UiO : University of Oslo Library

Module III:
Data organization, metadata, and documentation

Edina Pózer, Ivana Malovic, Solveig Isis Sørbø and Agata Bochynska

The University of Oslo Library

Materials developed as part of the *Skills development project for research data*:
https://www.ub.uio.no/english/about/projects/rdm-skills/

Today’s course

1) One hour lecture with a short break
2) Q&A session

Please feel free to write comments and questions in the Chat!

Data file structure

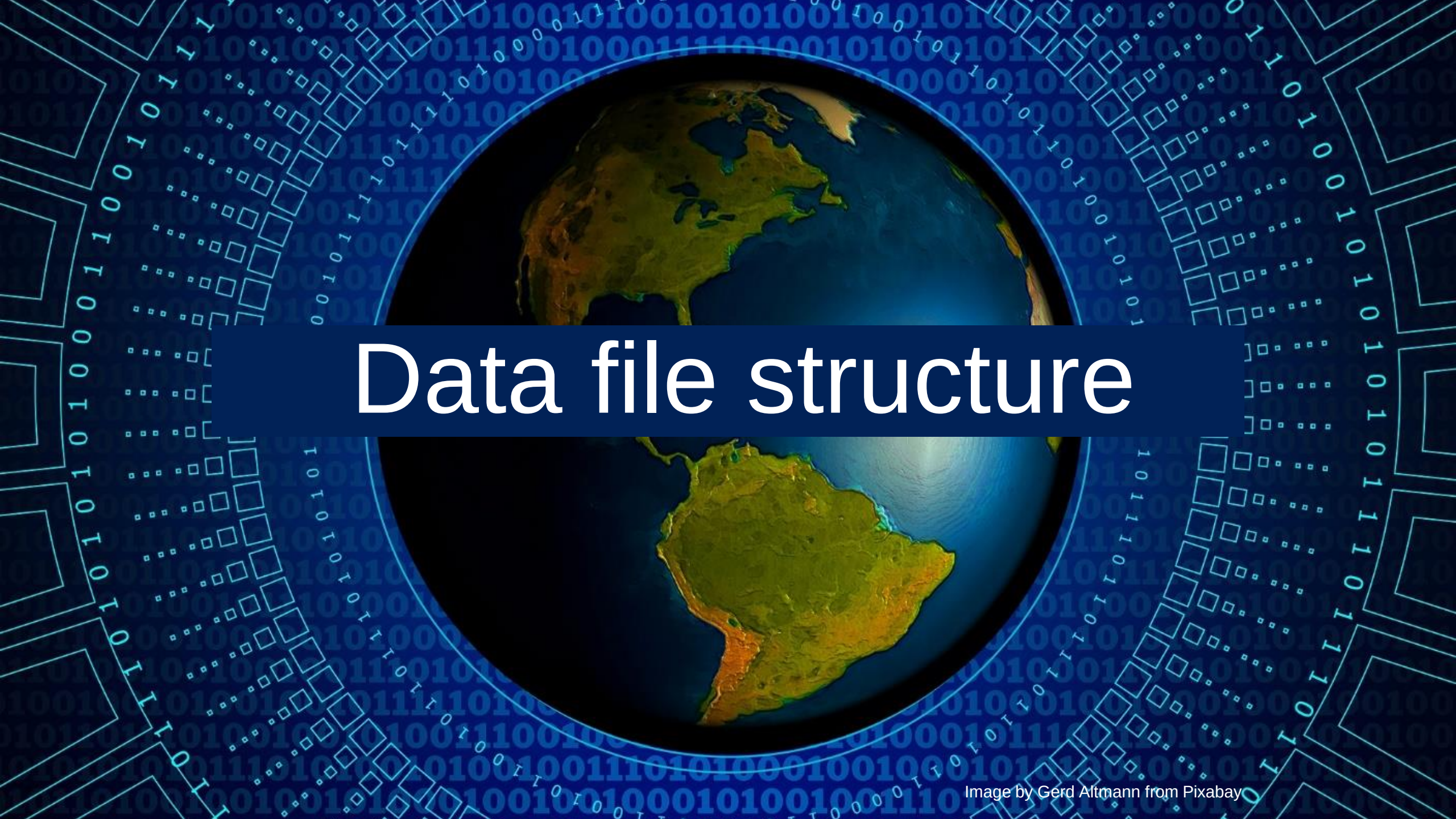


Image by Gerd Altmann from Pixabay

Design a data file structure

- In an early stage of your research, you are faced with the question of **what form your data files should take**. Your initial decision about the structure of your data files should be considered thoroughly.

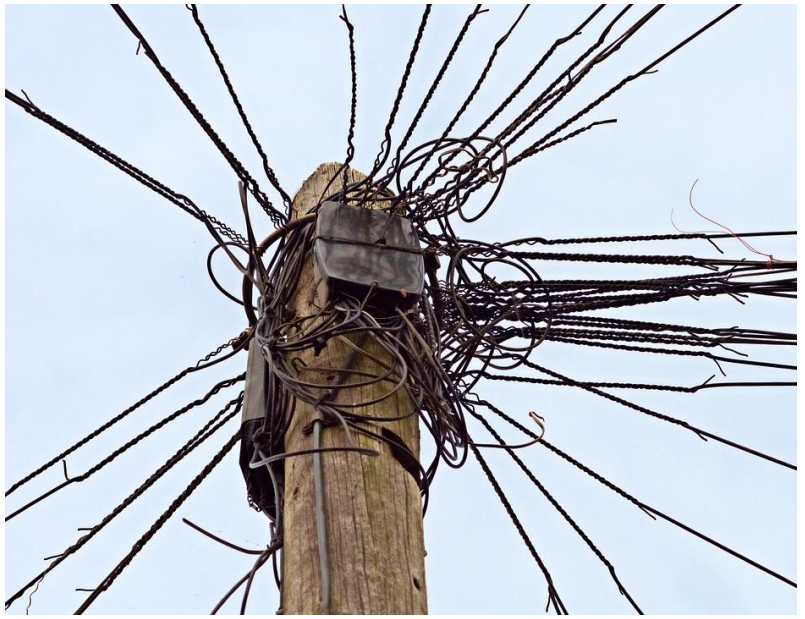
- The **data file structure** has a huge impact on the possible ways your files can be processed and analysed and once your structure has been filled with data, any changes to it are usually laborious and time-consuming.

Create a data organising system

- Follow your working pattern
- Systematic and logical
- Quick and easy to navigate
- Simple enough to be used all the time
- Considered and thought through before you start (!)
- Scaleable

Organise your data

- Do not use your computer desktop as a storage place
- Make a folder hierarchy and give your folders **descriptive** and **informative** names
- Avoid folders that become too broad or general, create more **subfolders** instead.
- Keep active and finished parts of your project separate, and take the time to tidy once in a while



Metadata

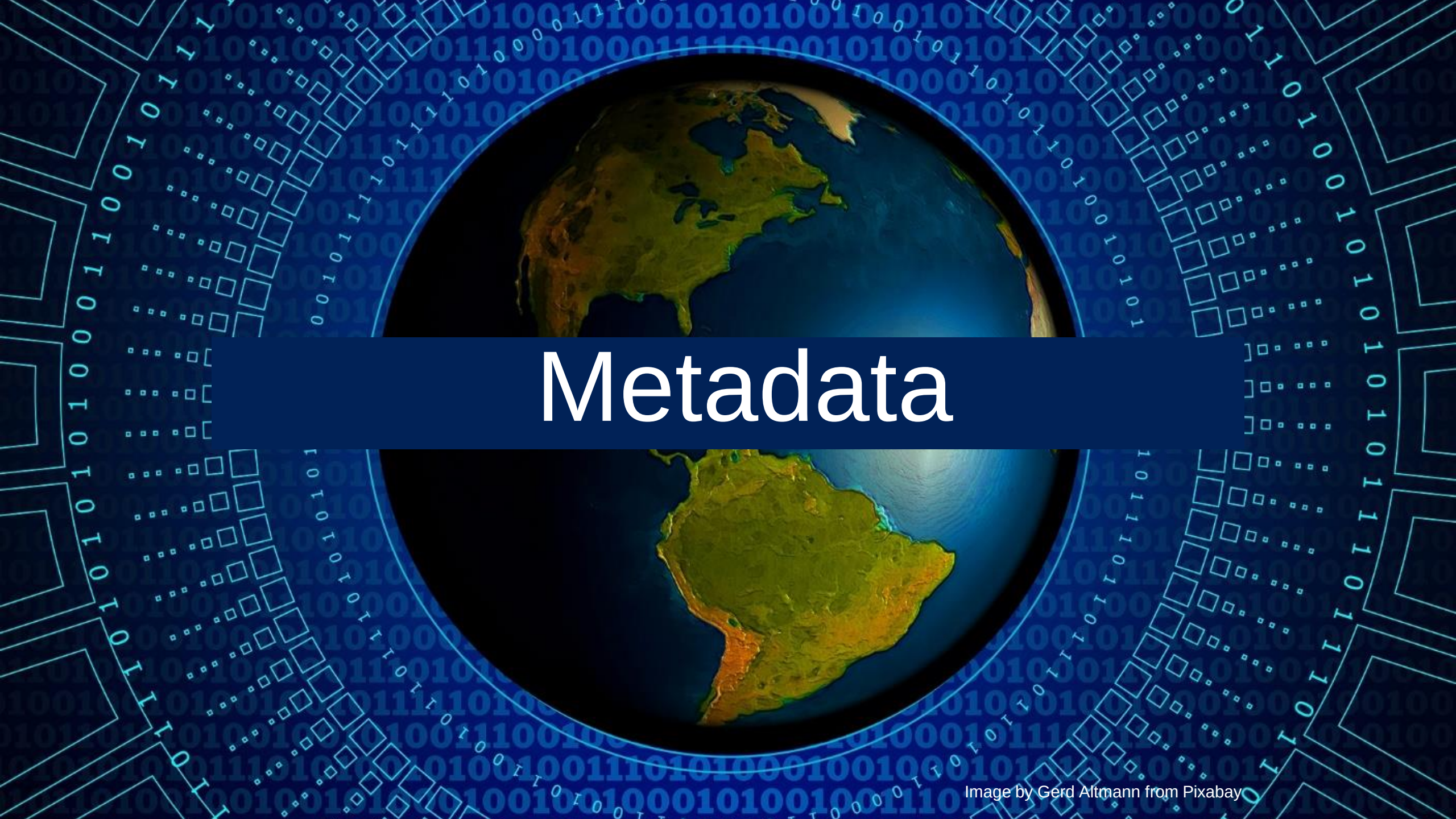
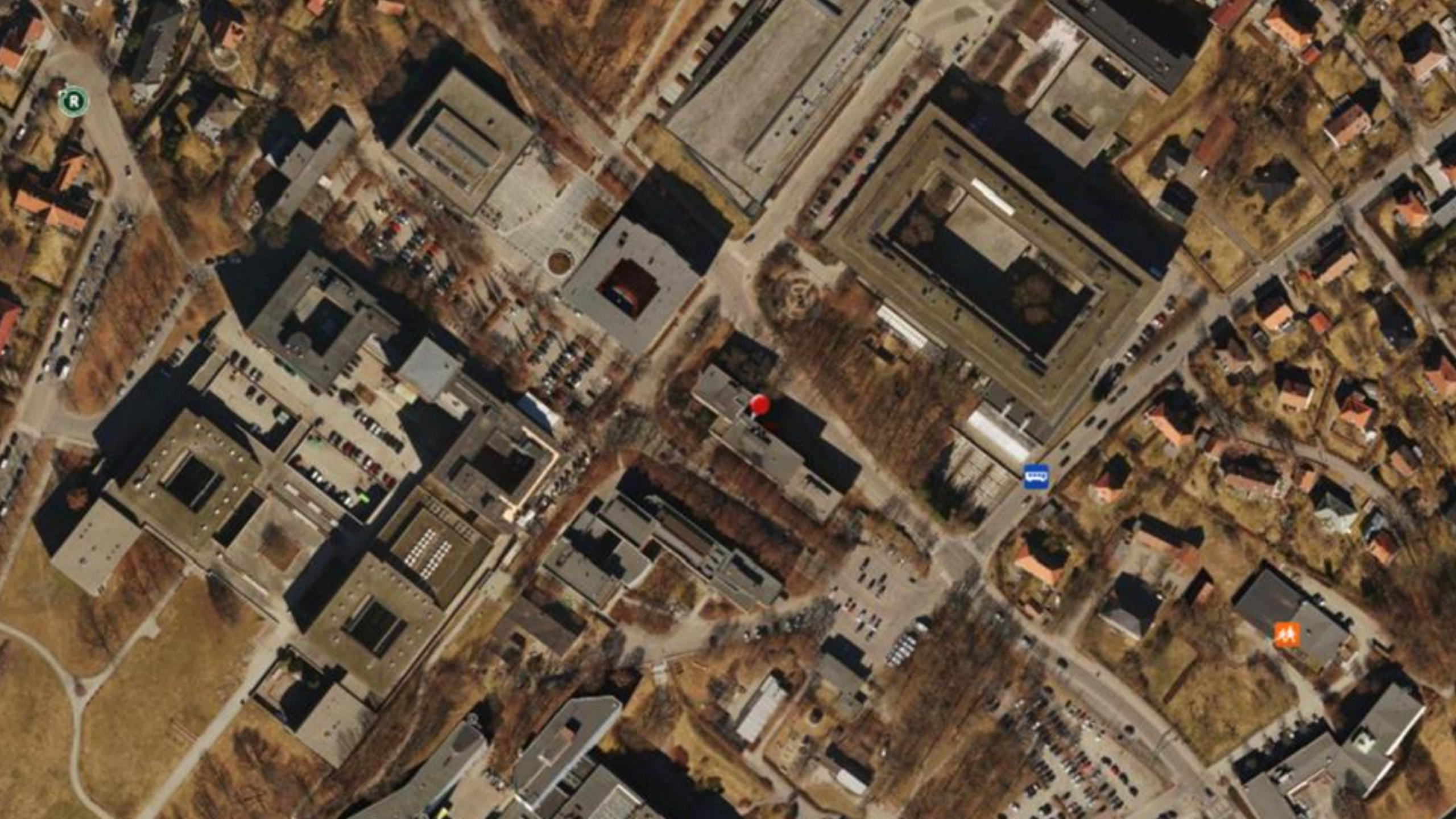


Image by Gerd Altmann from Pixabay

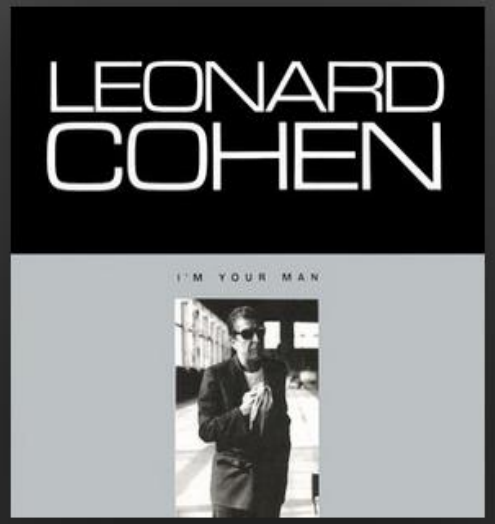


Metadata

File name: 2019-10-11 ScreenShot.png
Source: https://kart.finn.no
Year: 1937
Location: Oslo, Blindern
Location_GPS: 59.93760º N, 10.72122º Ø
59° 56′ 15.4″N, 10° 43′ 16.4″Ø
Altitude: 76m

Metadata

- Metadata or “data about data” are **descriptions that facilitate cataloguing data and data discovery**.
- Metadata are intended for machine-reading. When data is submitted to a trusted data repository, the archive generates machine-readable metadata.
- Machine-readable metadata help to explain the purpose, origin, time, location, creator(s), term of use, and access conditions of research data
- Your discipline very likely has standards for metadata! (see, for example: https://rd-alliance.github.io/metadata-directory/standards/)



ALBUM

I'm Your Man

Leonard Cohen • 1988 • 8 songs, 40 min 48 sec

#	TITLE	
1	First We Take Manhattan Leonard Cohen	6:00
2	Ain't No Cure for Love Leonard Cohen	4:49

Types of metadata

Descriptive metadata	For finding or understanding a resource
Administrative metadata - Technical metadata - Preservation metadata - Rights metadata	 - For decoding and rendering files - Long-term management of files - Intellectual property rights attached to content
Structural metadata	Relationships of parts of resources to one another
Markup languages	Integrates metadata and flags for other structural or semantic features within content

https://www.niso.org/publications/understanding-metadata-2017

Why do we need metadata?

- Discovery
- Aid in identification or understanding of a resource
- Interoperability
- Digital object management
- Preservation
- Navigation within parts of items

- **A huge help when making data FAIR!**

https://www.niso.org/publications/understanding-metadata-2017

Findable Accessible Interoperable Reusable

Dublin Core Metadata Element Set - a set of fifteen “core“ elements (properties) for describing resources

Dublin Core Metadata Element Set

01. **Contributor** – “An entity responsible for making contributions to the resource.”
02. **Coverage** – “The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.”
03. **Creator** – “An entity primarily responsible for making the resource.”
04. **Date** – “A point or period of time associated with an event in the lifecycle of the resource.”
05. **Description** – “An account of the resource.”
06. **Format** – “The file format, physical medium, or dimensions of the resource.”
07. **Identifier** – “An unambiguous reference to the resource within a given context.”
08. **Language** – “A language of the resource.”
09. **Publisher** – “An entity responsible for making the resource available.”
10. **Relation** – “A related resource.”
11. **Rights** – “Information about rights held in and over the resource.”
12. **Source** – “A related resource from which the described resource is derived.”
13. **Subject** – “The topic of the resource.”
14. **Title** – “A name given to the resource.”
15. **Type** – “The nature or genre of the resource.”

Source: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Darwin Core - extension of Dublin Core meant to provide a stable standard reference for sharing information on biological diversity

Darwin Core

Darwin Core: An Evolving Community-Developed Biodiversity Data Standard

John Wieczorek, David Bloom, ... David Vieglais

Record-level Terms	Dublin Core terms, institutions, collections, nature of data record	Simple Darwin Core (flat)
Occurrence	evidence of species in nature, observers, behavior, associated media, references.	
Event	sampling protocols and methods, date, time, field notes	
Location	geography, locality descriptions, spatial data	
Identification	linkage between Taxon and Occurrence	
Taxon	scientific names, vernacular names, names usages, taxon concepts, and the relationships between them	
GeologicalContext	geologic time, chrono-stratigraphy, biostratigraphy, lithostratigraphy	
ResourceRelationship	explicit relationships between identified resources (e.g., one organism to another, taxon to location, etc.)	Generic Darwin Core (relational)
MeasurementOrFact	measurements, facts, characteristics, assertions, references	

Reset zoom − + ‹ Previous Next › All

Figure 2. Darwin Core Categories: Simple Darwin Core is comprised of seven categories of terms (green).

This subset of Darwin Core terms represents descriptive data about organisms that can be represented in one file with one row... show more →

Show in Context

Download: PPT PNG

Source: https://doi.org/10.1371/journal.pone.0029715

Replication data for: "Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard"

Version 1.1

Koehl, Jean-Baptiste, 2021, "Replication data for: "Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard"", https://doi.org/10.18710/IIHGSH, DataverseNO, V1

Cite Dataset

Learn about Data Citation Standards.

Dataset Metrics

0 Downloads

Description High-resolution versions of the figures and supplements of the Koehl (2020) manuscript entitled "Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard", which could not be attached to the manuscript itself due to size limit but that are necessary to identify the main structures and interpretation. (2021-04-07)

Subject Earth and Environmental Sciences

Keyword Faults, Tectonics, High-resolution figures, Svalbard, Eurekan, Billefjorden Group, Strain partitioning, Bedding-parallel thrust, Décollement and duplexes

Related Publication Koehl, J.-B. P.: Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard, Solid Earth, 12, 2021.

Files Metadata Terms Versions

Citation Metadata

Dataset Persistent ID	doi:10.18710/IIHGSH
Publication Date	2021-04-07
Title	Replication data for: "Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard"
Author	Koehl, Jean-Baptiste (University of Oslo) - ORCID: 0000-0001-7189-1988
Contact	Use email button above to contact. Koehl, Jean-Baptiste P. (University of Oslo)
Description	High-resolution versions of the figures and supplements of the Koehl (2020) manuscript entitled "Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard", which could not be attached to the manuscript itself due to size limit but that are necessary to identify the main structures and interpretation. (2021-04-07)
Subject	Earth and Environmental Sciences
Keyword	Faults Tectonics High-resolution figures Svalbard Eurekan Billefjorden Group Strain partitioning Bedding-parallel thrust Décollement and duplexes
Related Publication	Koehl, J.-B. P.: Early Cenozoic Eurekan strain partitioning and decoupling in central Spitsbergen, Svalbard, Solid Earth, 12, 2021.

Export Metadata

Citation Metadata

Geospatial Metadata

Geographic Coverage	Norway, Svalbard
Geographic Unit	central Spitsbergen
Geographic Bounding Box	15.0 20.0 79.0 77.0

Dataverse –deposit your data

Enter metadata

Information about the various metadata fields can be obtained by placing the cursor on the field names (a roll-over window appears). Here is some more information about some of the fields:

- **Title:**
  - Enter a title for your dataset.
  - If your dataset is used in a publication, you may enter the title of the publication, and click on *Add "Replication Data for" to Title.*
- **Author**:
  - Enter your name as you use it in your publications. We recommend you to add your affiliation as well. For entering co-authors, click on the plus button. We also recommend you to add your ORCID (https://orcid.org).
- **Contact:**
  - Enter a contact email address. Also add the name of the contact person or research group/institution.
- **Description:**
  - Enter information about the data to be uploaded. Avoid using certain HTML tags and other special characters (e.g. [or]). If you need to add paragraphs, add the HTML tags and around each paragraph.
  - If relevant, enter information about the data collection/methodology here.
  - If applicable, also enter the publication abstract. The abstract should be entered into a second description field, which can be added by clicking the plus button to the right. NB! If your article is only submitted and not accepted (yet), DO NOT mention the name of the journal it has been submitted to.

https://site.uit.no/dataverseno/deposit//

Dataverse –deposit your data

- **Keyword:**
  - Information such as the subject area(s) (e.g. morphology or zoology) and the statistical method(s) may be entered into the keyword field.
  - Each keyword needs to be entered separately. Please click the plus button to enter more keywords.
  - Vocabulary and Vocabulary URL are not mandatory and may be left empty.
- **Related Publication:**
  - If the files you are depositing are the background data for a publication, you should include a reference to the publication here.
  - **Note!** If your manuscript has been submitted for review but has not yet been accepted, DO NOT list the name of the journal or publisher. Instead you may simply write "Submitted for review" or similar.
  - **Note!** If the review of your manuscript is going to be double blind (both author and reviewer are anonymous), you must add a note about it in the *Related Publication* field. This way, the curators can assist you in anonymizing the dataset.
  - (When adding more than one publication, only the first of them will be visible on the overview page of the dataset. If you don't want to highlight any of the publications in this way, you may add the following text in the first publication field: "Click Metadata tab

https://site.uit.no/dataverseno/deposit//

Design a data file structure

An important part of the **metadata** is often **embedded into the data file**, e.g. variable names and variable or value labels.

Therefore, the **structure** of your data also **contributes to the clarity** of your **data documentation**.

Documentation

Image by Gerd Altmann from Pixabay

Documentation vs. metadata?

"While data documentation is meant to be read and understood by humans, metadata (which are sometimes a part of the documentation) are primarily meant to be processed by machines."

https://howtofair.dk/how-to-fair/metadata/

Documentation: what, why and how?

- **What** documentation?
- **Why** documentation?
- **How** to create documentation?

What documentation?

- Systematically documented research data is key to making the data publishable, discoverable, citable, and reusable (and FAIR)

- Clear and detailed documentation **improve the overall data quality**

- It is vital to document **both the study** for which the data has been collected **and the data itself**. These two levels of documentation are called **project-level** and **data-level** documentation

https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata

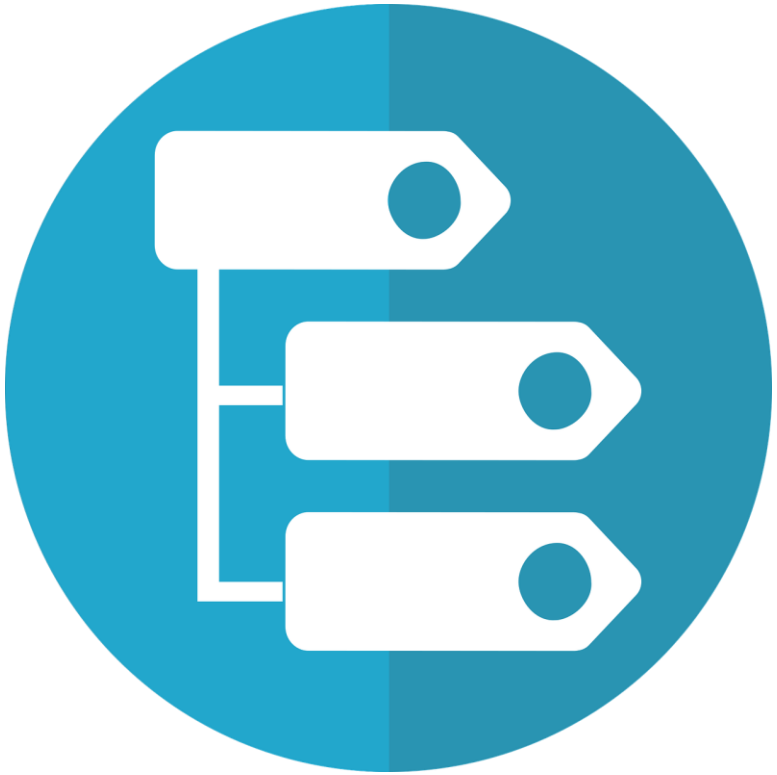
What documentation?

- The **project-level documentation** explains the aims of the study, what the research questions/hypotheses are, what methodologies were being used, what instruments and measures were being used, etc.

- **Data-level** or **object-level documentation** provides information at the level of individual objects such as images or variables in a database/table or transcripts, etc.

- It's become a convention to create multiple **README-files**, both for project-level documentation and for data-level documentation

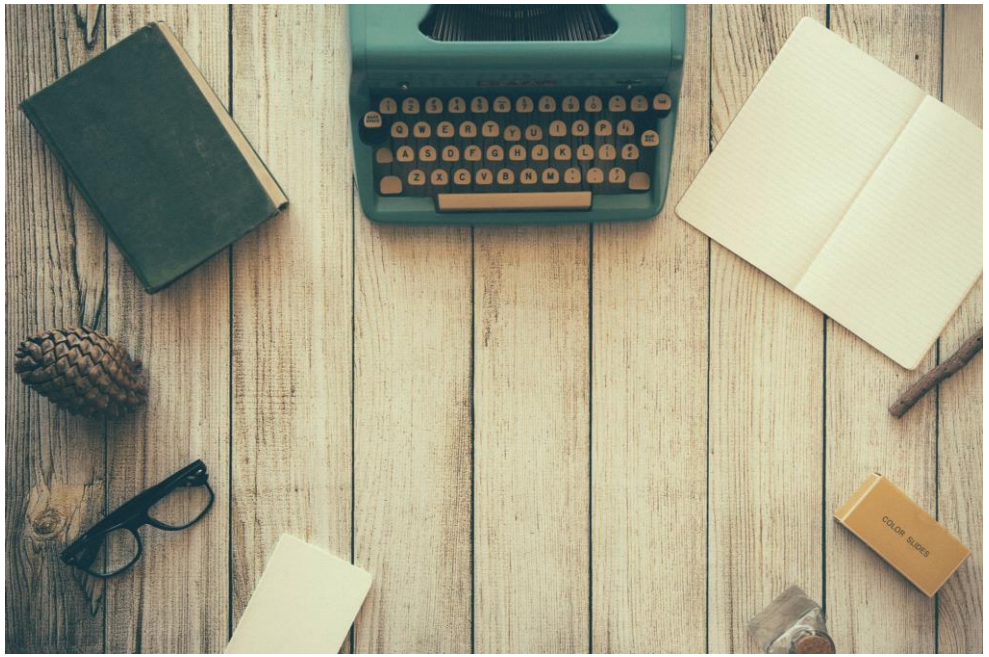
https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata



- **Project level**
  - General information about the project
  - Hypotheses
  - Methods for collecting and analyzing the data
- **File or folder level**
  - Folder system
  - Version control
  - Lists of files and how they were obtained
  - List of samples?
  - Explanation of scripts and sets of programs written
  - (Naming convention for files)
- **Variable or experiment level**
  - Description of each variable, with units

Document everything your data has been through

- Field journal
- Lab journals and experimental protocols
- Scripts for analysis
- Questionnaires, codebooks, data dictionaries
- Software syntax and output files
- Methodology reports
- Geolocation, orientation (e.g. when collecting a sample)
- Instrument settings and calibration



https://mantra.edina.ac.uk/documentation_metadata_citation/

Why documentation?

- Helps others understand your project and reuse your data

"It has been retracted because the results were not reproducible, and the authors found data missing from a lab notebook."

https://www.bbc.com/news/world-us-canada-50989423

Why documentation?

- Helps others understand your project and reuse your data
    - all you need to know about your data
    - all your collaborators need to know about your data
    - all you need to know about your collaborators data
    - all anyone else who wants to work with your data needs to know about it
    - more information is better than less
    - structured information is better than unstructured information
    - no documentation is the worst!
    - start early, save time & have less stress
    - re-use templates for good documentation, READMEs, etc.
    - you must have documentation when you archive your data

Rockenberger 2020, Shut Up And Write Documentation, README.txt https://zenodo.org/record/3778273

How to create documentation?

- README.txt-files:
  - Announce that they are the first file to open when looking through your old data
  - Provide a map for exploring your files
  - Create one README.txt file per folder in as many folders as you can
  - They do not need to be large, but their content should help navigation through digital files and folders
  - a project-level README.txt should give the general project information and a very coarse overview of file contents and locations
  - A data-(object-) level README.txt would be more specific as to what each file contains

Briney 2014, Wrapping Up A Project, http://dataabinitio.com/?p=344

How to create documentation?

README.txt-files: Cornell University template and guide:
https://data.research.cornell.edu/content/readme

README.txt-files: some examples
https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/LNCK8O

https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/Q3FZAN

https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/TYPJXU

Break

Image by Gerd Altmann from Pixabay

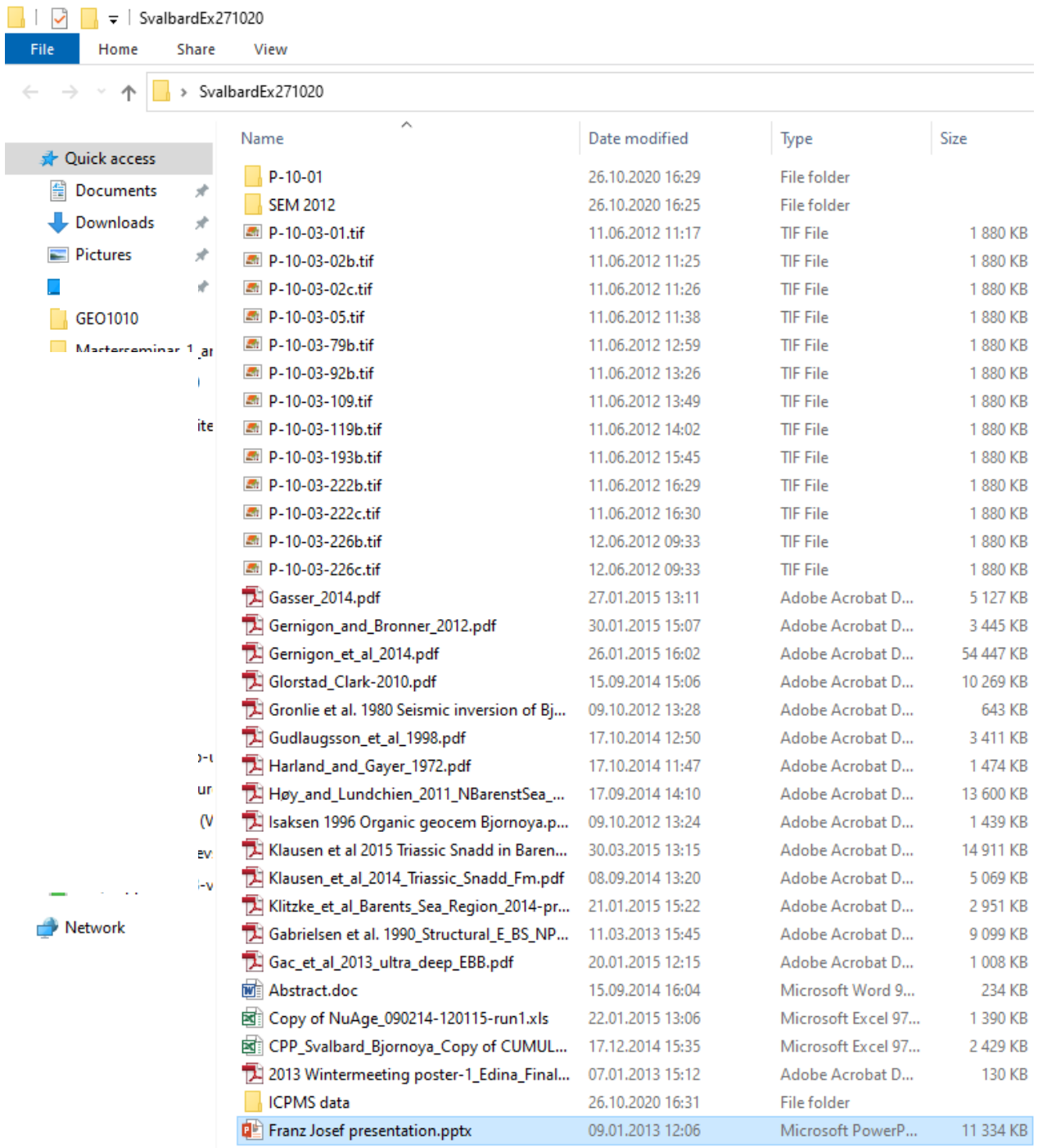
File naming & Folder structure

Image by Gerd Altmann from Pixabay

File naming & Folder structure

- Structuring your data files in folders is important for making it easier to **locate** and **organise** files and versions. A proper folder structure is especially needed when collaborating with others.

- The decision on how to organise your data files depends on the **plan and organisation of the study**. All material relevant to the data should be entered into the data folders, including detailed information on the data collection and data processing procedures.



- Example of a research project **without** a data file structure – Edina Pózer ©

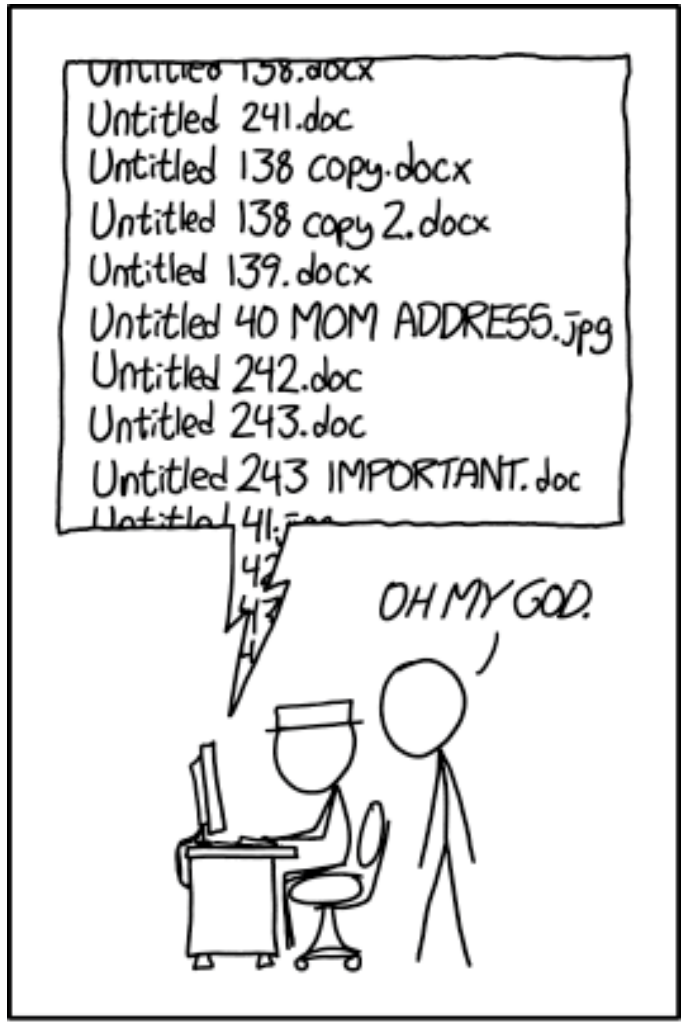
How it could look like:

```
project_name/
├── README.md             # overview of the project
├── data/                 # data files used in the project
│   ├── README.md         # describes where data came from
│   └── sub-folder/       # may contain subdirectories
├── processed_data/       # intermediate files from the analysis
├── manuscript/           # manuscript describing the results
├── results/              # results of the analysis (data, tables, figures)
├── src/                  # contains all code in the project
│   ├── LICENSE           # license for your code
│   ├── requirements.txt  # software requirements and dependencies
│   └── ...
└── doc/                  # documentation for your project
    ├── index.rst
    └── ...
```

Research project with a proper data file structure. Image taken from CodeRefinery, Lesson on Reproducible Research. https://coderefinery.github.io/reproducible-research/02-organizing-projects/. Shared under CC-BY 4.0.

Naming conventions

- Short names (but long enough that they still make sense)
- The most general information first, then add details to the name
- Underscore to separate words, DO NOT use space in file names!
- Dates backwards (YYYYMMDD)
- Numbers (e.g. version number) should have the same number of digits, use e.g. 01, not just 1.
- Version number at the end



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

https://www.explainxkcd.com/wiki/index.php/1459:_Documents

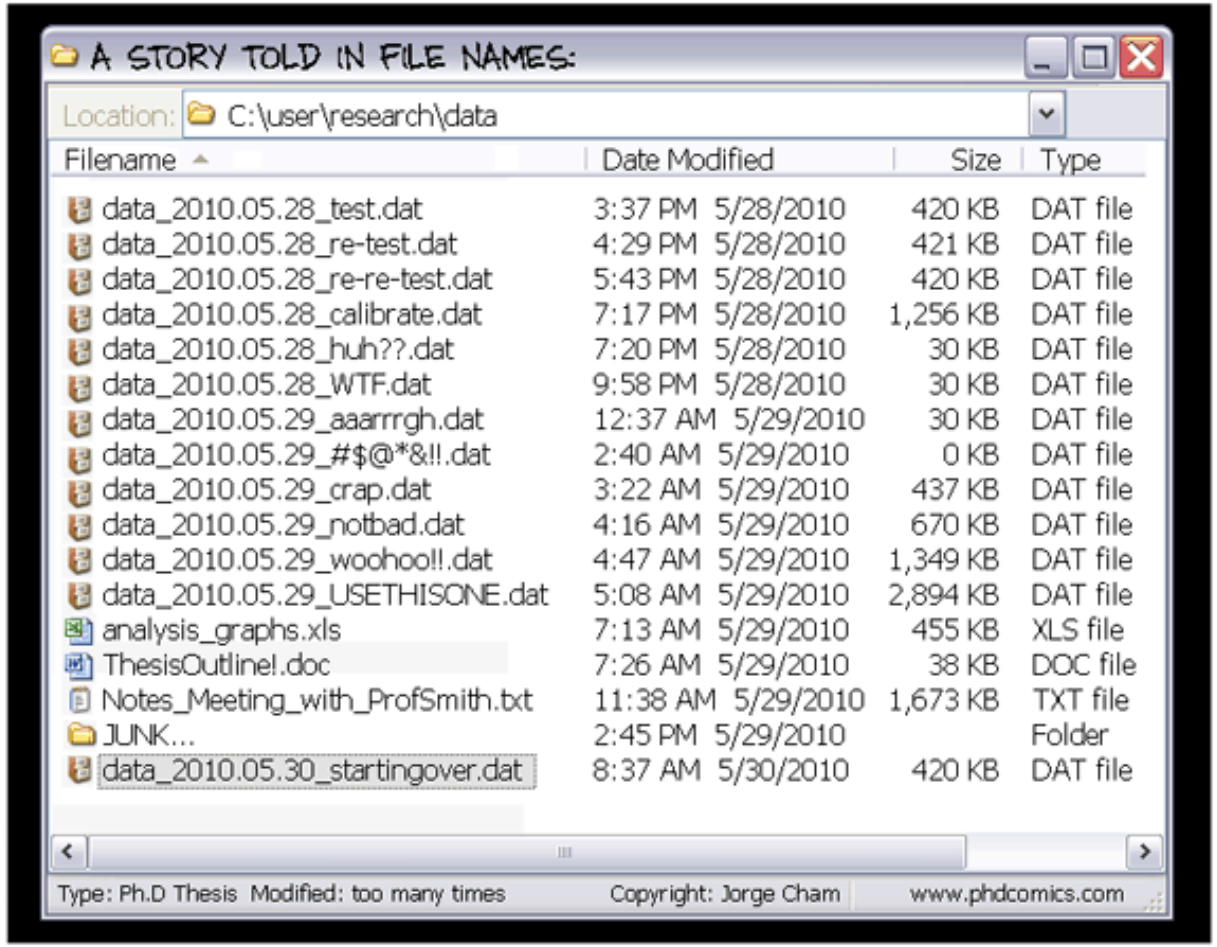
Avoid using the following characters in Folder and File names:

# pound	< left angle bracket	{ left curly bracket
% percent	> right angle bracket	} right curly bracket
& ampersand	/ forward slash	* asterisk
\ back slash	blank spaces	? question mark
$ dollar sign	' single quotes	= equal sign
! exclamation point	" double quotes	

Also, keep these rules in mind:

- Don’t start or end your filename with a space, period, hyphen, or underline
- Keep your filenames to a reasonable length
- Most operating systems are case sensitive; always use lowercase

Tidy every once in a while



http://phdcomics.com/comics.php?f=1323

Sustainable file formats

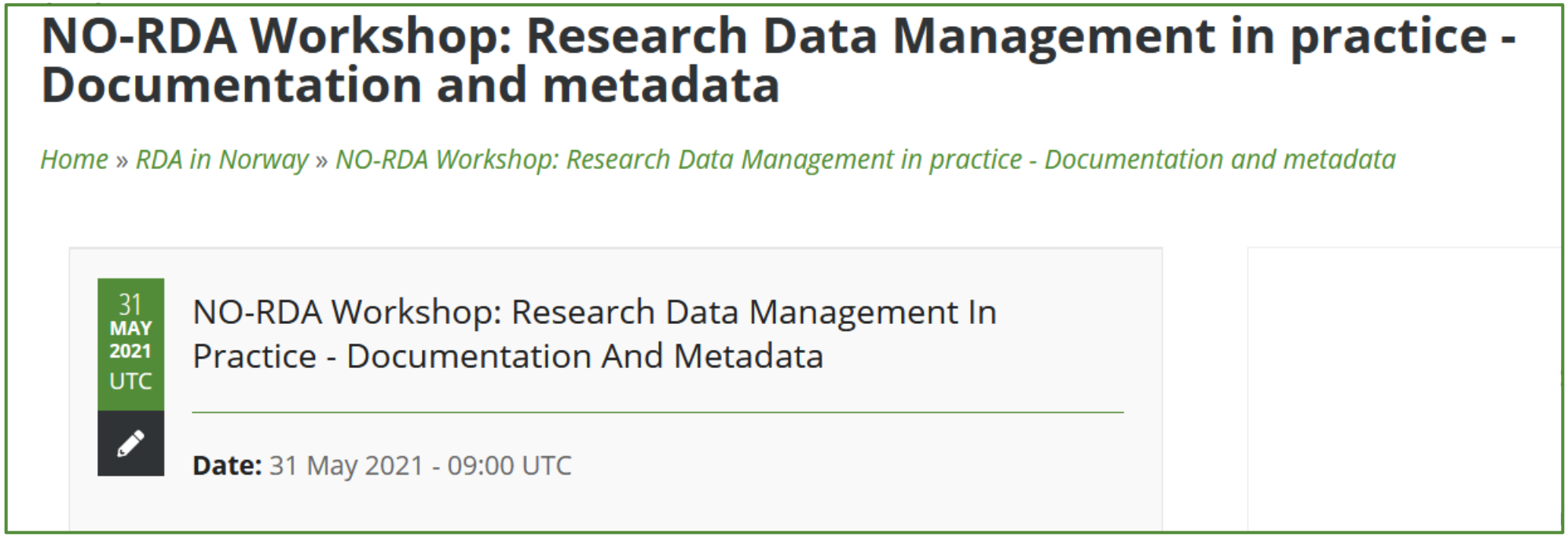
- When your project is finished and you plan on archiving your data, it is widely recommended to transfer data and accompanying files into a **more sustainable format.**

- Trusted data archives often have guidelines for this, e.g. the recommendations of Data Archiving and Networked Services (DANS)

https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats

Image by Gerd Altmann from Pixabay

Skills development: RDA Norway



https://www.rd-alliance.org/group/rda-norway/event/no-rda-workshop-research-data-management-practice-documentation-and-metadata

Skills development at UiO

Carpentry@UiO

Carpentries

www.uio.no/carpentry

CodeRefinery

https://coderefinery.org/

UiOs Digital Scholarship Center

https://www.ub.uio.no/english/writing-publishing/dsc/

Questions & Answers

Image by Gerd Altmann from Pixabay

Upcoming sessions…

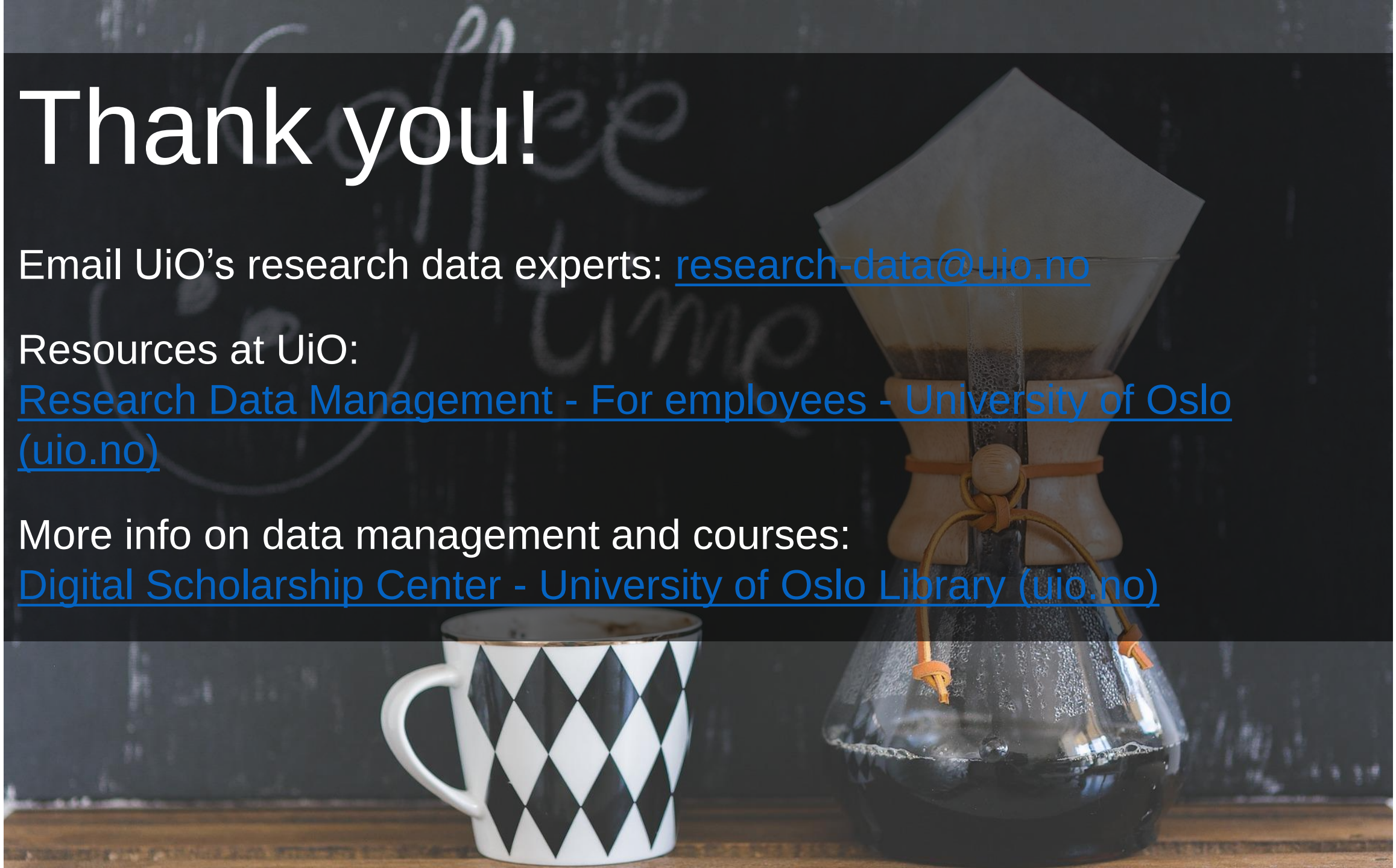
- Data Classification and Storage Selection
  12/05 9:00-10:30
- Ethics, Privacy and Data Protection
  19/05 9:00-10:30
- Sharing and Archiving Research Data
  20/05 9:00-10:30
- Data Discovery
  21/05 9:00-10:30

Thank you!

Email UiO's research data experts: research-data@uio.no

Resources at UiO:
Research Data Management - For employees - University of Oslo (uio.no)

More info on data management and courses:
Digital Scholarship Center - University of Oslo Library (uio.no)



Sources

- Deutz DB, Buss MCH, Hansen JS, Hansen KK, Kjelmann KG, Larsen AV, Vlachos E, Holmstrand KF (2020). *How to FAIR: a Danish website to guide researchers on making research data more FAIR*. https://howtofair.dk/how-to-fair/
- CESSDA Training Team (2017 - 2020). *CESSDA Data Management Expert Guide.* Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide
- *Research Data MANTRA* [online course, https://mantra.edina.ac.uk/] by the Research Data Service, University of Edinburgh
- Rockenberger A (2020) *Shut Up And Write Documentation: README.txt*. Retrieved from: https://zenodo.org/record/3778273
- K. Briney (2014) *Wrapping Up A Project*. Retrieved from: http://dataabinitio.com/?p=344
- *Guide to writing "readme" style metadata* by the Research Data management Service Group, Cornell University. Retrieved from: https://data.research.cornell.edu/content/readme]
- Riley J (2017). *Understanding Metadata: What is Metadata, and What is it For?: A Primer*. National Information Standards Organization (NISO). https://www.niso.org/publications/understanding-metadata-2017
- DCMI Metadata Terms (2020). https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715
- Data Archiving and Networked Services (DANS). *File formats*. Retrieved from: https://dans.knaw.nl/en/about/services/easy/information-about-depositing-data/before-depositing/file-formats