

Module III:

Data organization, metadata, and documentation

Edina Pózer and Ivana Malovic
The University of Oslo Library



Today's course

- One hour lecture with a short break
- Q&A session

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

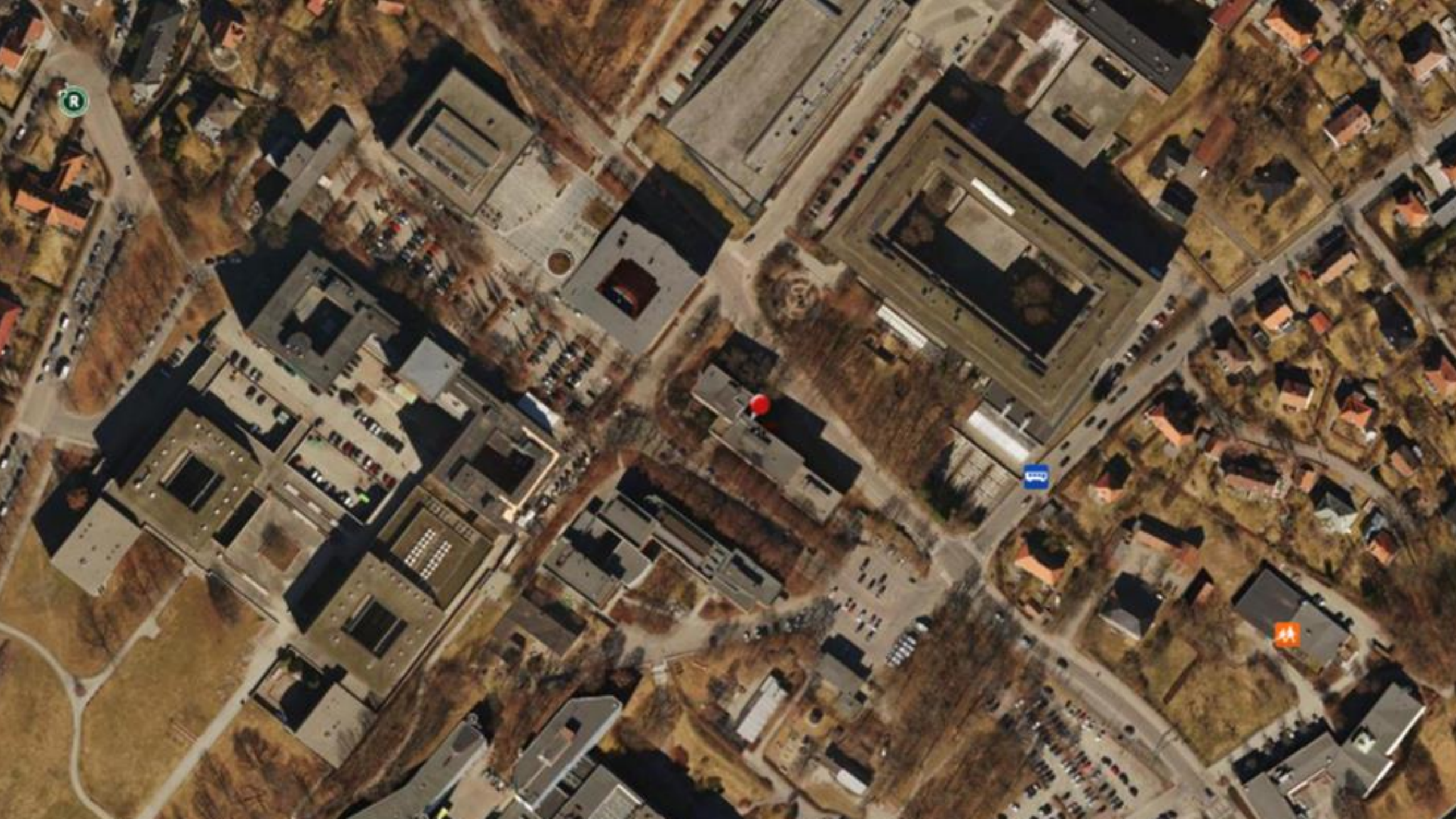
Agenda



- Metadata and why we need them
- Documentation and how to create it
- Data file structure
- File naming conventions
- Version control
- Sustainable file formats

Metadata





Living





Metadata

File name: 2019-10-11_ScreenShot.png

Source: <https://kart.finn.no>

Year: 1937

Location: Oslo, Blindern

Location_GPS: 59.93760° N, 10.72122° Ø

59° 56' 15.4"N, 10° 43' 16.4"Ø

Altitude: 76m

Metadata

- **“Data about data”**
- Descriptions that facilitate **cataloguing data and data discovery**
- Intended **for machine-reading**
- **Help to explain** the purpose, origin, time, location, creator(s), term of use, and access conditions of research data
- Your discipline might have **standards for metadata**

LEONARD
COHEN

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

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!**

F
indable



A
ccessible



I
nteroperable



R
eusable



Dublin Core Metadata Element Set

is a set of fifteen “core” elements (properties) for describing resources

Dublin Core Metadata Element Set

- | | |
|-----------------|---------------|
| 01. Contributor | 09. Publisher |
| 02. Coverage | 10. Relation |
| 03. Creator | 11. Rights |
| 04. Date | 12. Source |
| 05. Description | 13. Subject |
| 06. Format | 14. Title |
| 07. Identifier | 15. Type |
| 08. Language | |

Darwin Core

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

Darwin Core: An Evolving Community-Developed Biodiversity Data Standard

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	

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 per record and one column per term. Two additional categories (orange) expand Darwin Core with concepts that require a more complex data structure, such as multiple measurements from a single specimen, and cannot be represented easily in Simple Darwin Core.
doi:10.1371/journal.pone.0029715.g002



**UNIVERSITY
OF OSLO**

[DataVerseNO](#) > [University of Oslo](#) >

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

Version 1.4



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.](#)

[Access Dataset ▾](#)

[Contact Owner](#)

[Share](#)

Dataset Metrics ?

6 Views ?

0 Downloads ?

0 Citations ?

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, 1025–1049, <https://doi.org/10.5194/se-12-1025-2021>, 2021. doi: [10.5194/se-12-1025-2021](https://doi.org/10.5194/se-12-1025-2021)

[Files](#)

[Metadata](#)

[Terms](#)

[Versions](#)

<https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/IIHGSH>

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 (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, 1025–1049, <https://doi.org/10.5194/se-12-1025-2021>, 2021. doi: 10.5194/se-12-1025-2021 <https://doi.org/10.5194/se-12-1025-2021>

Language ?

English

Producer ?

University of Oslo (UiO) <https://www.uio.no/english>
Contributor ?

Data Collector : Jean-Baptiste P. Koehl
Research Group : Centre for Earth Evolution and Dynamics

[Files](#)[Metadata](#)[Terms](#)[Versions](#) 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 <p> and </p> 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.

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

Data Documentation Initiative (DDI) standard

- An international standard for describing the data produced by observational methods in the social, behavioural, economic, and health sciences
- Expressed in XML
- Supports the entire research data lifecycle

Data Documentation Initiative (DDI) standard

- For a single data set: [DDI Codebook](#) (DDI-C)
- For a longitudinal, linked and other complex datasets: [DDI Lifecycle](#) (DDI-L)
- For interdisciplinary research soon is coming [DDI - Cross Domain Integration](#) (DDI-CDI)

DDI standard's common fields

Title	Subject Terms	Sampling
Alternate Title	Geographic Coverage	Weights
Principal Investigator	Time Period	Mode of Collection
Funding	Date of Collection	Response Rates
Bibliographic Citation	Unit of Observation	Extent of Processing
Series Information	Universe	Restrictions
Summary	Data Type	Version History

DDI standard: Example

Home / Data catalogue / FSD2926

Display study description in English

FSD2926 Intercultural Urban Public Space in Toronto 2011-2013

[Overview](#)

[Detailed description](#)

[Questions](#)

[Publications](#)

[Download data](#)

Authors

- Galanakis, Michail (University of Helsinki)

Keywords

citizen participation, communities, cultural interaction, cultural pluralism, decision-making, ethnic groups, immigration, politicians, public spaces, services for young people, social inequality, urban development, urban environment, urban sociology, urban spaces

Abstract

The dataset contains transcripts of interviews conducted mainly in Toronto, Canada, during 2011 and 2012. A few interviews were conducted in Vancouver and Guelph as well. The main themes of the interviews were multiculturalism, interculturalism, diversity and public space, and how the participants' perceptions of interculturalism and public space. The interviewees were professionally or voluntarily involved in the physical or social planning process, in providing services for youth, or in dealing with managing diversity (in policy-making, planning, arts etc). They were community activists, professional designers, managers of public spaces, social services providers, or young persons who represented the users of services aimed at communities.

The interviews were reflective, and questions asked changed according to what the interviewees talked about. The three main research questions were what the participants considered public spaces to be, how they defined interculturalism and, for expert interviews, how they planned/designed for diversity. Toronto is a very multicultural city, and one of the main aims of the study was to learn how Toronto's public space is managed and how public space could be used more creatively for the benefit of diverse groups. Other topics that came up included exclusion of youth, crime, services and facilities for youth, social and educational inequality, unemployment, public transport, street art, safety, police harassment, and privatization of public space.

In addition to 25 one- and two-person interviews, there was one focus group interview of 13 young men and women. Interviewee age ranged from adolescents to senior citizens. Background variables included the interviewee's occupation, gender and age.

The dataset is (B) available **for research, teaching and study.**

[Download the data](#)

[Study description in other languages](#)

- [in Finnish](#)

[Related files](#)

- [Data example \(PDF file, in English\)](#)
- [Dataset guide \(PDF file, in English\)](#)

Permanent link to this dataset:

<http://urn.fi/urn:nbn:fi:fsd:T-FSD2926>

Direct link to this tab:

https://services.fsd.tuni.fi/catalogue/FSD2926?tab=summary&lang=en&study_language=en

[Study description in machine readable DDI 2.0 format](#)



Metadata record is licensed under a [Creative Commons Attribution 4.0 International](#) license.

<https://services.fsd.tuni.fi/catalogue/FSD2926>

DDI standard: Example

The machine-readable XML file:

<https://services.fsd.tuni.fi/catalogue/FSD2926/DDI/FSD2926e.xml>

```
<codeBook xml:lang="en">
  <docDscr>
    <citation>
      +<titlStm></titlStm>
      <prodStm>
        <producer abbr="FSD">Finnish Social Science Data Archive</producer>
        <copyright>Finnish Social Science Data Archive</copyright>
        <prodPlac>Finnish Social Science Data Archive</prodPlac>
      </prodStm>
    </biblClt>
    Intercultural Urban Public Space in Toronto 2011-2013 [data guide]. Finnish Social Science Data Archive [producer and distributor], 2018.
    </biblClt>
    <holdings URI="http://urn.fi/urn:nbn:fi:fsd:T-FSD2926" location="Finnish Social Science Data Archive FSD"/>
  </citation>
  <notes xml:lang="fi">
    FSD:n aineistokuvailut (FSD metadata records), jonka tekijä on Suomen yhteiskuntatieteellinen tietoarkisto (Finnish Social Science Data Archive), on lisensoitu Creative Commons Nimeä 4.0 Kansainvälinen (CC BY 4.0) -lisenssillä.
  </notes>
  <notes xml:lang="en">
    FSD:n aineistokuvailut (FSD metadata records) by Suomen yhteiskuntatieteellinen tietoarkisto (Finnish Social Science Data Archive) is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.
  </notes>
  <notes xml:lang="sv">
    FSD:n aineistokuvailut (FSD metadata records) av Suomen yhteiskuntatieteellinen tietoarkisto (Finlands samhällsvetenskapliga dataarkiv) är licensierad under en Creative Commons Erkännande 4.0 Internationell (CC BY 4.0) licens.
  </notes>
  </docDscr>
  <studyDscr>
    <citation>
      <titlStm>
        <titl>
          Intercultural Urban Public Space in Toronto 2011-2013
        </titl>
        <parTitl xml:lang="fi">
          Kulttuurienvälisyys ja julkinen tila Torontossa 2011-2013
        </parTitl>
        <IDNo agency="FSD">2926</IDNo>
        <IDNo agency="URN">urn:nbn:fi:fsd:T-FSD2926</IDNo>
      </titlStm>
      <rspStm>
        <AuthEnty affiliation="University of Helsinki">Galanakis, Michail</AuthEnty>
      </rspStm>
      <prodStm>
        <copyright>
          In accordance with the agreement between FSD and the depositor.
        </copyright>
      </prodStm>
      <distStm>
        <distrbtr URI="https://www.fsd.tuni.fi/" abbr="FSD">Finnish Social Science Data Archive</distrbtr>
      </distStm>
      <serStm URI="https://services.fsd.tuni.fi/catalogue/series/8?lang=en">
        <serName abbr="yks">Individual datasets</serName>
        <serInfo>
          <p>
            Individual datasets that do not belong to any series.
          </p>
        </serInfo>
      </serStm>
      <verStm>
        <version date="2014-02-13">1.0</version>
      </verStm>
      </biblClt>
      Galanakis, Michail (University of Helsinki): Intercultural Urban Public Space in Toronto 2011-2013 [dataset]. Version 1.0 (2014-02-13). Finnish Social Science Data Archive [distributor]. http://urn.fi/urn:nbn:fi:fsd:T-FSD2926
    </biblClt>
    <holdings URI="http://urn.fi/urn:nbn:fi:fsd:T-FSD2926" location="Finnish Social Science Data Archive FSD"/>
  </citation>
  <studyInfo>
    <subject>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">cultural pluralism</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">cultural interaction</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">public spaces</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">urban spaces</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">urban sociology</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">urban environment</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">urban development</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">social inequality</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">ethnic groups</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">decision-making</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">immigration</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">communities</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">citizen participation</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">politicians</keyword>
      <keyword vocab="ELSST" vocabURI="https://elsst.CESSDA.eu/id">services for young people</keyword>
      <topeClas vocab="OKM" vocabURI="http://www.yso.fi/onto/okm-tieteenala/conceptscheme">Social sciences</topeClas>
      <topeClas vocab="CESSDA Topic Classification" vocabURI="https://vocabularies.CESSDA.eu/urn/um:ddi:int.CESSDA.cv:TopicClassification:4.2">Housing</topeClas>
      <topeClas vocab="CFSSDA Topic Classification" vocabURI="https://vocabularies.CESSDA.eu/urn/um:ddi:int.CESSDA.cv:TopicClassification:4.2">1 and use and planning</topeClas>
    </subject>
  </studyInfo>
</codeBook>
```

What about non-digital data?

Osloslørsopp *Cortinarius osloensis* Brandrud, T.S. Jeppesen & Frøslev

KLASSIFISERING

Soppriket Fungi Rike
Stilksporesopper Basidiomycota Rekke
Hymeniesopper Agaricomycotina Underrekke
Ekte hymeniesopper Agaricomycetes Klasse
Agaricomycetidae Underklasse
Skivesoppordenen Agaricales Orden
Cortinariaceae Familie
Slørsopper *Cortinarius* Slekt
Osloslørsopp *Cortinarius osloensis*
Brandrud, T.S. Jeppesen & Frøslev Art

Vitenskapelig navn
Cortinarius osloensis

Bokmål
osloslørsopp ★

Nynorsk
osloslørsopp ★

[Gå til navnetreet](#)

BESKRIVELSE



Ungarn samt i Nord-Italia.

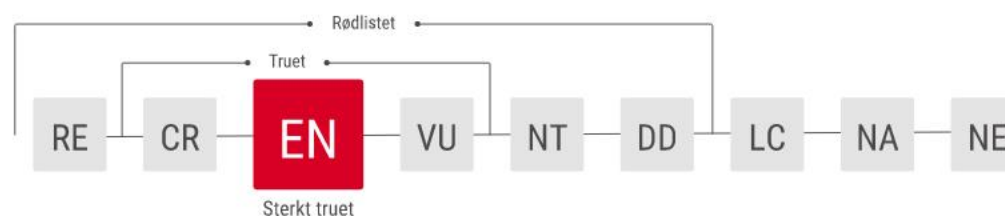
Osloslørsopp *Cortinarius osloensis* Brandrud, T.S. Jeppesen & Frøslev

Dette er en eksklusiv «aristokrat» med tilhold på beste vestkant i Oslo- og Asker-området. Lenge var den betraktet som en norsk endemisme; det vil si at arten kun er kjent fra Norge. Nylig er den imidlertid også funnet på én lokalitet i lindeskog i

KATEGORI

Fastlands-Norge med havområder

Gjeldende kategori er *sterkt truet* EN



[Se vurdering i Norsk rødliste for arter 2021](#)

REGISTRERTE OBSERVASJONER

Prikkene markerer registrerte funn av arten/artsgruppen.



Funndata i Artskart

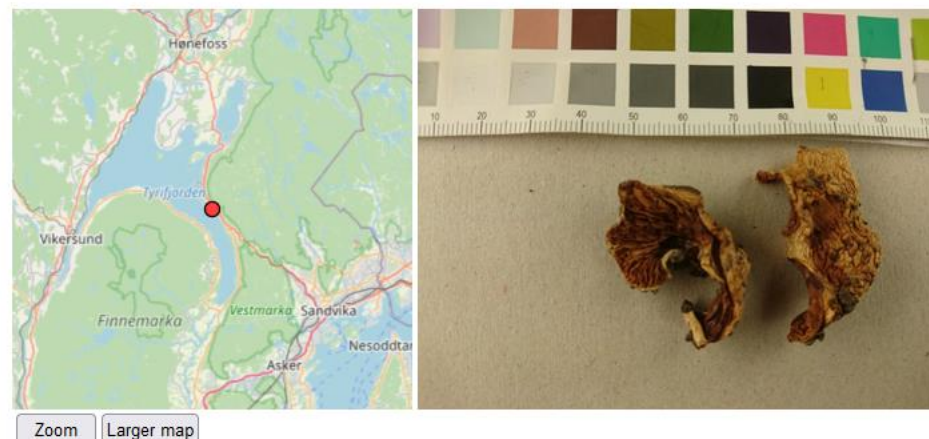
What about non-digital data?

Lukk side

The fungi collection

O-F-248784

Scientific name: *Cortinarius osloensis*
Uncertainty in id: IdentificationQualifierTaxonRank: Species
Determined by: Tor Erik Brandrud
Date of (last) determination: 2011-09-08
Collection date: 2011-09-08
Collector: Tor Erik Brandrud
Collecting number: TEB 737-11
Locality: Norway, Viken, Hole, Hole: Nesvika S
Coordinates: 59.996 N, 10.2854 E
Habitat: kalklindeskog, kalkhyller langs skar
Taxonomy: Agaricomycetes Agaricales Cortinariaceae



Object in [The fungi collection](#):

O-F-248784

Sample type: Preserved specimen

Item(s) in [The DNA bank](#):

[O-DFL-12453/2-D](#)

Sample type: DNA
Subtype: gDNA
Date of extraction:
Preservation: Dry
Method of extraction:
Prepared by: CCDB
DNA concentration
Genbank Acc.No:
BOLD ProcessID: [NOBAS7496-19](#)
Sequence length 667[0n]

What about non-digital data?

Lukk side

The DNA bank

O-DFL-12453

Scientific name: *Cortinarius osloensis* Brandrud, T.S.Jeppesen & Frøslev
Uncertainty in id:
Determined by: Tor Erik Brandrud
Date of (last) determination: 2011-09-08
Collection date: 2011-09-08
Collector: Tor Erik Brandrud
Locality: Norway, Buskerud, Hole, Nesvika S
Coordinates: 59.585627 N, 10.269668 E
Taxonomy: Agaricales Cortinariaceae



Photo not available

Zoom Larger map

Object in [The fungi collection:](#)

[O-F-248784](#)

Sample type: Preserved specimen

Item(s) in [The DNA bank:](#)

[O-DFL-12453/2-D](#)

Sample type: DNA
Subtype: gDNA
Date of extraction:
Preservation: Dry
Method of extraction:
Prepared by: CCDB
DNA concentration
Genbank Acc.No:
BOLD ProcessID: [NOBAS7496-19](#)
Sequence length
Validated: Unknown

What about non-digital data?



Metadata standards overviews

- The Digital Curation Center:

<https://www.dcc.ac.uk/guidance/standards/metadata>

- Research Data Alliance Metadata Directory:

<https://rd-alliance.github.io/metadata-directory/standards/>

Metadata is still work-in-progress!

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

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**."

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**

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

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



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

How to create documentation?

- README.txt-files:
 - **The first file to open**
 - **Provide a map** for exploring your files
 - Create **one** README.txt file **per folder**
 - 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

How to create documentation?

- Cornell University template and guide README.txt-files:

<https://data.research.cornell.edu/content/readme>

<https://cornell.app.box.com/v/ReadmeTemplate>

- README.txt-files: some examples

<https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/BZNYPF>

- Earth and Environmental Sciences / Engineering

<https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/QWIDIT>

- Agricultural Sciences; Medicine, Health and Life Sciences

<https://dataverse.no/dataset.xhtml?persistentId=doi:10.18710/IQGCHB> - Social Sciences

Questions?

Data file structure & File naming

Designing 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.

Designing a data file structure (cont.)

- 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.

Create a data organising system that...

- follows your working pattern
- is systematic and logical
- is quick and easy to navigate
- is simple enough to be used all the time
- is considered and thought through before you start (!)
- is scaleable

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
    └── ...
```

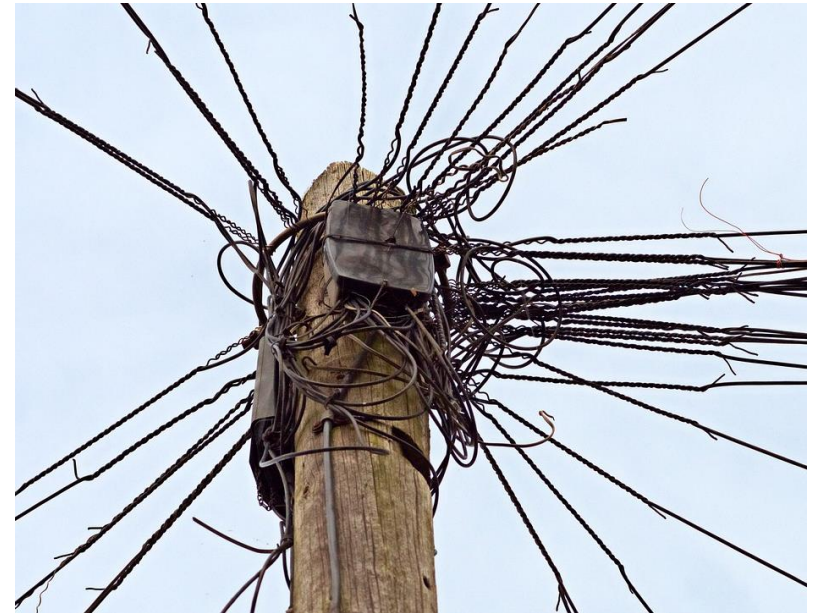
Research project with a proper data file structure. Image taken from CodeRefinery, Lesson on Reproducible Research. Shared under CC-BY 4.0.

SvalbardEx271020				
File	Home	Share	View	
SvalbardEx271020				
Name	Date modified	Type	Size	
P-10-01	26.10.2020 16:29	File folder		
SEM 2012	26.10.2020 16:25	File folder		
P-10-03-01.tif	11.06.2012 11:17	TIF File	1 880 KB	
P-10-03-02b.tif	11.06.2012 11:25	TIF File	1 880 KB	
P-10-03-02c.tif	11.06.2012 11:26	TIF File	1 880 KB	
P-10-03-05.tif	11.06.2012 11:38	TIF File	1 880 KB	
P-10-03-79b.tif	11.06.2012 12:59	TIF File	1 880 KB	
P-10-03-92b.tif	11.06.2012 13:26	TIF File	1 880 KB	
P-10-03-109.tif	11.06.2012 13:49	TIF File	1 880 KB	
P-10-03-119b.tif	11.06.2012 14:02	TIF File	1 880 KB	
P-10-03-193b.tif	11.06.2012 15:45	TIF File	1 880 KB	
P-10-03-222b.tif	11.06.2012 16:29	TIF File	1 880 KB	
P-10-03-222c.tif	11.06.2012 16:30	TIF File	1 880 KB	
P-10-03-226b.tif	12.06.2012 09:33	TIF File	1 880 KB	
P-10-03-226c.tif	12.06.2012 09:33	TIF File	1 880 KB	
Gasser_2014.pdf	27.01.2015 13:11	Adobe Acrobat D...	5 127 KB	
Gernigon_and_Bronner_2012.pdf	30.01.2015 15:07	Adobe Acrobat D...	3 445 KB	
Gernigon_et_al_2014.pdf	26.01.2015 16:02	Adobe Acrobat D...	54 447 KB	
Glorstad_Clark-2010.pdf	15.09.2014 15:06	Adobe Acrobat D...	10 269 KB	
Gronlie et al. 1980 Seismic inversion of Bj...	09.10.2012 13:28	Adobe Acrobat D...	643 KB	
Gudlaugsson_et_al_1998.pdf	17.10.2014 12:50	Adobe Acrobat D...	3 411 KB	
Harland_and_Gayer_1972.pdf	17.10.2014 11:47	Adobe Acrobat D...	1 474 KB	
Høy_and_Lundchien_2011_NBarenstSea_...	17.09.2014 14:10	Adobe Acrobat D...	13 600 KB	
Isaksen 1996 Organic geocem Bjornoya.p...	09.10.2012 13:24	Adobe Acrobat D...	1 439 KB	
Klausen et al 2015 Triassic Snadd in Baren...	30.03.2015 13:15	Adobe Acrobat D...	14 911 KB	
Klausen_et_al_2014_Triassic_Snadd_Fm.pdf	08.09.2014 13:20	Adobe Acrobat D...	5 069 KB	
Klitzke_et_al_Barents_Sea_Region_2014-pr...	21.01.2015 15:22	Adobe Acrobat D...	2 951 KB	
Gabrielsen et al. 1990_Structural_E_BS_NP...	11.03.2013 15:45	Adobe Acrobat D...	9 099 KB	
Gac_et_al_2013_ultra_deep_EBB.pdf	20.01.2015 12:15	Adobe Acrobat D...	1 008 KB	
Abstract.doc	15.09.2014 16:04	Microsoft Word 9...	234 KB	
Copy of NuAge_090214-120115-run1.xls	22.01.2015 13:06	Microsoft Excel 97...	1 390 KB	
CPP_Svalbard_Bjornoya_Copy of CUMUL...	17.12.2014 15:35	Microsoft Excel 97...	2 429 KB	
2013 Wintermeeting poster-1_Edina_Final...	07.01.2013 15:12	Adobe Acrobat D...	130 KB	
ICPMS data	26.10.2020 16:31	File folder		
Franz Josef presentation.pptx	09.01.2013 12:06	Microsoft PowerP...	11 334 KB	

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

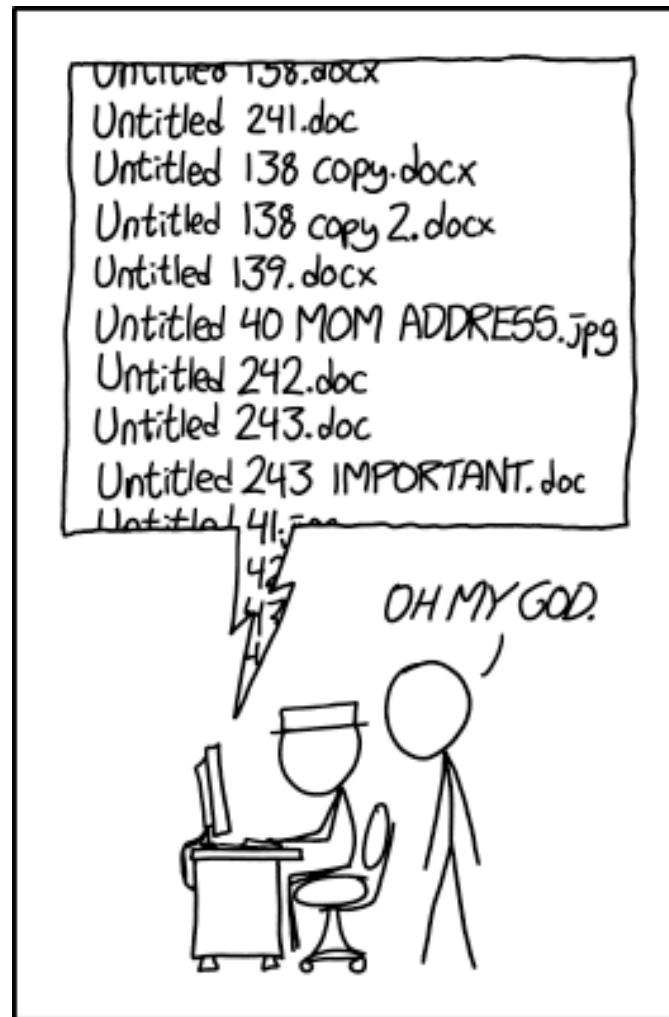
While organising your data...

- 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



File naming conventions (FNC)

- **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.

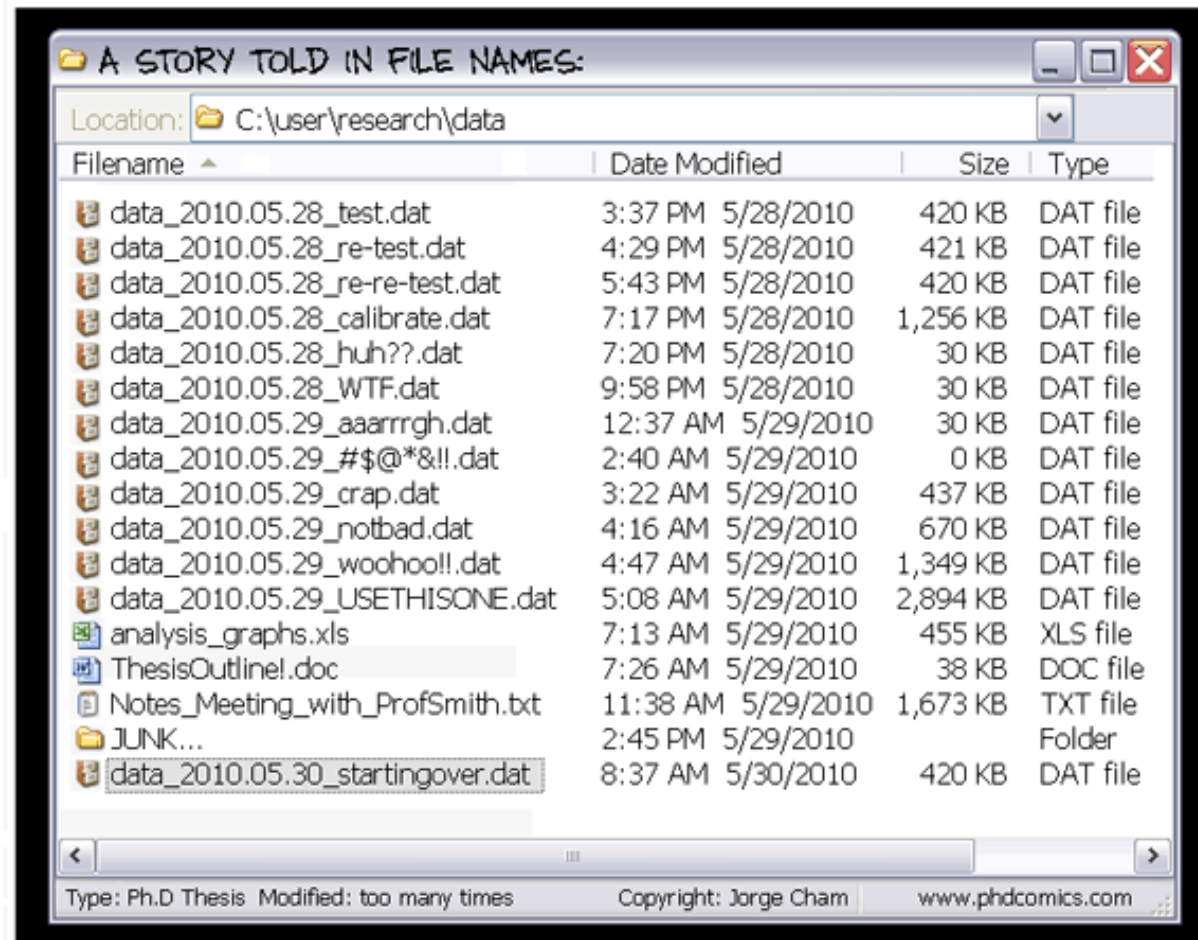
Avoid using the following characters in Folder and File names:

# 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!



Versioning

- Refers to saving new copies of your files when you make changes so that later you can go back and retrieve specific versions of your files
 - DataFileName_1.0 = original document
 - DataFileName_1.1 = original document with minor revisions
 - DataFileName_2.0 = document with substantial revisions

Versioning

- Refers to saving new copies of your files when you make changes so that later you can go back and retrieve specific versions of your files

- image1_v1.jpg
- image1_v2.jpg
- image2_v1.jpg
- image2_v2.jpg
- ...

- image1_v1.jpg
- image1_v10.jpg
- image1_v2.jpg
- ...

Versioning

- Refers to saving new copies of your files when you make changes so that later you can go back and retrieve specific versions of your files

- image1_20151021
- image1_20151214
- image1_20160123
- ...

- dataset1_20160402_KES
- dataset1_20160301_WTC
- dataset1_20160814_GSC
- ...

Version control

- Perfect for collaboration
- Tracking and managing changes to a file or set of files over time
- You will be able to recall any version at any time
- It can be used for documents, software development, large websites etc.

Version control system

- Git: Free and open source version control system



Version control system

- Git: Free and open-source version control system
- GitHub: is an internet hosting service for software development and version control using Git



git



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 DataverseNO and Data Archiving and Networked Services (DANS):
 - <https://site.uit.no/dataverseno/deposit/prepare/#preferred-file-formats>
 - <https://dans.knaw.nl/en/file-formats/>

Services and tools for RDM support at UiO:

- Research support and IT staff at your faculty/institute
- Subject librarian for your discipline
- [UiO Research data management website](#) (English)
- Research data group at DSC: research-data@uio.no
- Norwegian Centre for Research Data (NSD, now Sikt)



Questions?
Comments?

Upcoming sessions

Workshop on Data Management Plans

Apr. 19, 2:00 PM, Georg Sverdrups hus: DSC-Oasen

Data classification and storage selection

Apr. 21, 9:00 AM, Zoom

Copyright and licensing

Apr. 21, 1:00 PM, Zoom

Workshop on Data Management Plans

Apr. 24, 12:00 PM, Georg Sverdrups hus: DSC-Oasen

Sharing and archiving research data

Apr. 26, 9:00 AM, Zoom

Finding and reusing research data

Apr. 27, 9:00 AM, Zoom

Digital Scholarship Centre

At the Digital Scholarship Centre (DSC) you get guidance on how you can make the best possible use of digital tools and methods in your research and communication activities.

Open Access →

Information about open access publishing, publisher agreements, self-archiving, requirements, and guidelines.

Open and reproducible research →

Make your research more transparent and reproducible.

Research Data Management →

Managing your data both during and after a research project.

Text-mining →

Information about digital tools for searching, mining, and analysing textual data.

Systematic search →

Information about systematic literature searching, how to get started, and how to get help.

Visualisation →

Use of visual methods to explore, communicate and understand data.

Carpentry@UiO →

Offers workshops in foundational digital skills such as coding and data management.

Reference management →

Styles, tools, and information on reference management.

Digiforsk blog →

A blog about work and resources at UiO where digital research methods are involved. Posts in both Norwegian and English.

Research data management

Welcome to UiO's data management pages maintained by the research data group at [Digital Scholarship Center](#).

[Data management plans \(DMPs\)](#) >

[Data classification and storage](#) >

[Data organization](#) >

[Data documentation and metadata](#) >

[Data sharing and publishing](#) >

[Finding and reusing data](#) >

Glossary

An overview of [often used terms](#) in research data management.

Laws and ethics

[Information on privacy and data protection at UiO](#)

[Intellectual property rights at UiO](#)

[National guidelines from Research Ethics Committees](#)

Courses and support

[Currently offered courses](#)

Need advice? Feel free to [book an individual appointment](#) or contact us at **research-data@uio.no**

Are you a data manager? Join our [data managers network!](#)



Carpentry@UiO

Carpentry@UiO is a community of people who are passionate about learning, teaching, and sharing best practices and digital skills for making the research process more reproducible and effective. If you want to get involved, or join one of our workshops, check us out!



The Unix Shell

Shell speeds up repetitive and tedious processes. It is also essential skills needed to use high-performance computing (HPC) resources.



Version Control with Git

Git helps you to keep track of what you've done, for a better collaboration and for yourself in future. In the workshop we use GitHub as well.



Programming in Python

Python is now widely used in scientific computing with various powerful packages. Carpentry@UiO runs workshops for participants with no programming experience ("Plotting and Programming in Python" lesson) and for participants at intermediate level ("Programming with Python" lesson, episodes 10-12).



R for Reproducible Scientific Analysis

R is commonly used for statistical analysis, but it is also a powerful programming language. Workshops on R focuses on teaching best practices for scientific computing: breaking down analyses into modular units, task automation, and encapsulation. Workshops on R may use lessons from Data Carpentry instead.



Using Databases and SQL

Databases include powerful tools for search and analysis, and can handle large, complex data sets. The lesson will show how to use a database to explore research data by using SQL.

<https://www.ub.uio.no/english/courses-events/courses/other/Carpentry/>



University of Oslo Library

Digital Scholarship Center and Open Research

Open Science Lunch 2023

Each last Thursday of the month at 12.00 we invite you to join us virtually for an online open lunch to hear about how to make your research more open.

February 23rd

Open your data with DataverseNO

March 30th

Public health reporting: an open source approach

April 27th

Open knowledge resources: Store norske leksikon and Wikipedia

May 25th

Keep the rights to your work: UiO's Rights Retention Policy



Time and place: Apr. 27, 2023 12:00 PM–1:00 PM, Hybrid: Georg Sverdrups hus and Zoom

Open knowledge resources: Store norske leksikon and Wikipedia

Join us for a discussion on using, maintaining and contributing to freely available open knowledge resources.



Time and place: May 25, 2023 12:00 PM–1:00 PM, Hybrid: Georg Sverdrups hus and Zoom

Keep the rights to your work: UiO's Rights Retention Policy

Learn about how you can retain the rights to your published work with the new Rights Retention Policy at UiO.

ReproducibiliTea

Journal Club

**JOIN IN AND DISCUSS WITH FELLOW
STUDENTS AND RESEARCHERS**

**OPEN RESEARCH, REPRODUCIBILITY
and RESEARCH IMPROVEMENT**



Join us

Everyone is welcome to join us - whether you are an enthusiast of open and reproducible research, a skeptic, or a cautious explorer. Currently, all meetings are hybrid with the possibility of joining on-site at Blindern or via Zoom. Grab a cup of tea (coffee?) and join us!

Subscribe to our mailing list





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

Thank you!

Contact us at:

research-data@uio.no

Sources

- Deutz DB, Buss MCH, Hansen JS, Hansen KK, Kjellmann 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>