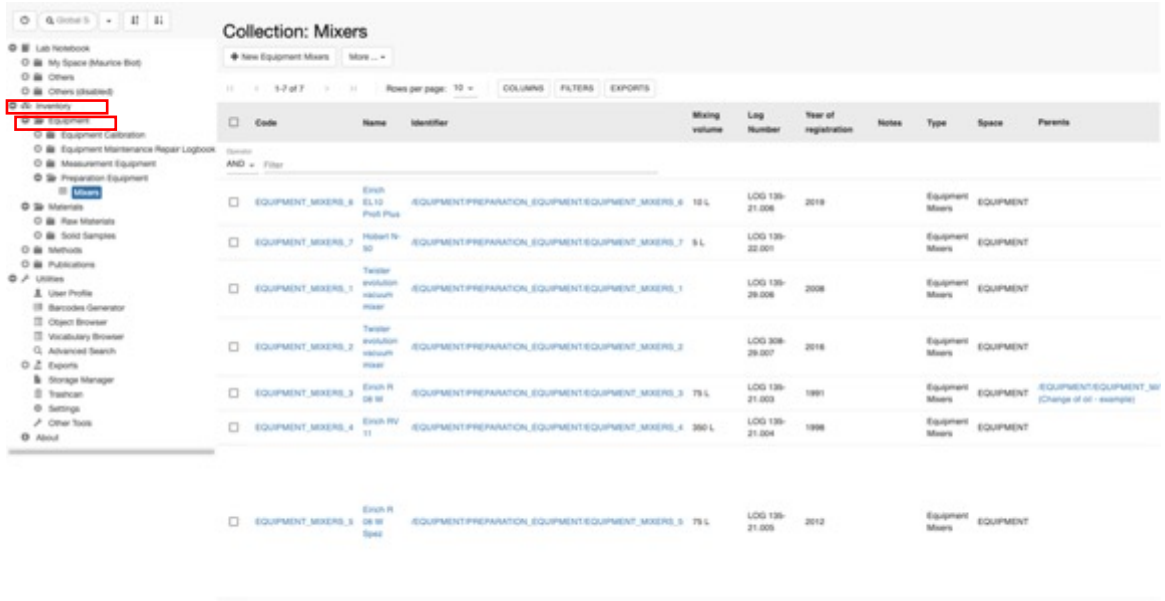


https://openbis.ch

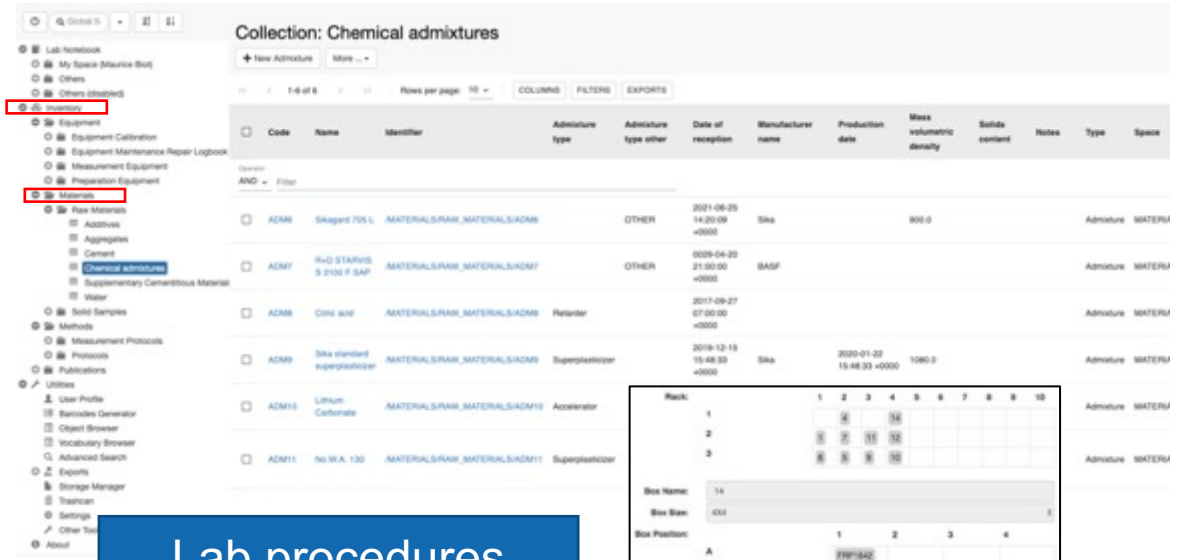
Inventory management

openBIS

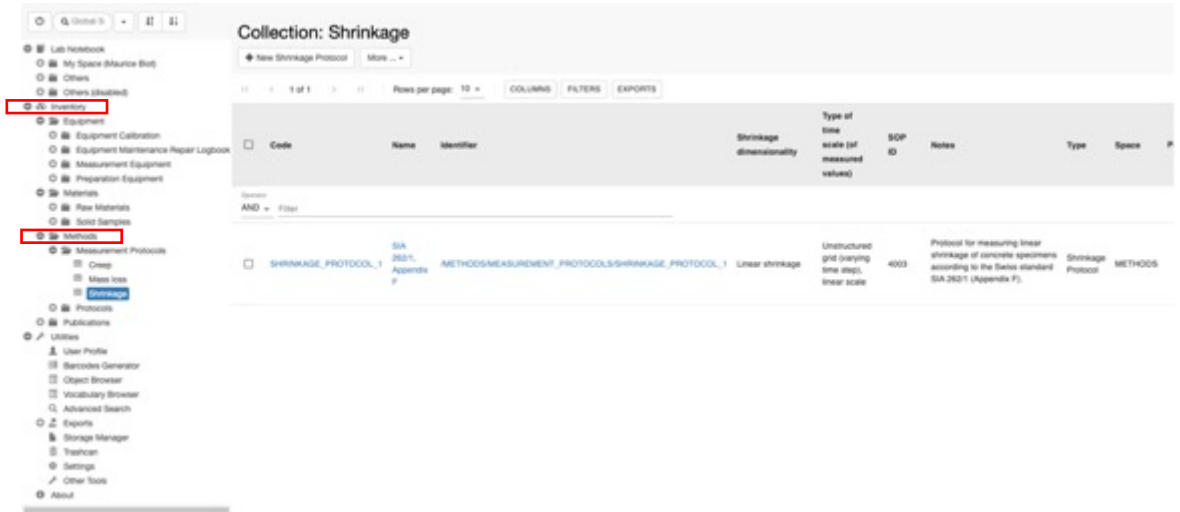
Lab equipment



Lab samples & materials

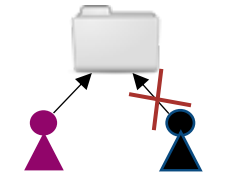


Lab procedures



Samples' storage manager

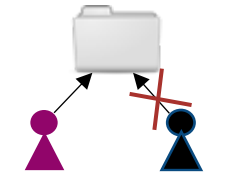
Barcode reader



User rights management

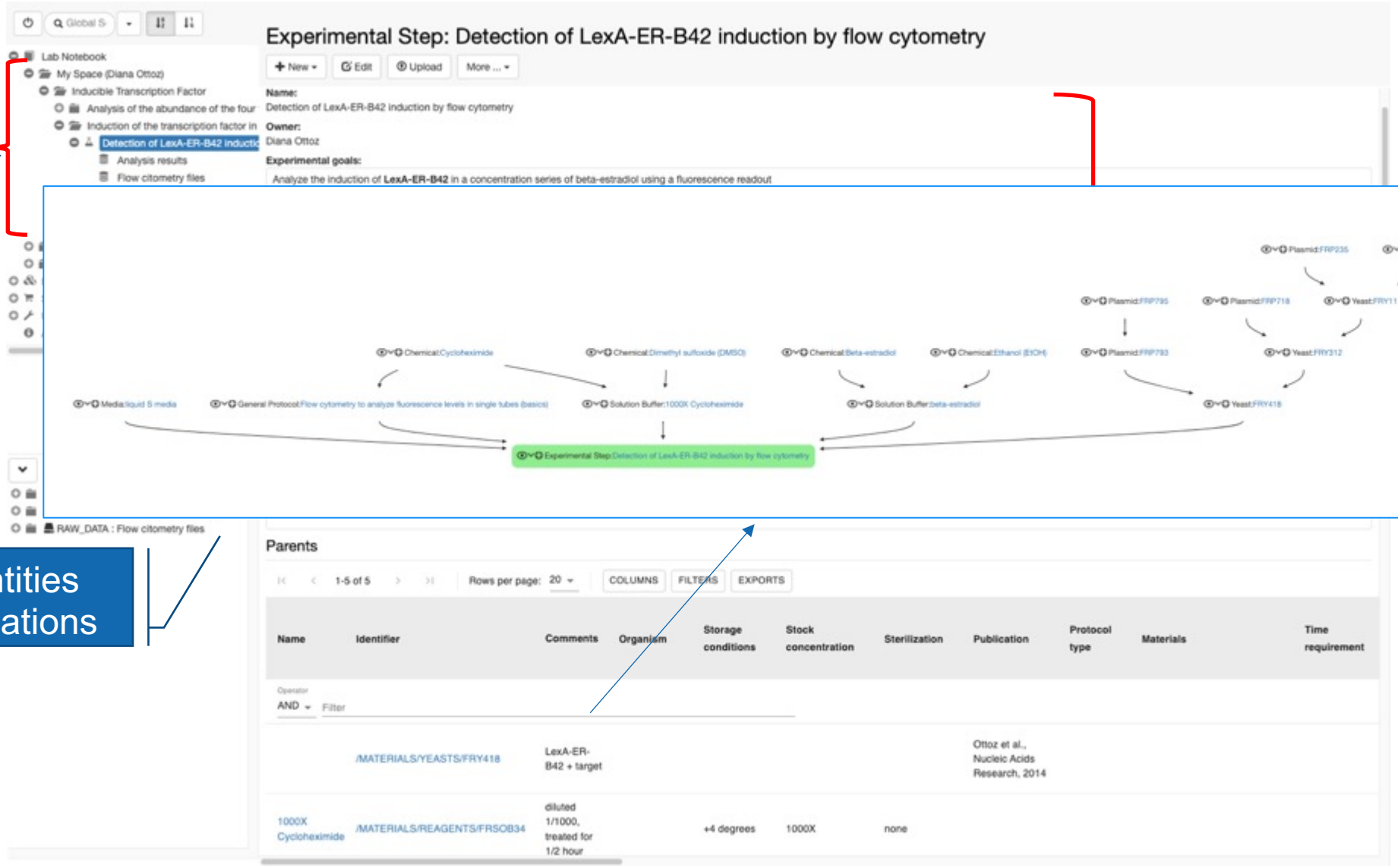
Electronic Lab Notebook

Personal folder



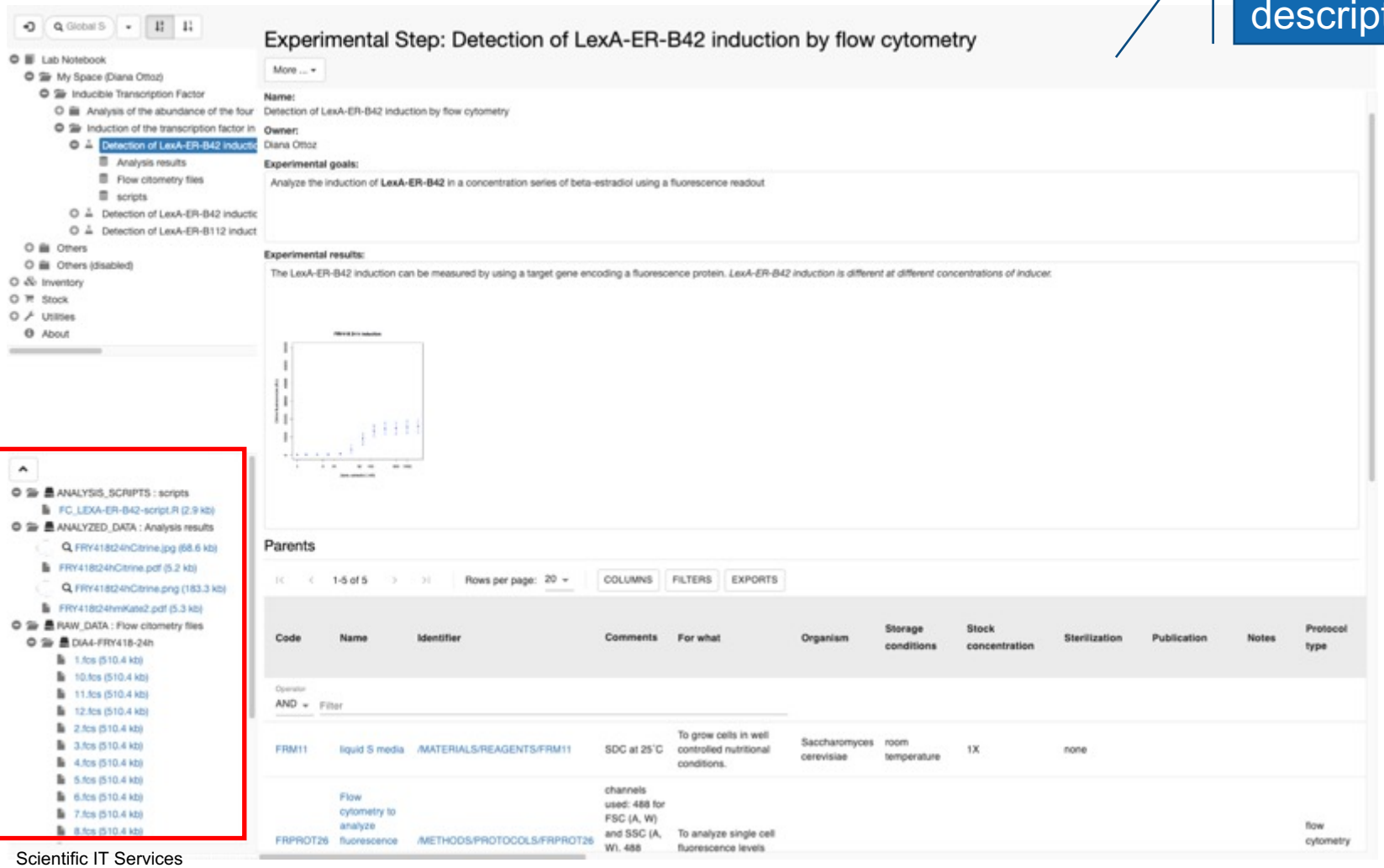
User rights management

Entities relations

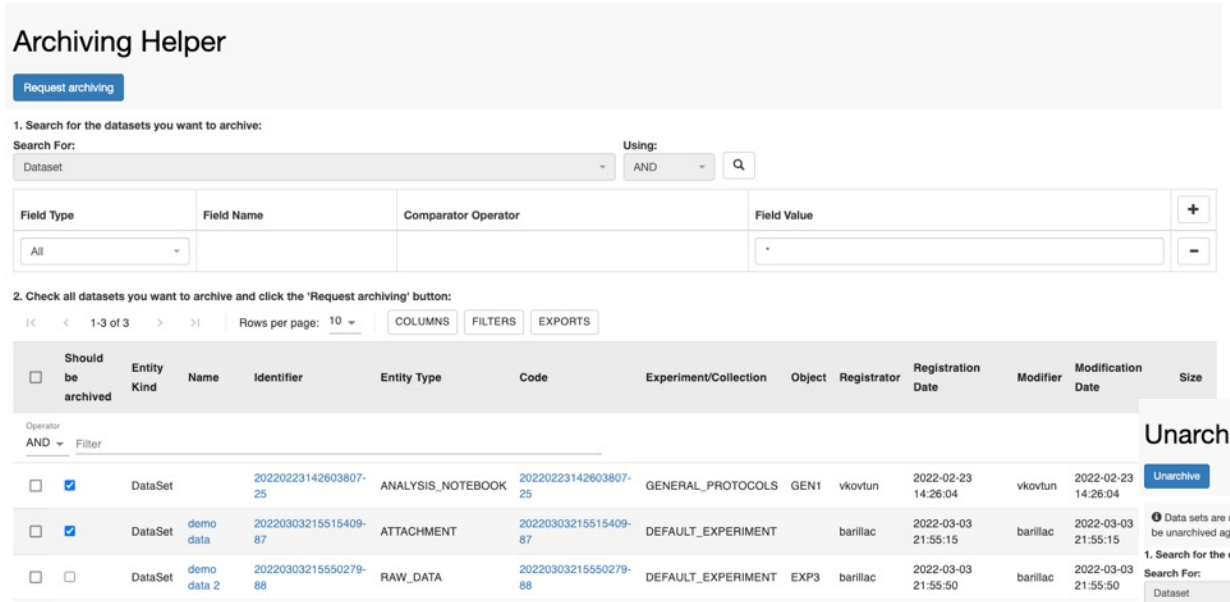


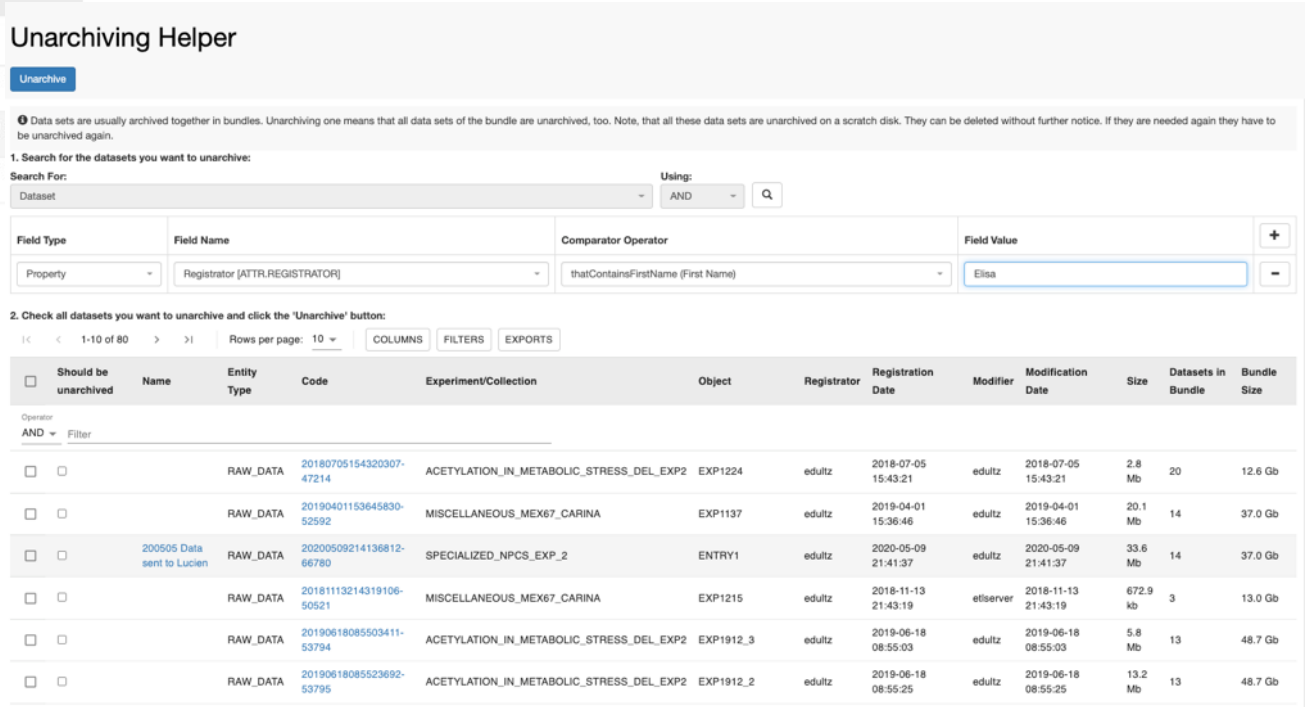
Data management

- Data are always connected to experimental descriptions

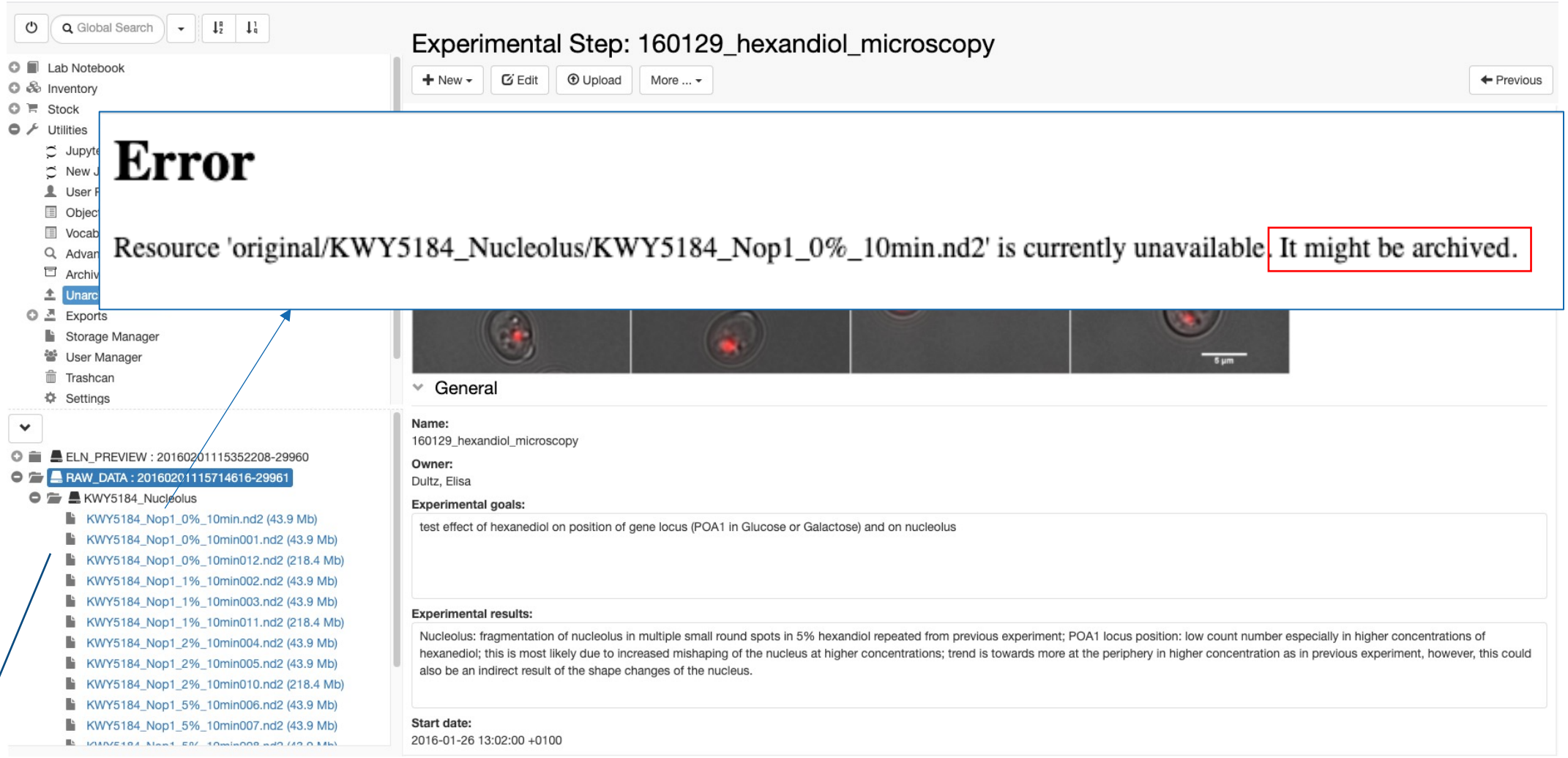


Data archiving and unarchiving from LTS via openBIS (for ETHZ)



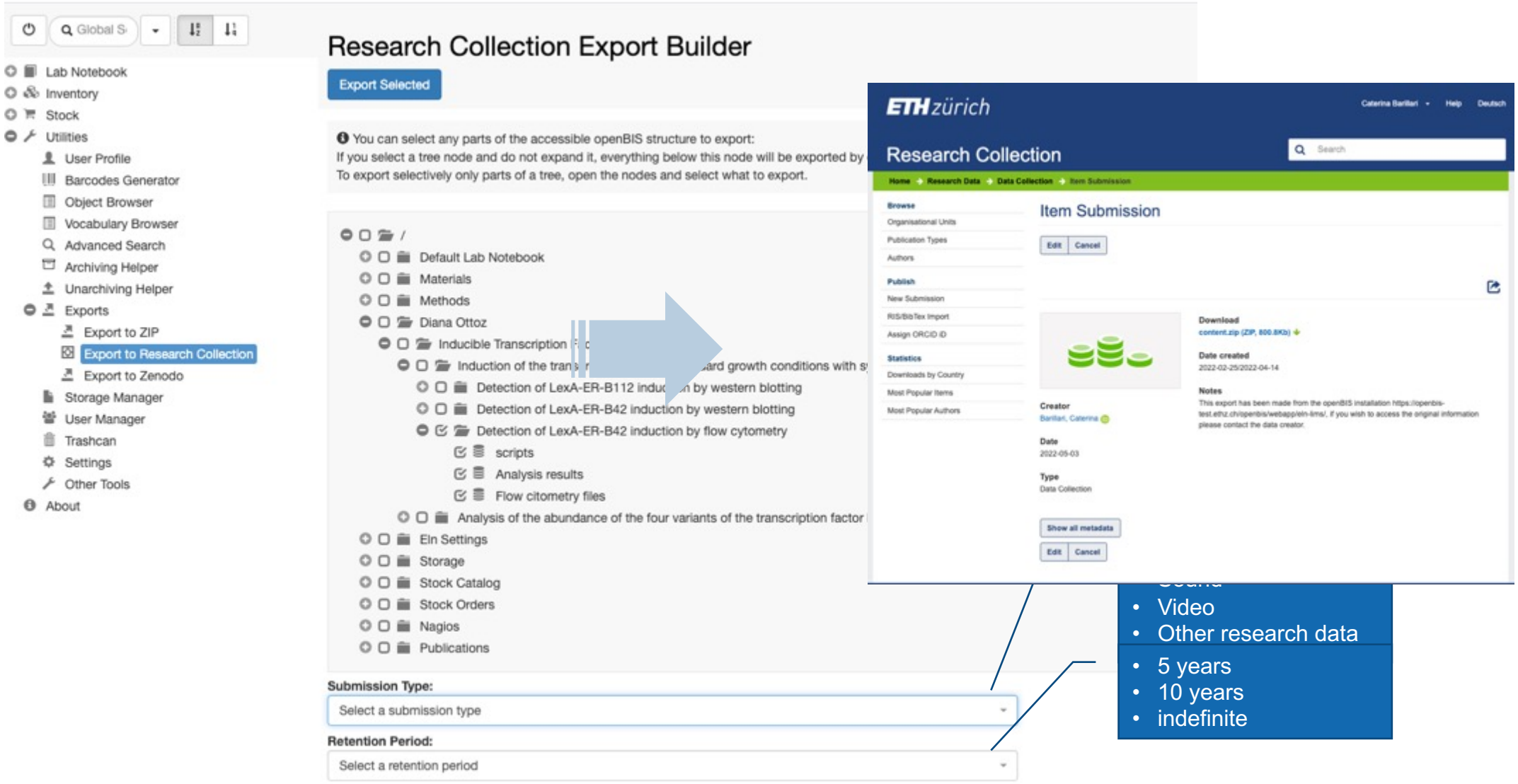


Access to archived data (for ETHZ)

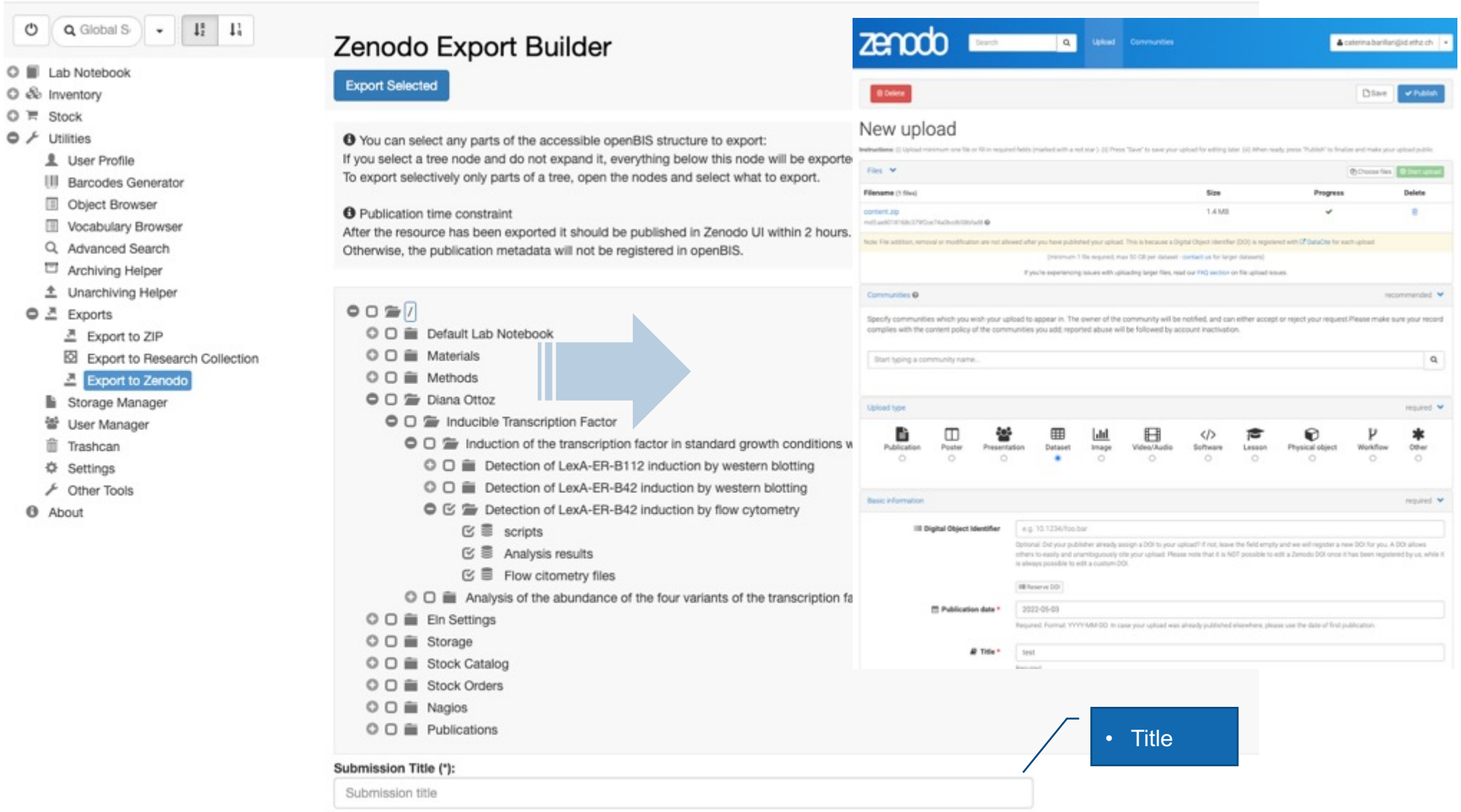


Archived data still visible

Data publication: export to ETH Research Collection (for ETHZ)



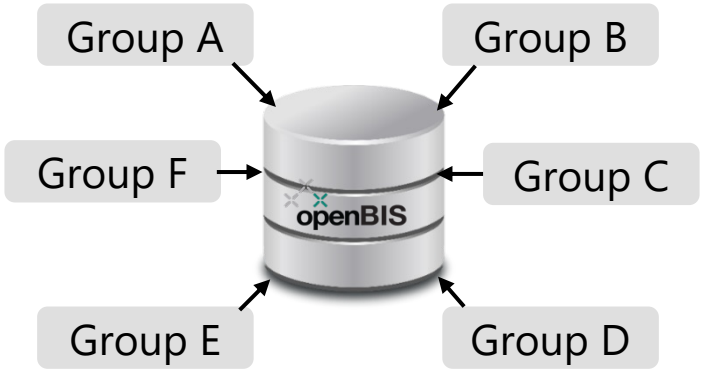
Data publication: export to Zenodo



RDM services @ ETHZ

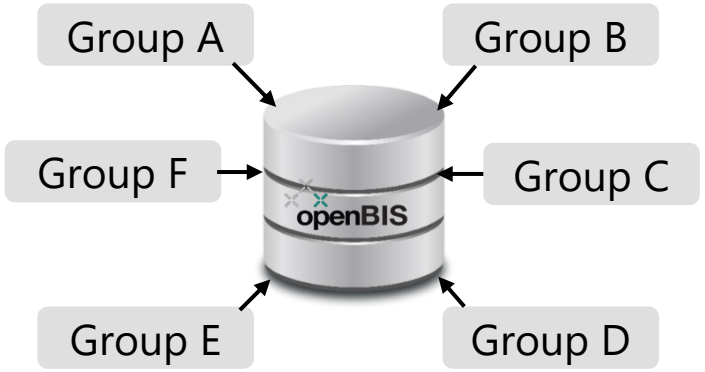
- ETH-wide services provided since 2018

ETH Research Data Hub (RDH)



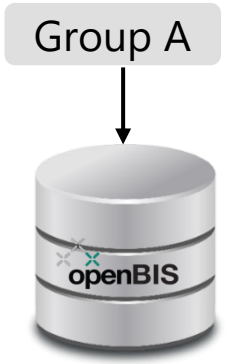
For all ETH groups
- Centrally managed
- Shared resource
- Limited Customization
- No sensitive data

ETH Departmental Data Hub (DDH)



For groups of a certain department/institute
- Centrally managed
- Shared resource
- Department customization

ETH Research Data Node (RDN)



For single ETH research groups
- Individually managed
- Dedicated resource
- Individual customization

A national RDM service for the academic community

- Service establishment funded by a swissunsiversities P5 project
- Project: 2018-2020
- Service start: 2021

Cloud-hosted openBIS

- Virtual servers per research group, institute or institution
- Optionally with JupyterHub server for analytics

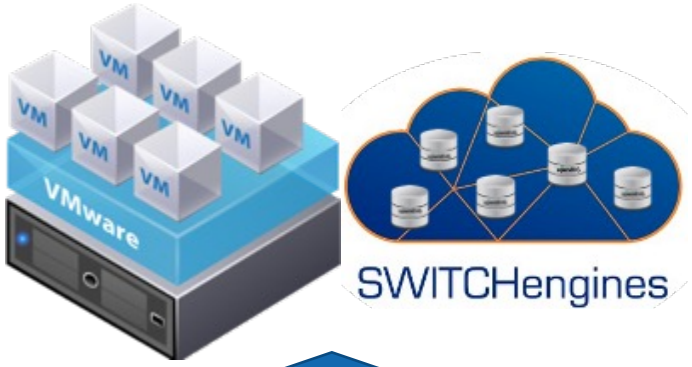
Self-hosted openBIS
Support for set up on local IT infrastructure

Current customers

Balgrist University Hospital
Empa
University of Zurich
UNIVERSITÄT BERN
PAUL SCHERRER INSTITUT PSI
zhaw

https://openrdm.swiss/

What do the SIS's RDM services provide?



Installation on ETH ID infrastructure or SWITCHEngines

Maintenance + upgrades

Consulting

Tailored data modelling

- Trainings
- Users' support

How do the SIS's RDM services help researchers?

Support for all phases of the data life cycle

Single platform for data & information management

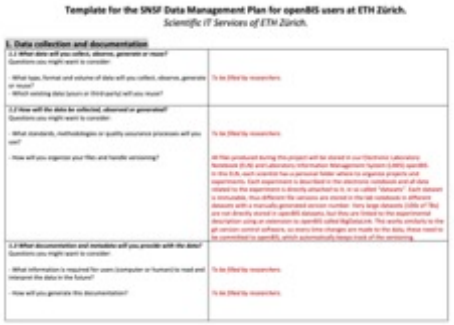
Structured & standardized data organization

Easy access to LTS (ETHZ)

ETH zürich
Research Collection

zenodo

Simplified data publication & sharing



Support for DMP writing

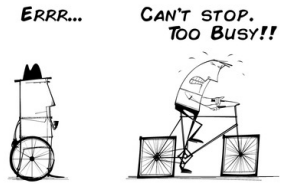
https://sis.id.ethz.ch/services/rdm/SNSF-DMP-openBIS-template.pdf

Questions on openBIS and RDM services provided by SIS?

Code management

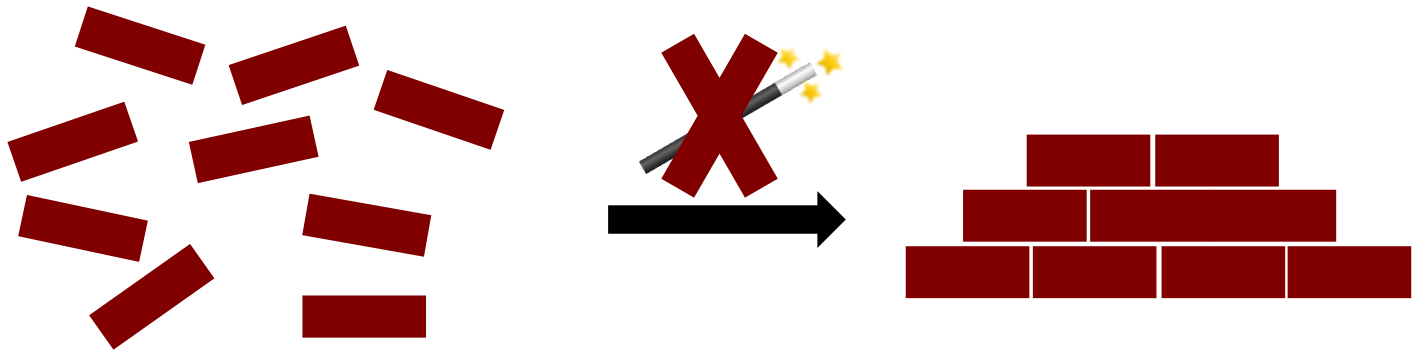
- Code management is an integral part of active data management
- We did not forget about it, it is a topic for next week's workshop on ***Reproducible scientific computing and data analysis***

Take home messages



TOO BUSY TO IMPROVE?
WorkC@mpass

- ❑ Data management should be an integral part of you work
- ❑ Try to establish standard procedures in the group, if not yet in place
- ❑ RDM requires **WORK** & **TIME**, but the time spent on this is an **investment** for the future!



Contact us for consultations / trainings on: data management, version control, reproducible computational workflows or data science support

sis.helpdesk@ethz.ch

Why is this all important for me?

Ok, time to start writing up my **PhD thesis**. Wow, 3 years of work. Let's begin..

I have a request from a reader to check the analysis I presented in this paper. Now, where are they?

image: Freepik.com

These results will make an excellent **publication**! Let me find all the data I need

My ex-PhD supervisor wants to know where my data are, so their new student can carry on my work

- ❑ If you have well annotated, organized and backed up you data over time, all this will be easy!

Contacts

Caterina Barillari

caterina.barillari@id.ethz.ch

Henry Lütcke

henry.lutcke@id.ethz.ch

sis.helpdesk@ethz.ch
https://sis.id.ethz.ch/


https://twitter.com/ETH_SIS?lang=en

Any final questions on what we have seen today?

Feedback: https://www.umfrageonline.ch/s/a13b937