University of Basel

BIOZENTRUM
The Center for Molecular Life Sciences

Understanding the Research Data Lifecycle

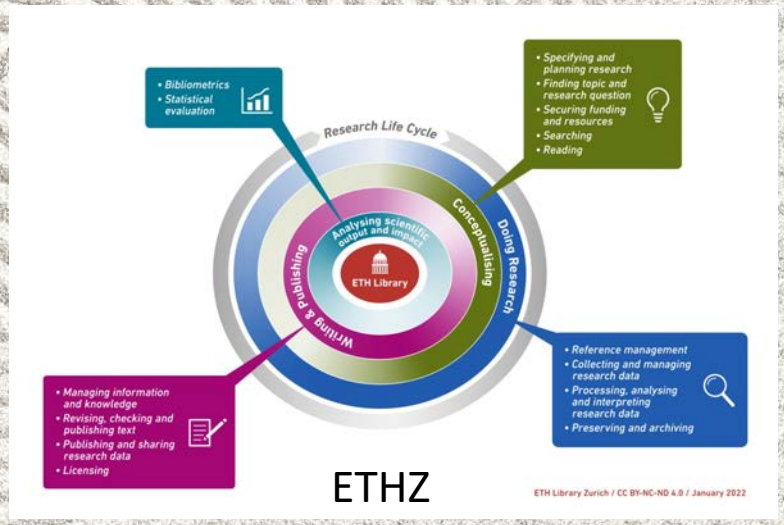
Dr. Michael Podvinec

Biozentrum Research IT

Image: http://www.voicendata.com/wp-content/uploads/2015/07/Data-Management.png

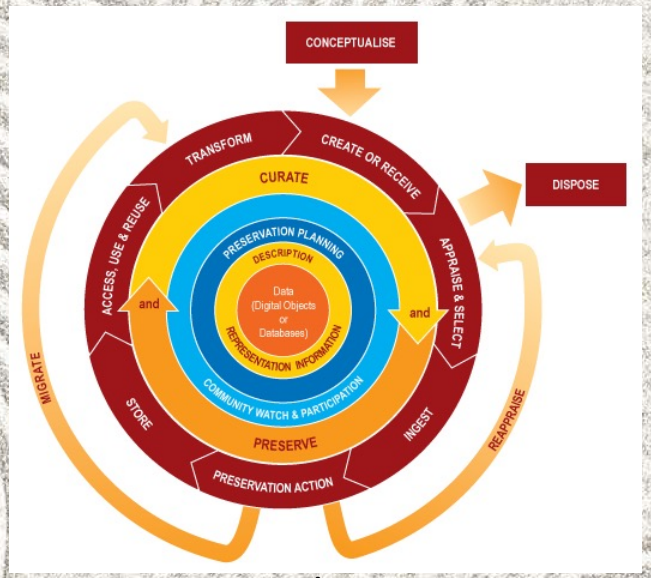
Research Data Lifecycle

Harvard University

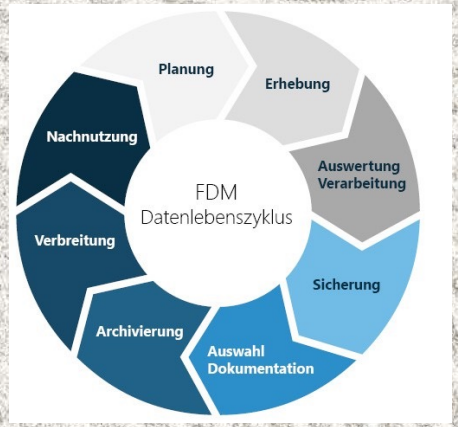


ETHZ

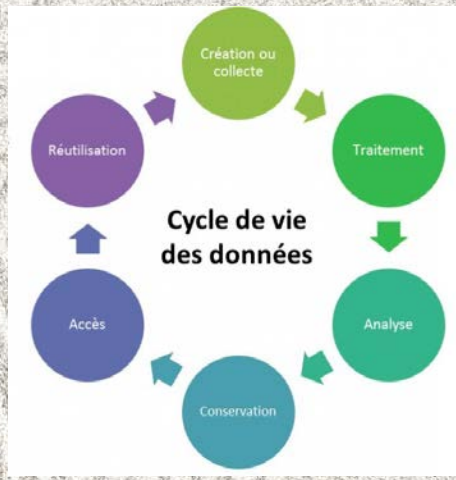
Princeton University



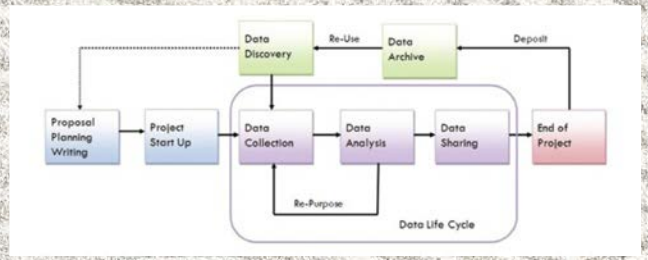
UK Digital Curation Centre



Uni Paderborn

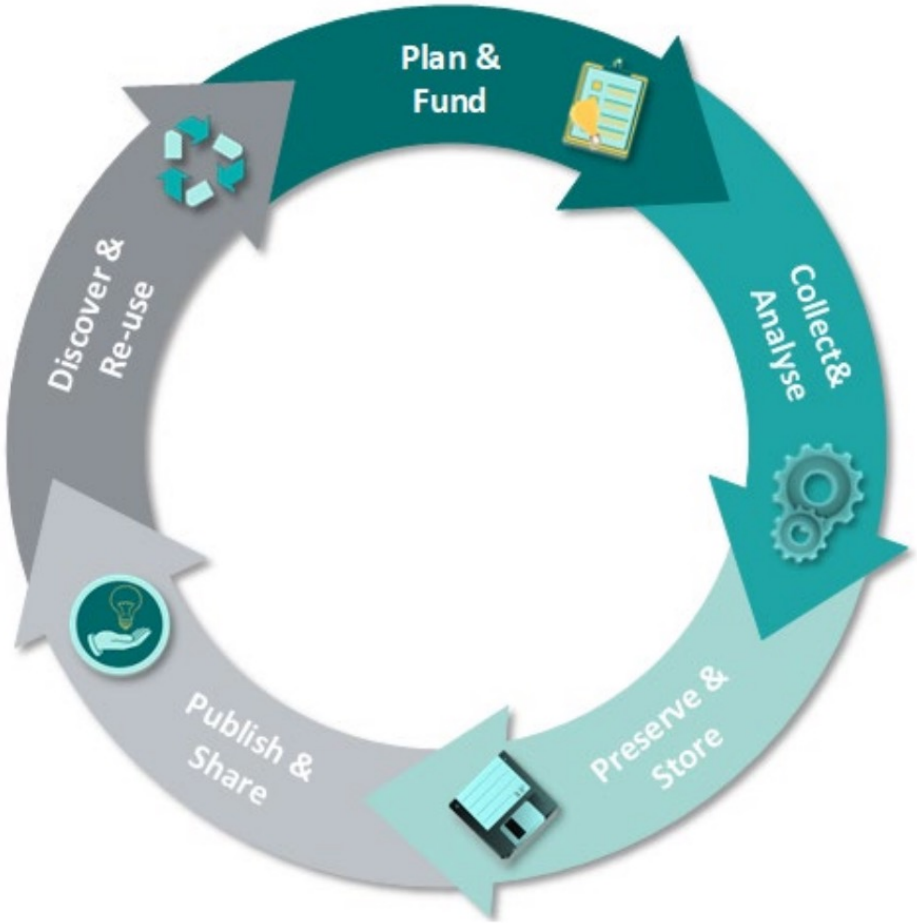


Institut Pasteur



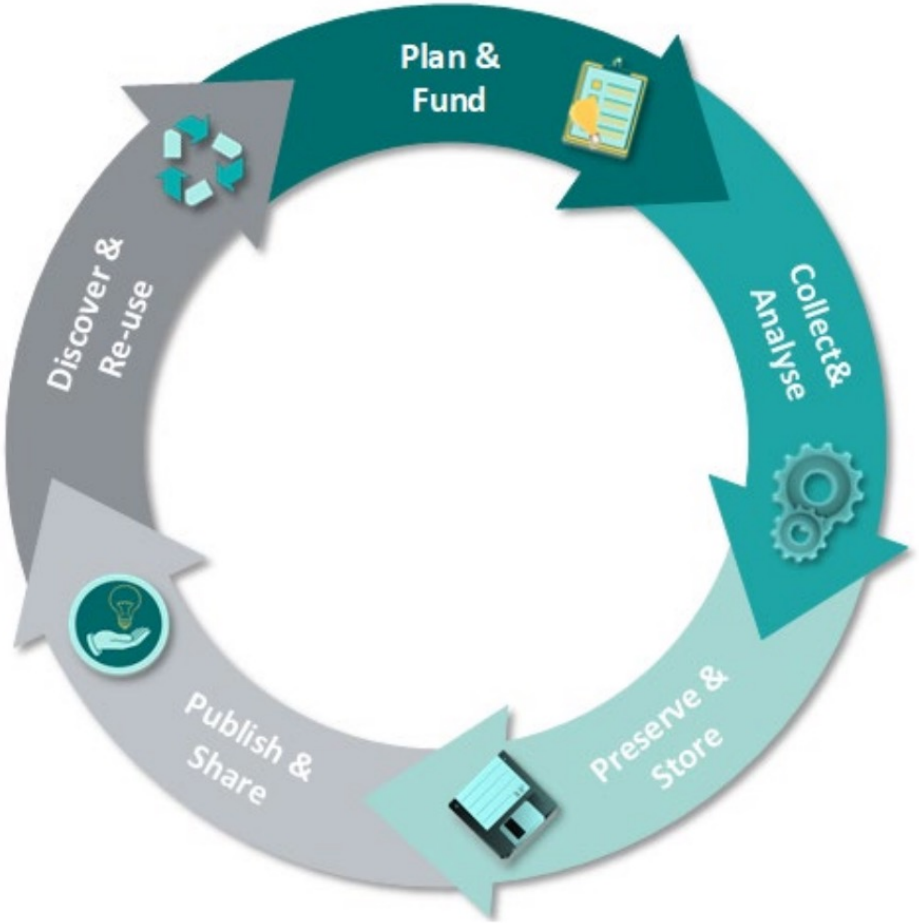
University of Oslo

Research Data Lifecycle: Uni Basel Version



See also: researchdata.unibas.ch

First phase: Conceptualize and Receive



Project Planning Phase

What data? How much data?

What data is existing / can be re-used?

How will data be collected, organized, managed?

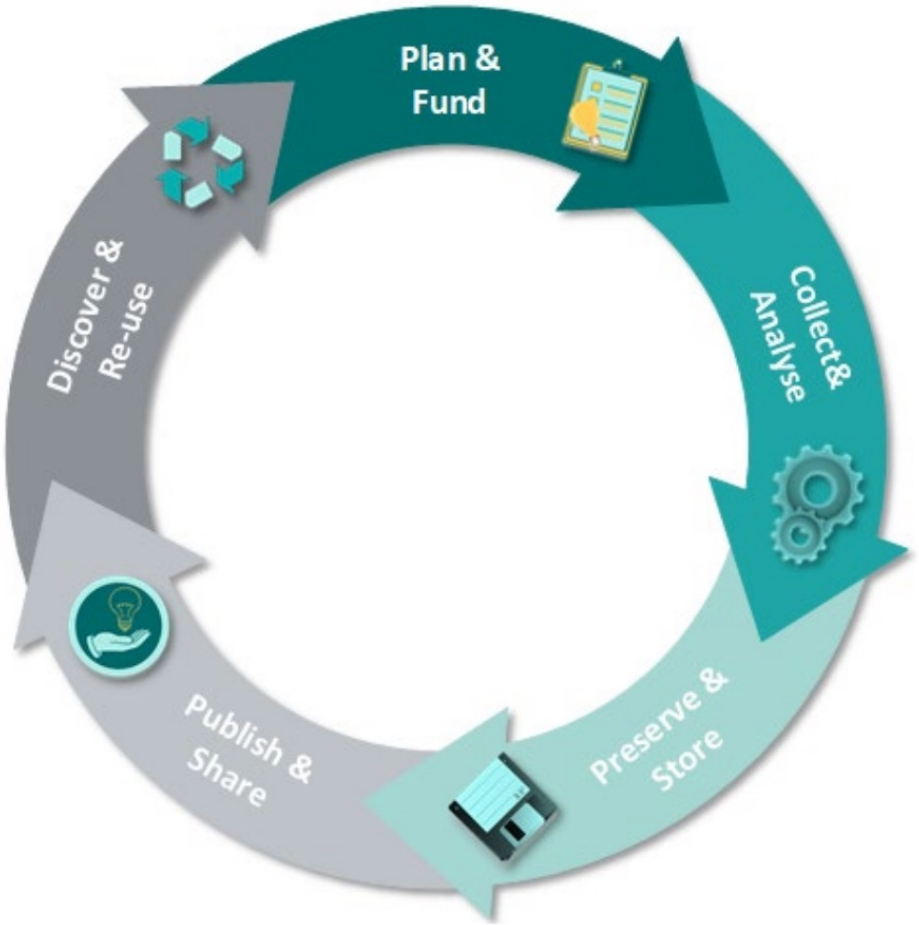
Ethical, legal, security concerns

Data exchange, re-use?

-> Data Management Plans

-> Part of planning a funded project

Phase 2: Create, Evaluate, Collaborate, Store



Active Research

Experiments, observations, secondary materials collected

Data is being processed (transformed, collated, converted, validated, cleaned, etc.)

Metadata is created: Data sets are being described

Data is being analyzed and interpreted: Raw materials are turned into insights. Instruments, code, protocols need to be documented

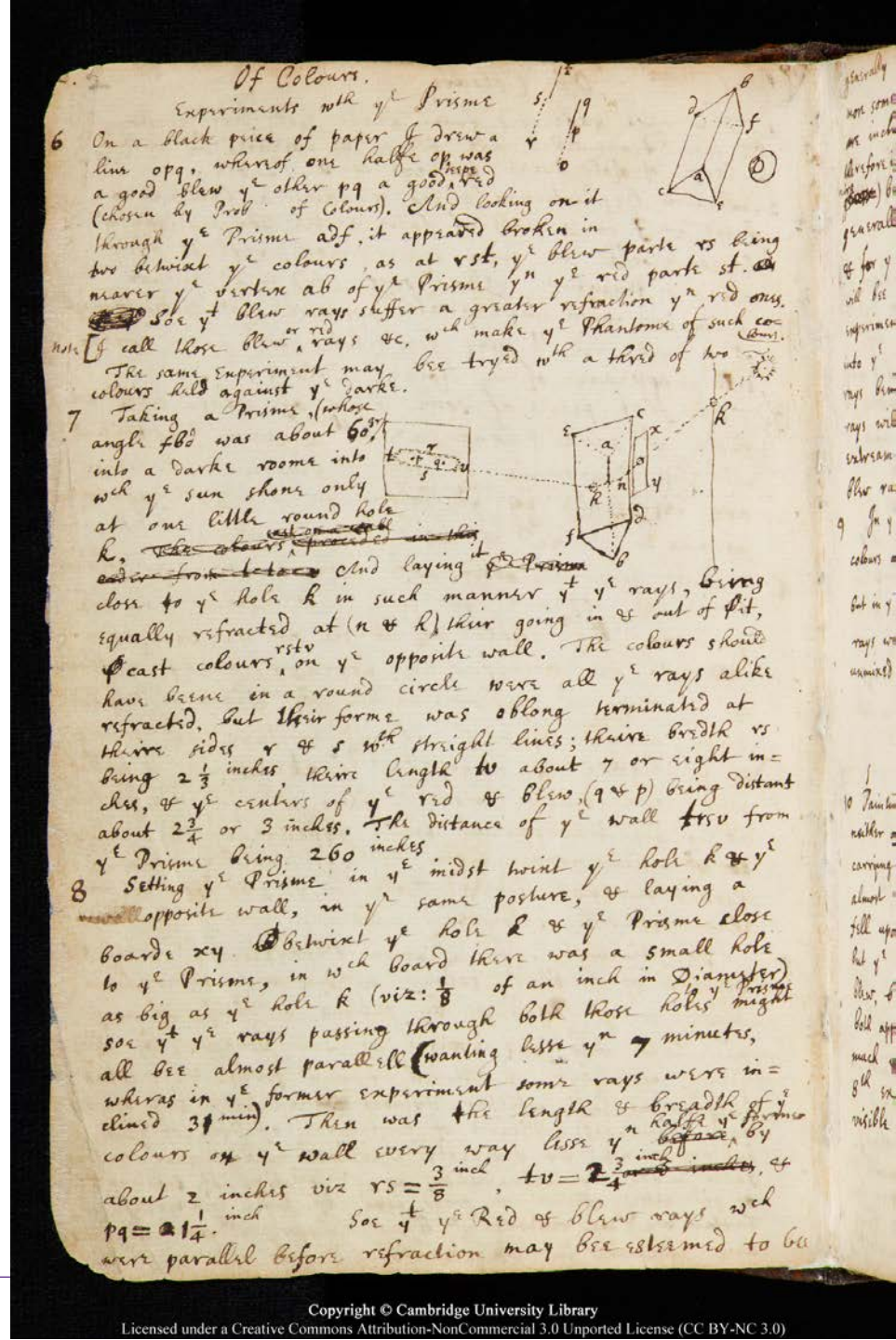
Scientific results are published or patented

Data is stored in active storage.

Documenting your research is part and parcel of the scientific method

Document your work, your observations both expected and unexpected, and your data.

- Laboratory Notebooks
- Drawings, paintings
- Numeric data in tables, books



https://commons.wikimedia.org/w/index.php?curid=822362
https://commons.wikimedia.org/w/index.php?curid=136451

The key to scientific record keeping:

Document your own research in a way that someone else with a comparable education/skill set as yours can understand it

Such as:

- A collaborator
- Your group leader
- Someone who picks up the project after you
- Someone in another group who wants to extend the research you published earlier
- ...
- Or you yourself, 7 years later, trying to write a publication

Scientific Record Keeping Best Practices

Useful & good research records should include these details:

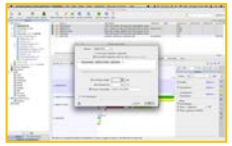
- **What** you did – experimental protocol
- **When** you did it - date
- **Why** you did it – objective
- **How** you did it – methods
- **Who** you are (the person creating the record)
- **What project**(s) this work was part of
- Who conceived of the study (if not yourself)
- Special materials & instruments utilized
- Source of materials & instruments
- Discussion of data – results – expected and unexpected
- Data handling and analyses
- Data interpretation by yourself (and others if pertinent)
- Next steps based on reported results

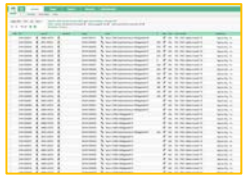
Janet Clark, Ph.D. Director, Office of Fellowship Training NIMH DIRP

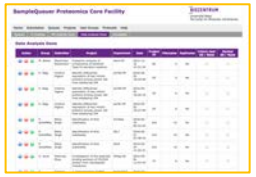
Where to store your data: Data management systems


ELNwiki

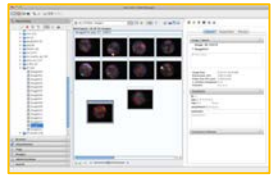

BEAST

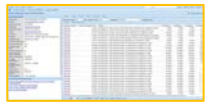

Geneious


PyRAT


SampleQueuer


LabKey


Omero


openBIS

ELN
Electronic Lab Notebooks

- **Information about experiments**
- *Who* did *what when* and *why*?
- What was the *outcome*?
- Protocols
- Descriptions
- Gel pictures, …

LIMS
Laboratory Information Management System

- **Track samples and materials**
- Track provenance
- Traceability of processes
- Procedures and workflows
- Automation is relevant
- Strains
- Vectors, …

SDMS
Scientific Data Management System

- **Store and annotate large datasets**
- Provide programmatic access
- Make datasets searchable
- Analysis and visualization may be built-in
- Microscopy images
- Sequencing data, …

Where to store your data: Storing files safely

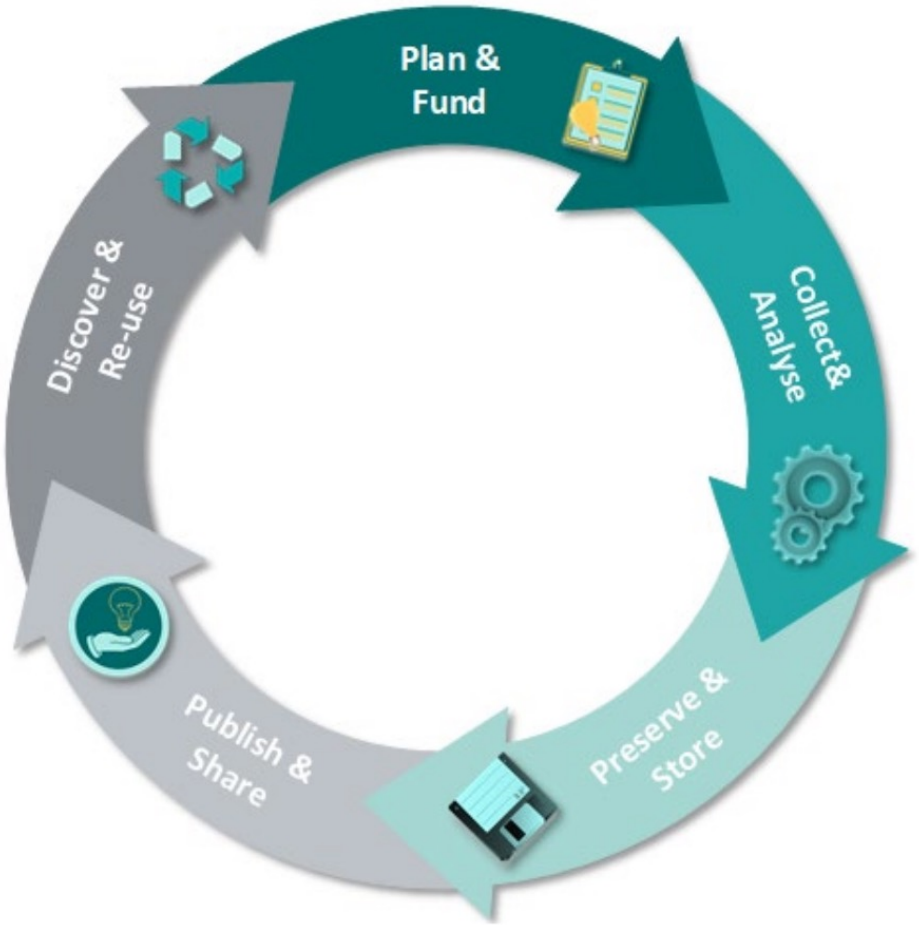
180TB of genomic data on USB disks.
What could possibly go wrong...

When storing data in files, consider the following at the beginning of a project:

- Use centrally managed storage.
- What type, size is your data?
  Which storage system is best used?
- Use a hierarchical and logical folder structure
- Separate active and completed projects,
  delete unused temporary files
- **Version your files** using a convention or a SCM system (git)
- Use a **naming convention for your file names**
- Use well-known data formats
  (ideally non-proprietary, open-source formats)
- Describe your data with proper documentation and metadata.

USB Disk image: © Chris Dagdigian, BioTeam

Phase 3: Conserve, Annotate, Curate, Archive



Preserve & Store

Active storage vs. preservation for long term:

Completion of metadata annotation

Selection of data with long term value

Convert files into the best file format

Prepare for deposition

Archive data to be retained but not published


SEAL OF THE PRESIDENT OF THE UNITED STATES

Size of this preview: 800 × 536 pixels. Other resolutions: 320 × 214 pixels | 640 × 429 pixels | 1,024 × 686 pixels | 1,280 × 857 pixels | 7,000 × 4,687 pixels.

This file has annotations. Move the mouse pointer over the image to see them.

Original file (7,000 × 4,687 pixels, file size: 2.66 MB, MIME type: image/jpeg); ZoomViewer: flash/no flash

Open in Media Viewer

Technical Data: Type, File Format, Resolution ...

Summary [edit]

Description	President Lyndon B. Johnson signs the 1964 Civil Rights Act as Martin Luther King, Jr., and others, look on
Date	2 July 1964
Source	http://photolab.lbjlib.utexas.edu/detail.asp?id=18031
Author	Cecil Stoughton, White House Press Office (WHPO)
Permission (Reusing this file)	"This image is in the public domain and may be used free of charge without permissions or fees."

Content Description

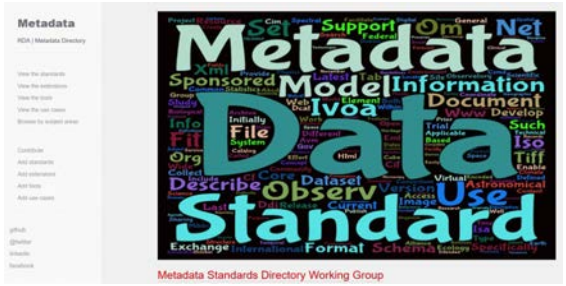
Origin, Creator, License

This is a **featured picture** on the English language Wikipedia (Featured pictures) and is considered one of the finest images

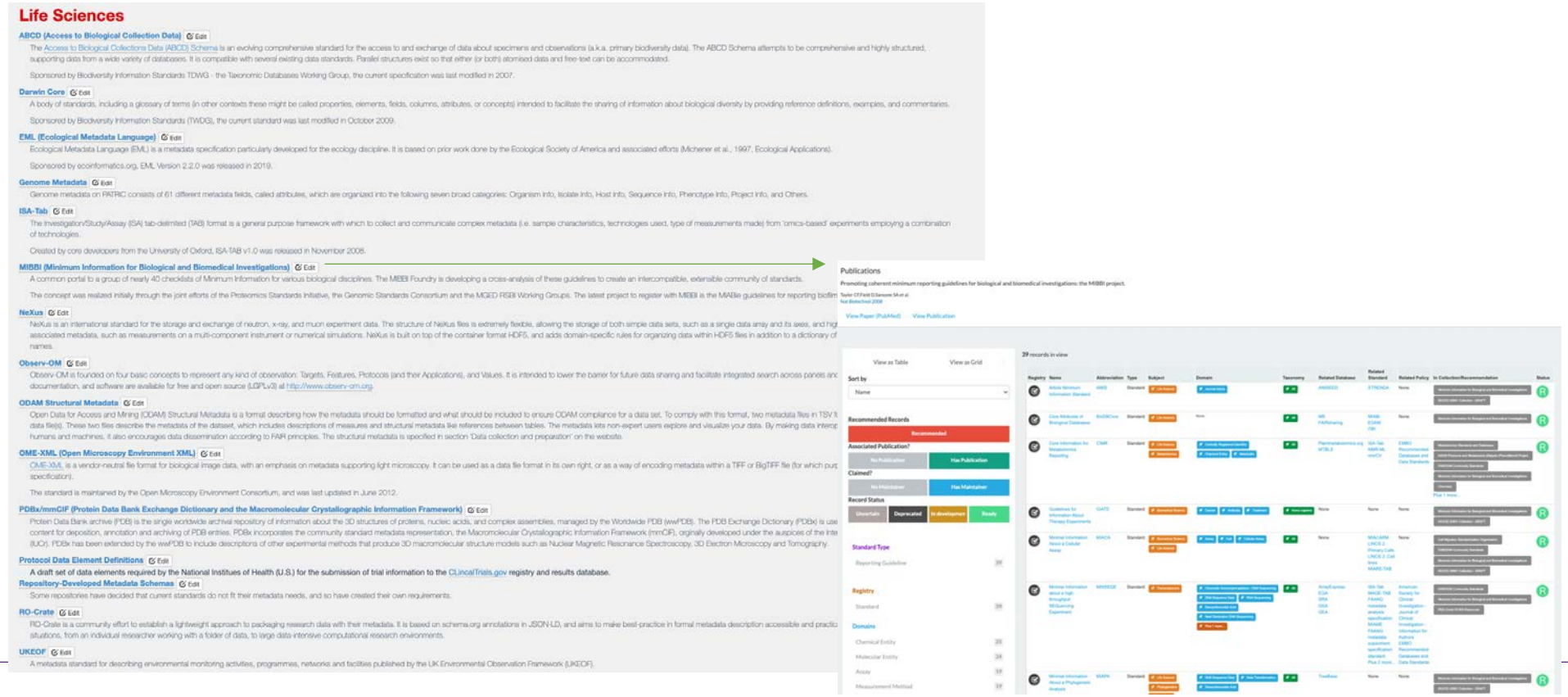
https://commons.wikimedia.org/wiki/File:Lyndon_Johnson_signing_Civil_Rights_Act,_July_2,_1964.jpg

Discipline-specific metadata standards

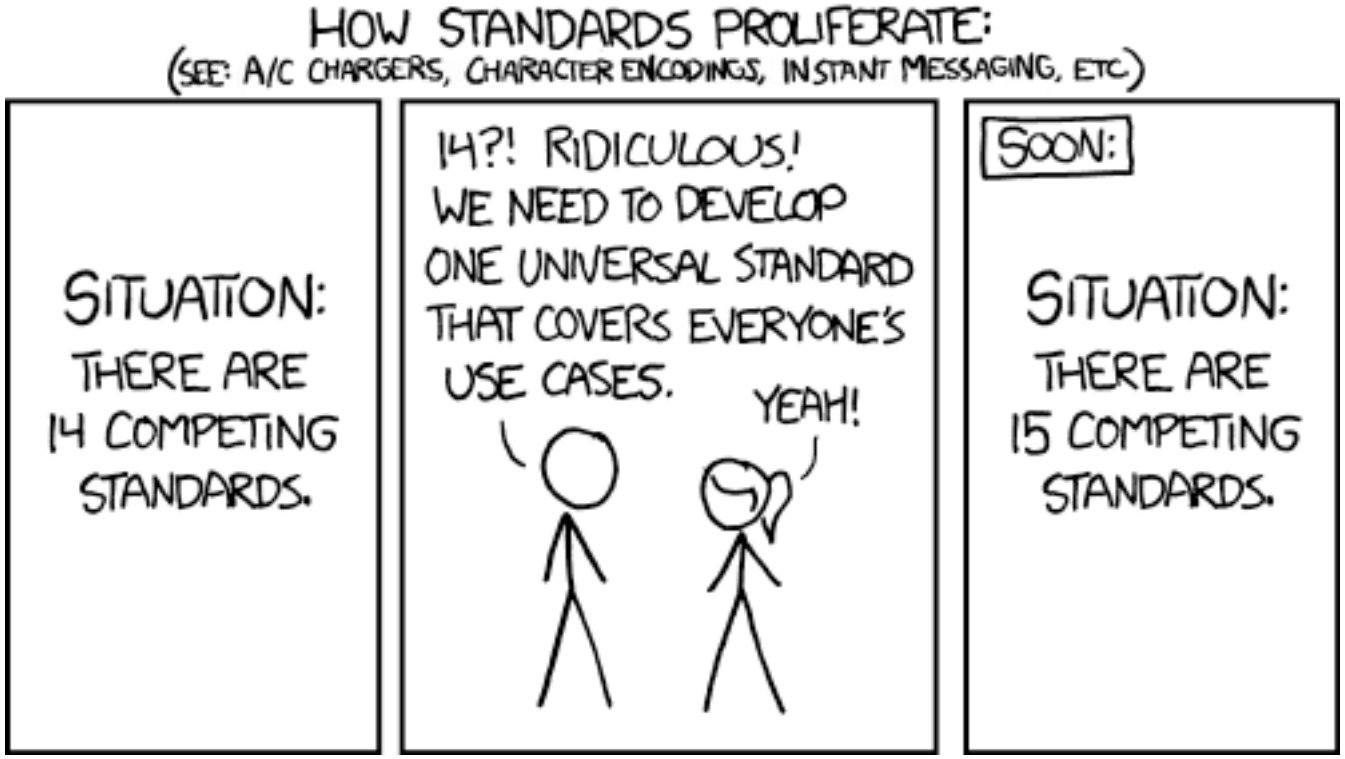
https://rd-alliance.github.io/metadata-directory/



https://fairsharing.org/



Why are there so many metadata standards?



The good news is:

As research fields mature, standards will eventually become generally accepted, e.g PDB files, fasta files, pepXML, OME-TIFF, …

README files

README content:

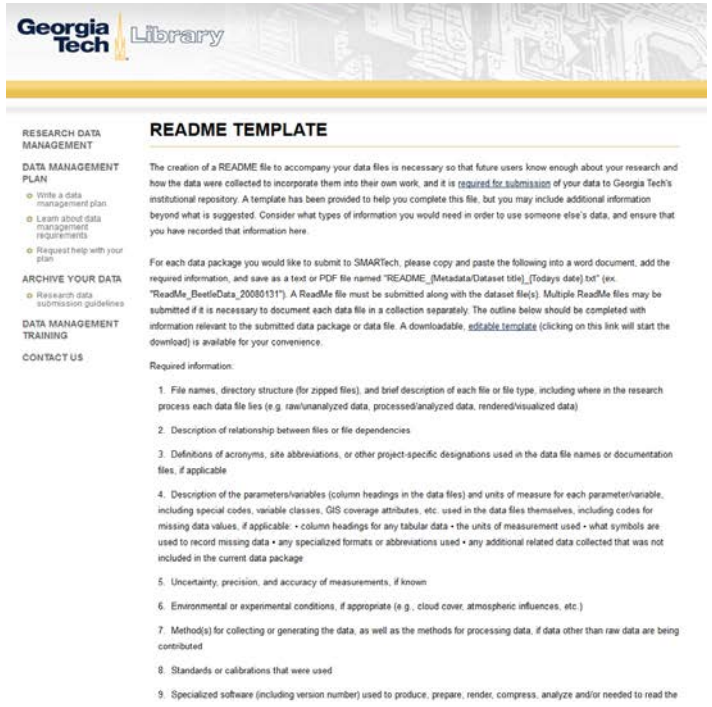
- General information
- Data and file overview
- Sharing and access information
- Methodological information

Best practices

- 1 README / 1 data folder (whenever possible)
- Name the README in accordance with described files
- Write the README as a plain text file (open format)
- Use the same format for multiple READMEs
- Follow your discipline's scientific taxonomic conventions
- Use standardized date formats.
  Suggested: *W3C/ISO 8601 date standard*, `YYYY-MM-DD` or `YYYY-MM-DD-hh:mm:ss`.

A README provides information about data file(s), granting better reusability.

README Template


Georgia Tech Library

RESEARCH DATA MANAGEMENT
DATA MANAGEMENT PLAN
- Write a data management plan
- Learn about data management requirements
- Request help with your plan

ARCHIVE YOUR DATA
- Research data submission guidelines

DATA MANAGEMENT TRAINING
CONTACT US

README TEMPLATE

The creation of a README file to accompany your data files is necessary so that future users know enough about your research and how the data were collected to incorporate them into their own work, and it is required for submission of your data to Georgia Tech's institutional repository. A template has been provided to help you complete this file, but you may include additional information beyond what is suggested. Consider what types of information you would need in order to use someone else's data, and ensure that you have recorded that information here.

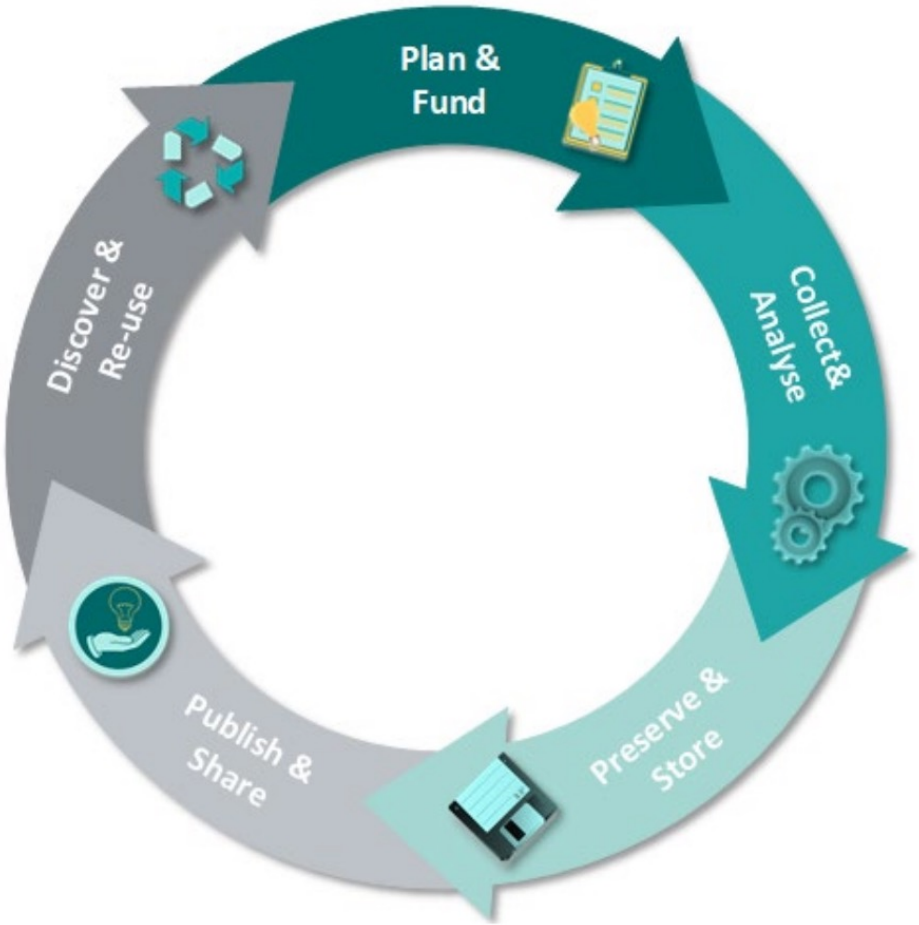
For each data package you would like to submit to SMARTech, please copy and paste the following into a word document, add the required information, and save as a text or PDF file named "README_[Metadata/Dataset title]_[Todays date].txt" (ex. "ReadMe_BeetleData_20080131"). A ReadMe file must be submitted along with the dataset file(s). Multiple ReadMe files may be submitted if it is necessary to document each data file in a collection separately. The outline below should be completed with information relevant to the submitted data package or data file. A downloadable, editable template (clicking on this link will start the download) is available for your convenience.

Required information:

1. File names, directory structure (for zipped files), and brief description of each file or file type, including where in the research process each data file lies (e.g. raw/unanalyzed data, processed/analyzed data, rendered/visualized data)
2. Description of relationship between files or file dependencies
3. Definitions of acronyms, site abbreviations, or other project-specific designations used in the data file names or documentation files, if applicable
4. Description of the parameters/variables (column headings in the data files) and units of measure for each parameter/variable, including special codes, variable classes, GIS coverage attributes, etc. used in the data files themselves, including codes for missing data values, if applicable: • column headings for any tabular data • the units of measurement used • what symbols are used to record missing data • any specialized formats or abbreviations used • any additional related data collected that was not included in the current data package
5. Uncertainty, precision, and accuracy of measurements, if known
6. Environmental or experimental conditions, if appropriate (e.g., cloud cover, atmospheric influences, etc.)
7. Method(s) for collecting or generating the data, as well as the methods for processing data, if data other than raw data are being contributed
8. Standards or calibrations that were used
9. Specialized software (including version number) used to produce, prepare, render, compress, analyze and/or needed to read the

http://d7.library.gatech.edu/research-data/readme

Phase 4: Publish and Share



Publish & Share

Publications based on data should include a data citation / reference how the data can be accessed

Selection of suitable repository and deposition

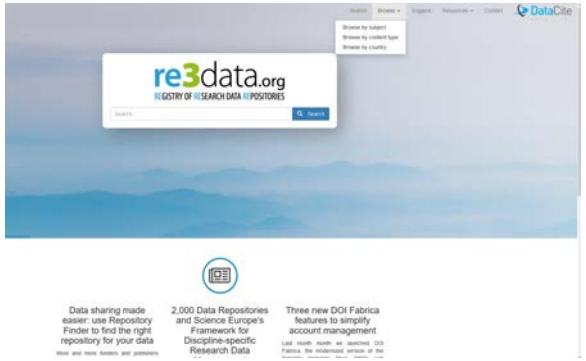
Assignment of Digital Object Identifiers

Decide on a data access model (“As open as possible, as restricted as needed”) and re-use conditions (e.g. CC-licenses)

FAIR data

How to find the best repository for your data:

- Repository **must support FAIR data**
  (Findable, Accessible, Interoperable, Reproducible)
- Choose **domain-specific** rather than general repositories
- Choose **common** rather than institutional repositories
- Choose what the **community** is using (foster re-use)
  - See example list of repositories and re3data.org

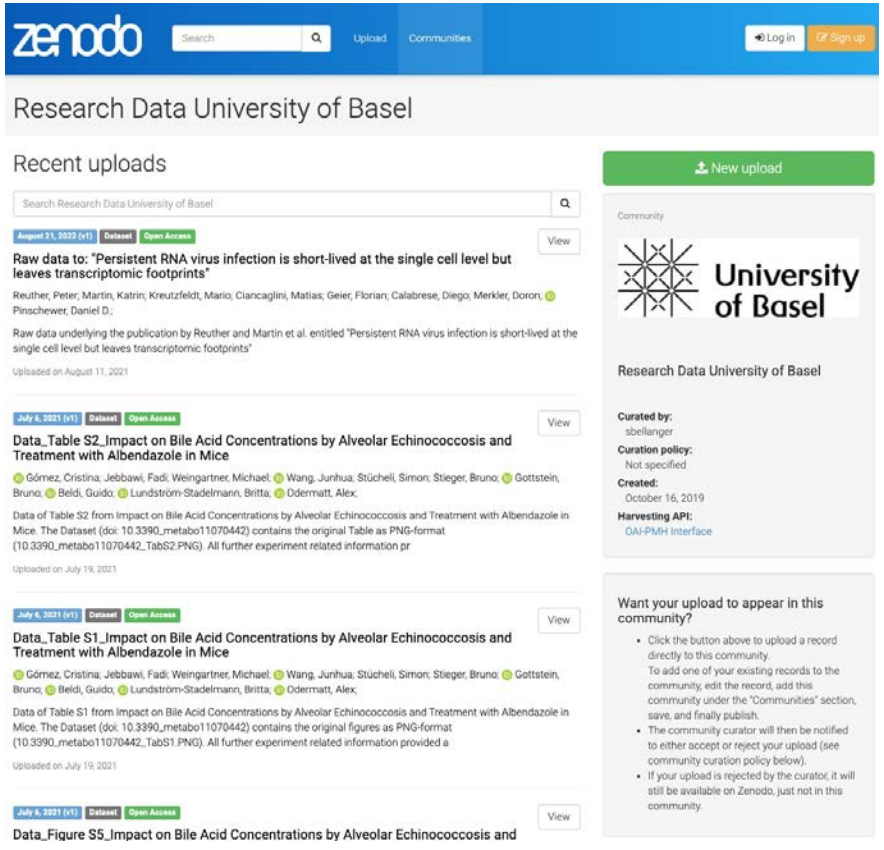


SNSF uses this checklist to validate suitability:

- Datasets have globally unique and persistent identifiers (e.g. DOI)?
- Support intrinsic and submitter-defined metadata?
- Clarity under which license data will be available (e.g. CC0, CC BY, etc.)
- Citation information and metadata always publicly accessible (even for restricted datasets)
- Submission requests intrinsic metadata in a specific format (to ensure machine readability/interoperability)
- Repository has a long-term preservation plan for the archived data

https://www.snf.ch/en/FAiWVH4WvpKvohw9/topic/research-policies, **see points 5.1 / 5.2**

Zenodo: Everybody's general purpose repository



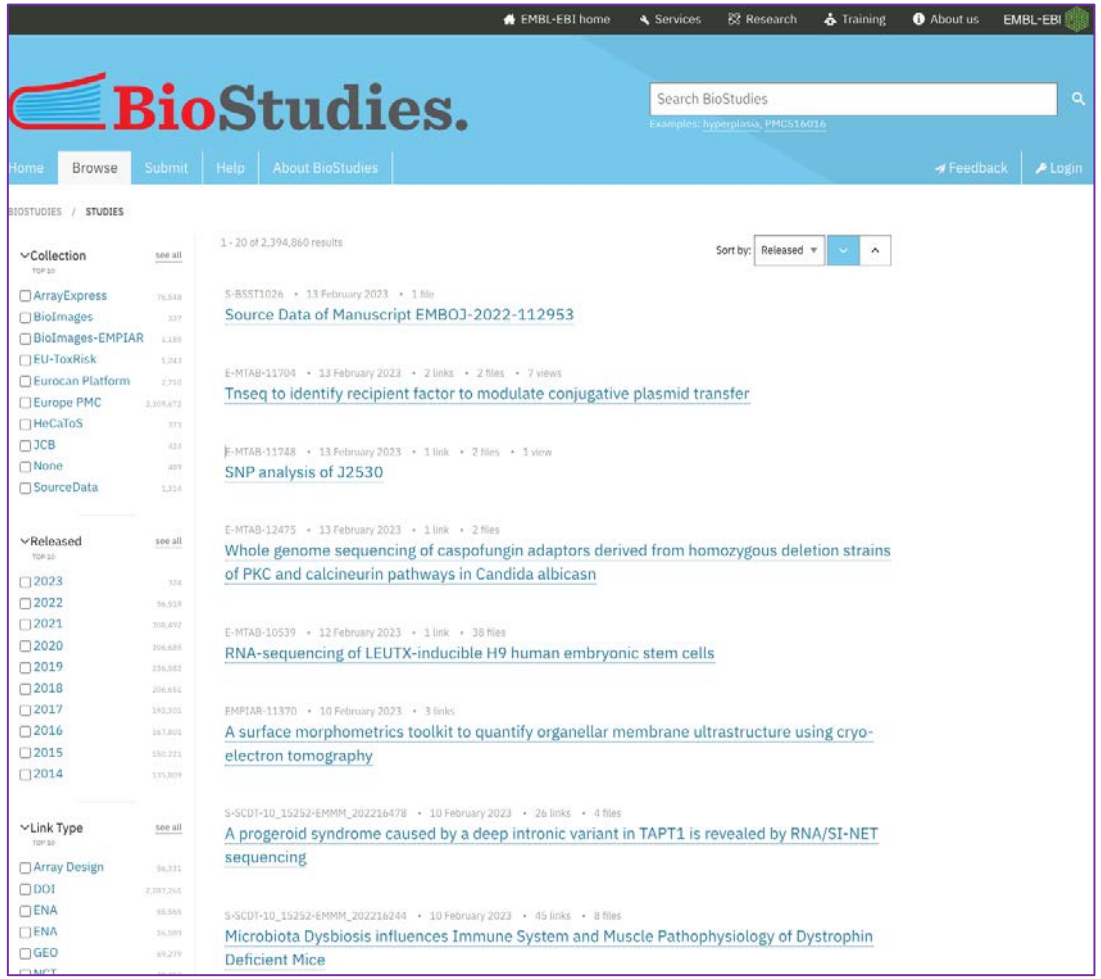
- **Hosted by the CERN**
- **Free of charges**
- **Max 50GB/dataset (unlimited datasets)**
- **Automated DOI assignment**

- **OpenAIRE integration (EC reporting)**
- **GitHub integration**
- **ORCID integration**

- All file formats accepted
- Usage statistics interface
- OAI-PMH protocol (content harvesting)
- 18 petabytes disk cluster
- Each file has 2 replicas on different servers
- 2 independent MD5 checksums per file
- Metadata 12-hourly backup cycle
- ...

sandbox.zenodo.org

EBI BioStudies: a new general repository for biological studies and their data

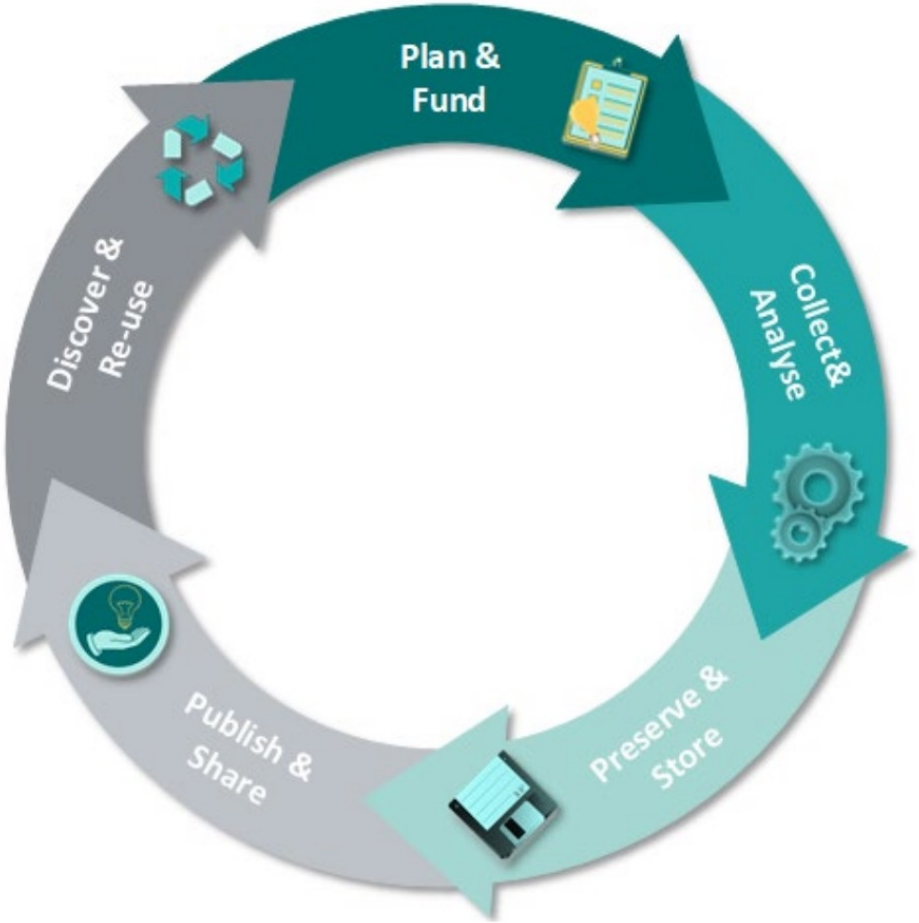


- **Hosted by EBI**

- **Integration with other EBI databases**
- **Integration with PMC**
- **Link to data or contain the data itself**

- Recommended ELIXIR Deposition Database
- Deposition can be manual or automated
- Accompanies structured EBI databases
- Aims to accept all «other» data generated from a biological study as well as a study description
- ...

Phase 5: Discover, Re-use, Learn and Teach



Discover & Re-use

Prepare new studies

Reexamine existing data

Mining data sets for new insights

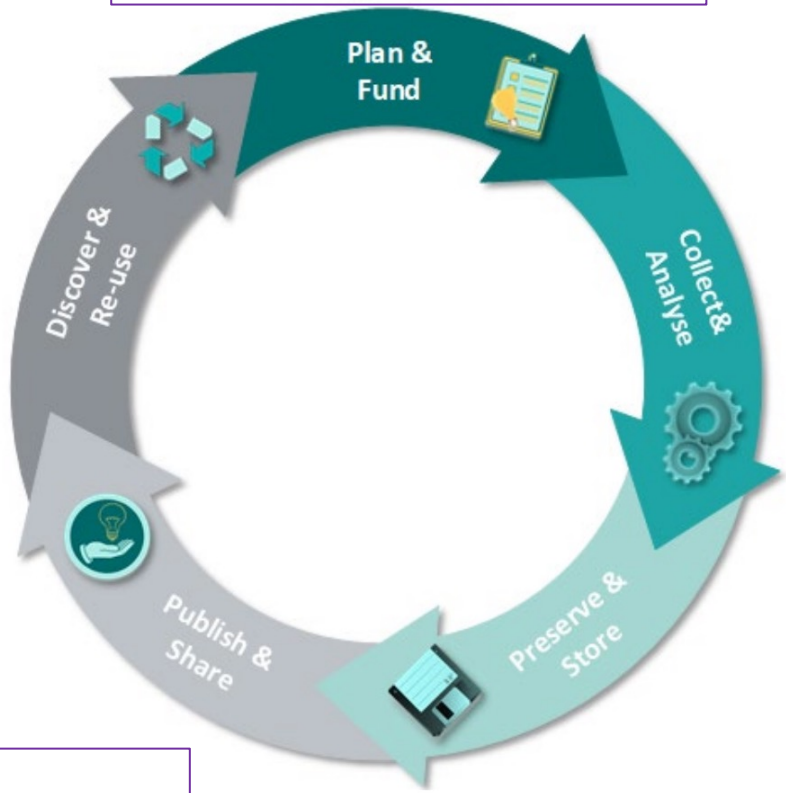
Use data sets for method development

Training and Teaching with existing data sets

Respect the usage license

Quality checkpoints in the research data life cycle

Write a Data Management Plan (DMP)
- Required by funding agencies
- What kinds of data, origin?
- How are they managed/made available during and after project?



Data deposition
- **Curation and transfer:**
  Requires time & effort
  - Complete metadata annotation
  - Collect data sets
  - Ensure open file formats
  - Select repository & upload (SNF guidelines, optimize reuse)

Data annotation and organization
- Done by experimenter
- Metadata about experiment
- Identify data for a paper

… According to the project DMP (good) or a lab/group DMP (better)

Data deposition requirements (SNSF):
- Latest at time of scientific publication
- **All data** underlying paper (i.e. the figures)
- **NOT** protected data (IP, data protection, etc)
- Must be documented with adequate **metadata** that other can understand and reuse


DATA CURATION:
DOES IT SPARK JOY?
NETFLIX | NEW YEAR'S DAY
TIDYING UP
WITH MARIE KONDO

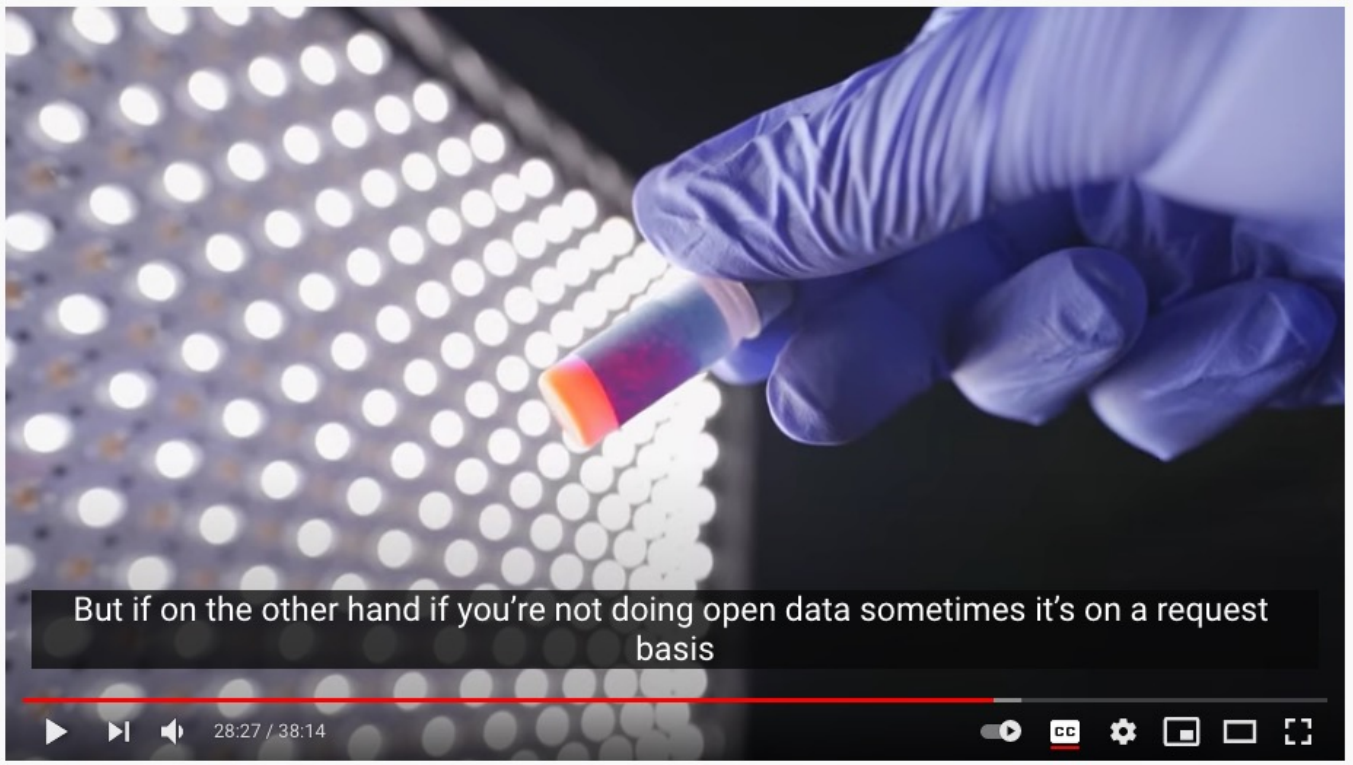
Thank you.

Dr. Michael Podvinec
Biozentrum Research IT

Image: http://www.voicendata.com/wp-content/uploads/2015/07/Data-Management.png

Forgotten Data: A film by Simone Pengue

Prix Média Newcomer 2021 of the Swiss Academy of Sciences



SCAN ME

About this material/acknowledgements

This material is licensed under the CC-BY-SA license

- Exception:
  - Adobe Stock Images under license to University of Basel (Slide 2)

Adobe Stock Images licensed to University of Basel

This presentation is standing on the shoulders of others:

- University of Basel Research Data Network
- Aude Bax de Keating, EPFL Library (at the time)
- Others (references provided wherever possible)

«Data available on request»

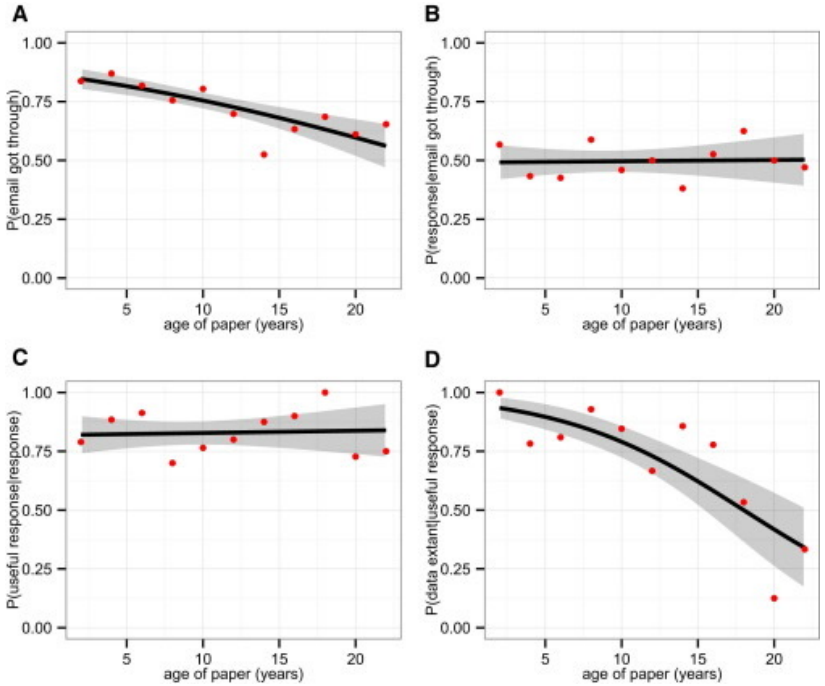
Current Biology
Volume 24, Issue 1, 6 January 2014, Pages 94-97

Report

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines [1, 2], Arianne Y.K. Albert [3], Rose L. Andrew [1], Florence Débarre [1, 4], Dan G. Bock [1], Michelle T. Franklin [1, 5], Kimberly J. Gilbert [1], Jean-Sébastien Moore [1, 6], Sébastien Renaut [1], Diana J. Rennison [1]

Show more



"Policies ensuring that research data are available on public archives are increasingly being implemented(...). These policies are predicated on the idea that authors are poor stewards of their data, particularly over the long term [7], and indeed many studies have found that authors are often unable or unwilling to share their data [8, 9, 10, 11].

However, there are no systematic estimates of how the availability of research data changes with time since publication. We therefore requested data sets from a relatively homogenous set of 516 articles published between 2 and 22 years ago, and found that availability of the data was strongly affected by article age.

For papers where the authors gave the status of their data, the odds of a data set being extant fell by 17% per year. In addition, the odds that we could find a working e-mail address for the first, last, or corresponding author fell by 7% per year.

Our results reinforce the notion that, in the long term, research data cannot be reliably preserved by individual researchers, and further demonstrate the urgent need for policies mandating data sharing via public archives."

What is a Data Management Plan?

Required by all major funders (SNSF, Horizon Europe, NIH) - with minor variations

Example: SNSF Data Management Plan questions

- 1.1 What data will you collect, observe, generate or re-use?
- 1.2 How will the data be collected, observed or generated?
- 1.3 What documentation and metadata will you provide with the data?
- 2.1 How will ethical issues be addressed and handled?
- 2.2 How will data access and security be managed?
- 2.3 How will you handle copyright and Intellectual Property Rights issues?
- 3.1 How will your data be stored and backed-up during the research?
- 3.2 What is your data preservation plan?
- 4.1 How and where will the data be shared?
- 4.2 Are there any necessary limitations to protect sensitive data?
- 4.3 I will choose digital repositories that are conform to the FAIR Data Principles.
- 4.4 I will choose digital repositories maintained by a non-profit organisation.

Help:

- Websites:
  - researchdata.unibas.ch
  - https://wiki.biozentrum.unibas.ch/x/X8KBAw
- Uni Basel Research Data Network / Data Stewards

A living document:

The direction of research is not completely predictable.

The data management challenges are expected to evolve during a project.

Likewise, the DMP can evolve. It has to be finalized at the conclusion of the project

Data Management Requirements from Funders

SNSF:

DMPs required since 2017, format unchanged so far.

Data policy: All data pertaining to publications within the project must be made available in recognized repositories meeting FAIR principles
(see SNSF IR 11.8.1)

DMP (for now) not part of the scientific evaluation, but must conclusively describe data collected, generated and observed at the project end time. (SNSF IR 11.8.3, sep. guidelines).
New: Must only be submitted upon successful scientific evaluation

2025-2028 Priorities: Open research data, infrastructure and services + Skills one of four pillars.

Note: SNSF allows for CHF10'000 for data curation efforts

EU FPs: FAIR DMPs

Horizon 2020 2014-2017: ORD Pilot (Opt-in) – 2017-2020: All projects

Horizon Europe: Open Data Policy ("as open as possible, as closed as necessary")

NIH:

2023 policy: All grants must provide DMS plan. Compliance can be reviewed during reporting intervals and provisions to terminate the agreement.
Established repositories strongly encouraged

Why do I need to keep old data and how long?

Data retention serves you and others:

- Continue working internally with the same data / internal re-use
- Ensure reproducibility of research results
- Keep the data for legal requirements
  - To defend yourself against accusations of plagiarism, fraud, unethical conduct
  - Intellectual Property: Prove priority of invention
- Re-use within the scientific community
  - SNSF requires Data Management Plans with grant proposals
  - SNSF expects funded researchers to generate FAIR data
  - Publishers require publication of data underlying a publication

How long do I have to keep my data?

- Uni Basel: 5 years minimum
- SNSF: 10 years
- -> 10 years is a good general rule

Where do I keep/publish my data?

- Active storage?
- Archival systems / “Deep Storage”?
- Data repositories? -> Allow reproducibility and re-use by applying the FAIR principles

A simple guide to repositories for common life science data types

Data type	Pre-publishing/Internal	Post-publishing/External
Imaging data	OMERO	IDR Bioimage Archive* EMPIAR ??? (size-dependent)
Proteomics/Mass Spectrometry	SampleQueuer PUMA storage	PRIDE & others (during publication)
Genomics	sciCORE HPC storage LabKey PUMA storage	NCBI and EBI sequence databases
"Everything else": Western Blots qPCR Sample information Electrophysiology Video microscopy (…)	PUMA storage JUMBO storage Annotations in ELN LabKey Sample Manager(?) (…)	EBI BioStudies* Zenodo** Dryad***

More:
See *re3data.org* for other relevant research data repositories. The SNSF uses re3data as a reference when deciding on the FAIRness of a repository.

- re3data shows that no persistent identifier is used by these databases. This seems strange – and likely will be updated.

** Zenodo accepts any research data sets up to 50GB per submission and is funded by the EU through OpenAIRE.
There is a «**Research Data University of Basel**» community you can associate your research with.

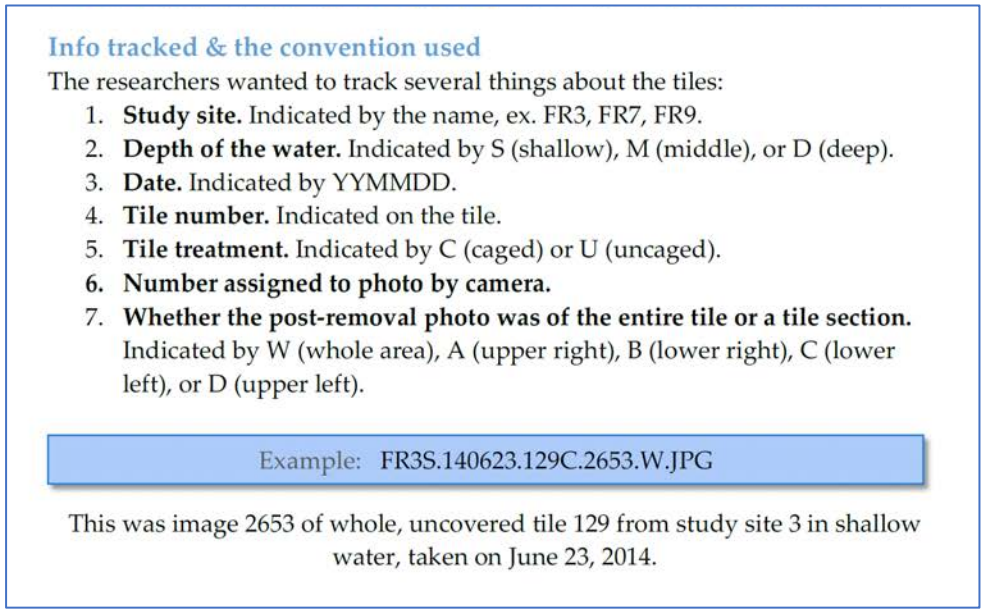
*** Dryad is comparable to Zenodo (but from the US). It has a financial model for submissions > 50GB

Naming your file

File Naming Caveats

Please keep in mind that not all characters are allowed in file names and name length restrictions on file and path level also do apply.
We recommend to keep the path length **below 260 characters and only use safe characters**. Keep the file name at **32 characters or less**.

Develop a file naming convention:



Info tracked & the convention used

The researchers wanted to track several things about the tiles:

1. **Study site.** Indicated by the name, ex. FR3, FR7, FR9.
2. **Depth of the water.** Indicated by S (shallow), M (middle), or D (deep).
3. **Date.** Indicated by YYMMDD.
4. **Tile number.** Indicated on the tile.
5. **Tile treatment.** Indicated by C (caged) or U (uncaged).
6. **Number assigned to photo by camera.**
7. **Whether the post-removal photo was of the entire tile or a tile section.** Indicated by W (whole area), A (upper right), B (lower right), C (lower left), or D (upper left).

Example: FR3S.140623.129C.2653.W.JPG

This was image 2653 of whole, uncovered tile 129 from study site 3 in shallow water, taken on June 23, 2014.

Safe characters:

```
ABCDEFGHIJKLMNOPabcd
efghijklmnop01234567
89-_
```

Don't use:

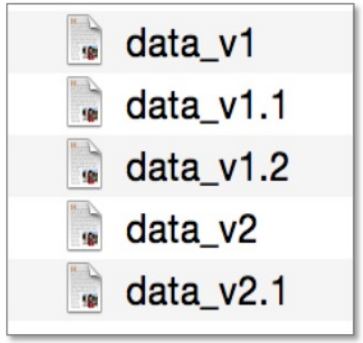
```
< > : ” / \ | ? *
```

Another example: https://www.unibas.ch/de/Universitaet/Administration-Services/Generalsekretariat/Archive-Sammlungen/Universitaetsarchiv1.html

Versioning your file

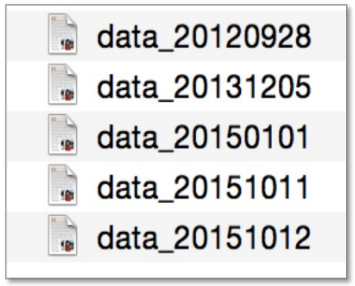
Try to:

Sort



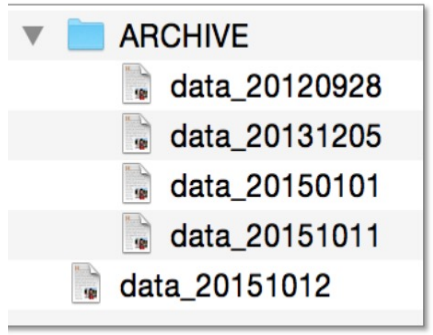
Major changes: ordinal numbers
Minor changes: decimal numbers

Distinguish



Dates distinguish between the different versions.

Separate



Archive older files in a separate folder.

Version control software:

The FAIR Data Principles

Findable

To aid automatic discovery of relevant datasets, (meta)data should be easy to find by both humans and machines and be assigned a persistent identifier.

Accesible

Limitations on the use of data, and protocols for querying or copying data are made explicit for both humans and machines.

Interoperable

(Meta)data should use standardised terms (controlled vocabularies), have references to other (meta)data and be machine actionable.

Reusable

(Meta)data are sufficiently well described for both humans and computers to be able to understand them and have a clear and accessible data usage license.