UiT The Arctic University of Norway

How to structure and document research data

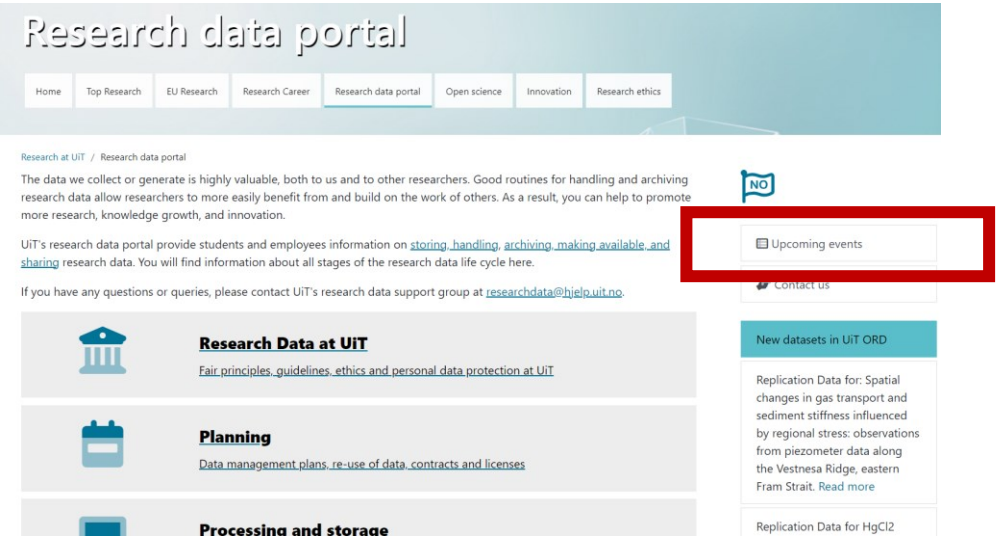
Noortje Haugstvedt, PhD

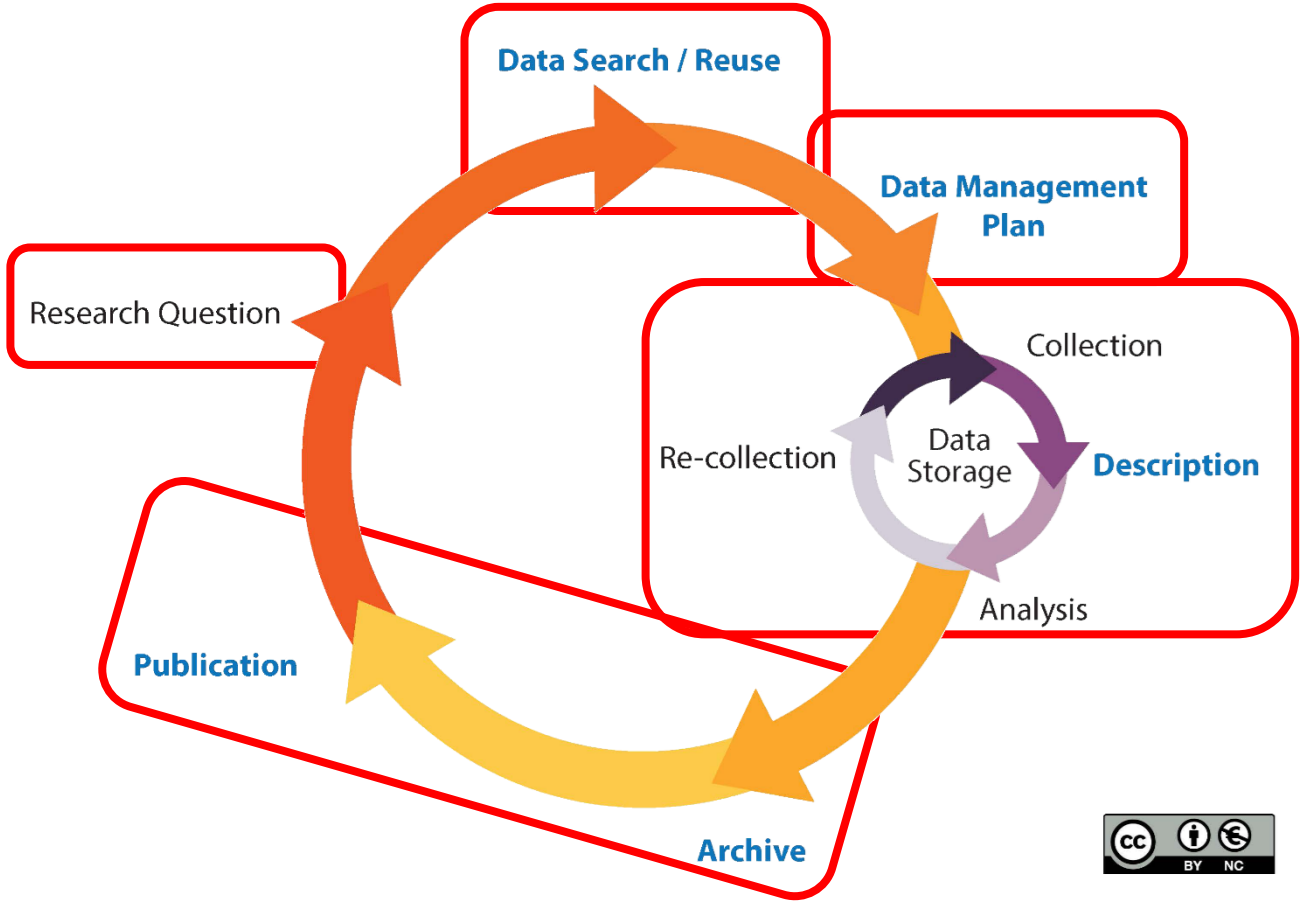
Adrian Verhoef, PhD

17.04.2023 – University Library

The research data management lifecycle & webinars at UiT

Find the webinars on
the Research Data Portal
under [Upcoming events (link)]





Adapted original source:
The University of California, Santa Cruz,
Data Management LibGuide, Research Data Management Lifecycle, diagram,
viewed May 2, 2016 at <http://guides.library.ucsc.edu/datamanagement>

Learning objectives for this webinar

Why is it important to **structure** and **document** your data?

How to do it in a **persistent** way?

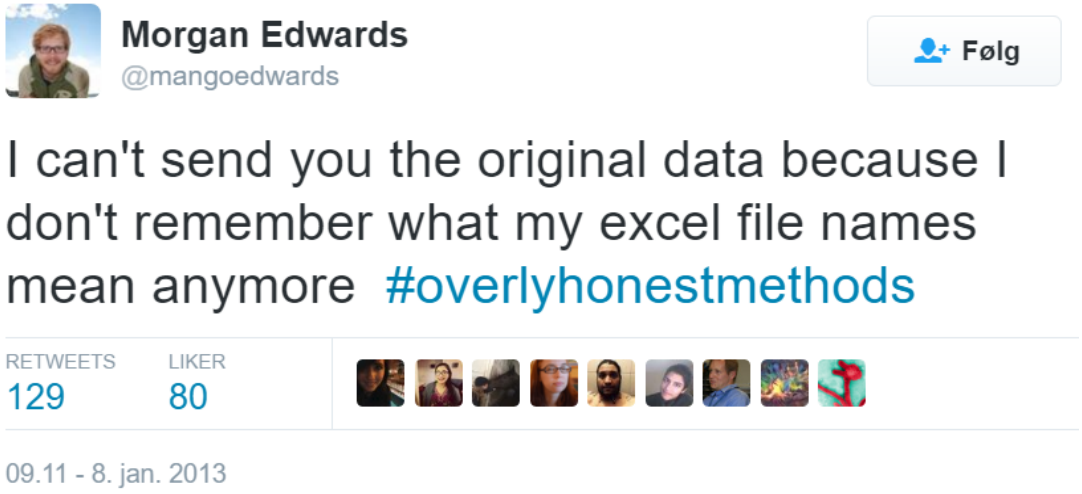
Where to find more **information** and **help?**

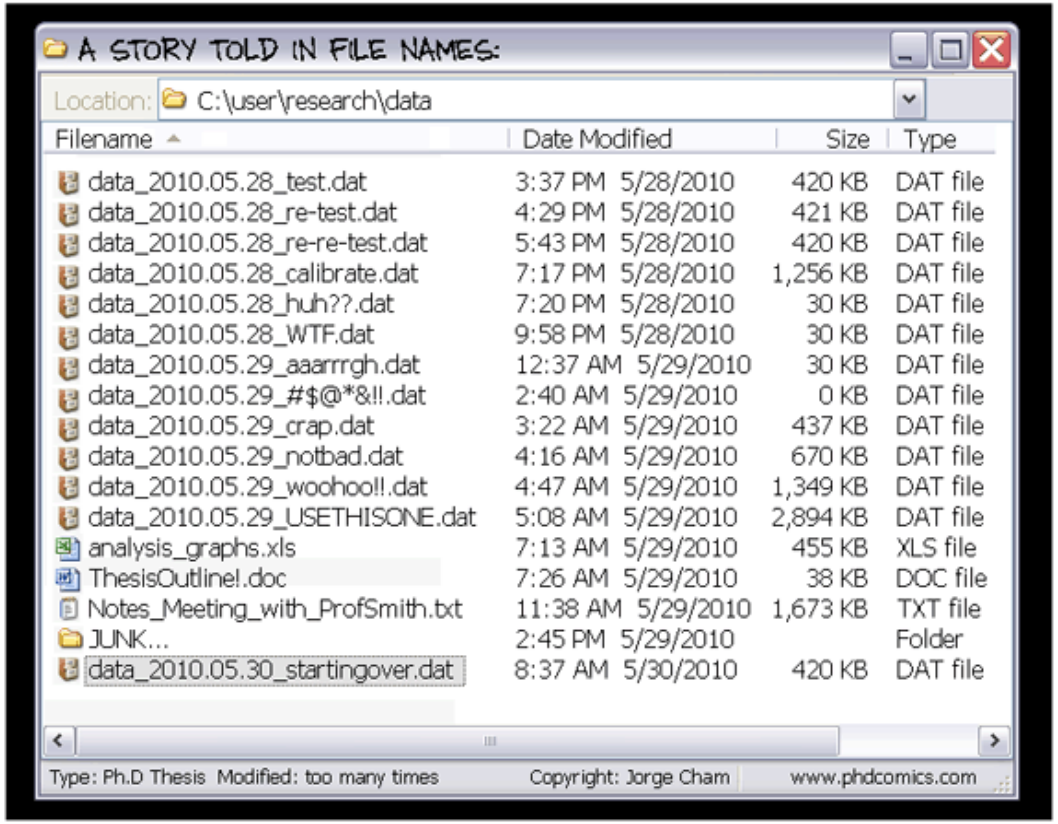
Please feel free to interrupt for questions or comments!

Or use the chat to post questions and comments.

Why?

Do you know where your data are?





http://phdcomics.com/comics/archive.php?comicid=1323

Why?

Retraction Watch

Tracking retractions

Doing the right thing: Authors retract brain paper with "systematic human error in coding"

with one comment

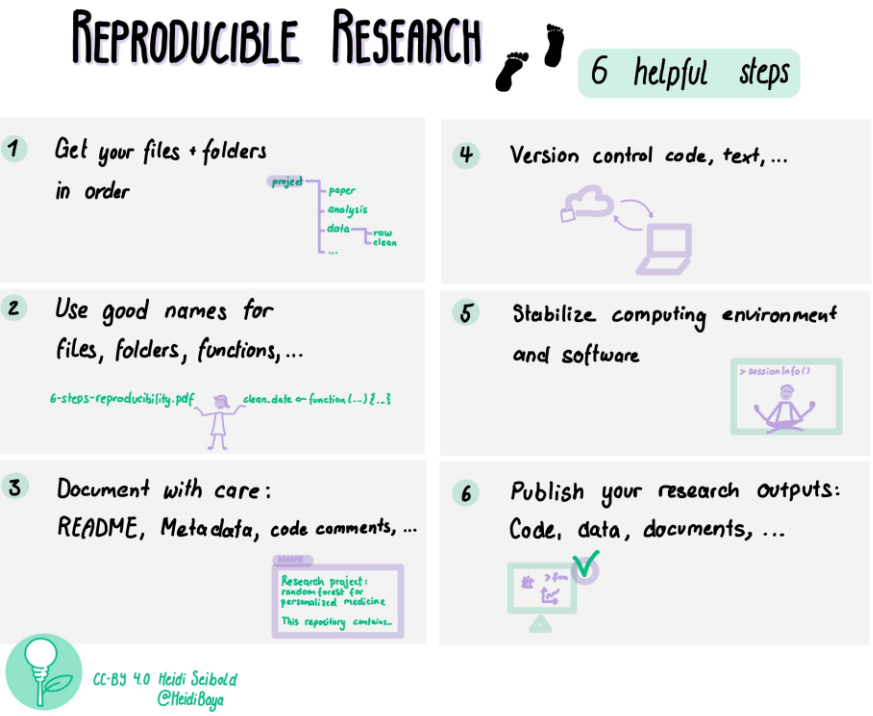
A group of Swiss neurologists have lost their 2013 article in *Frontiers in Human Neuroscience* after reporting that their data were rendered null by coding errors.

frontiers in HUMAN NEUROSCIENCE 2.9

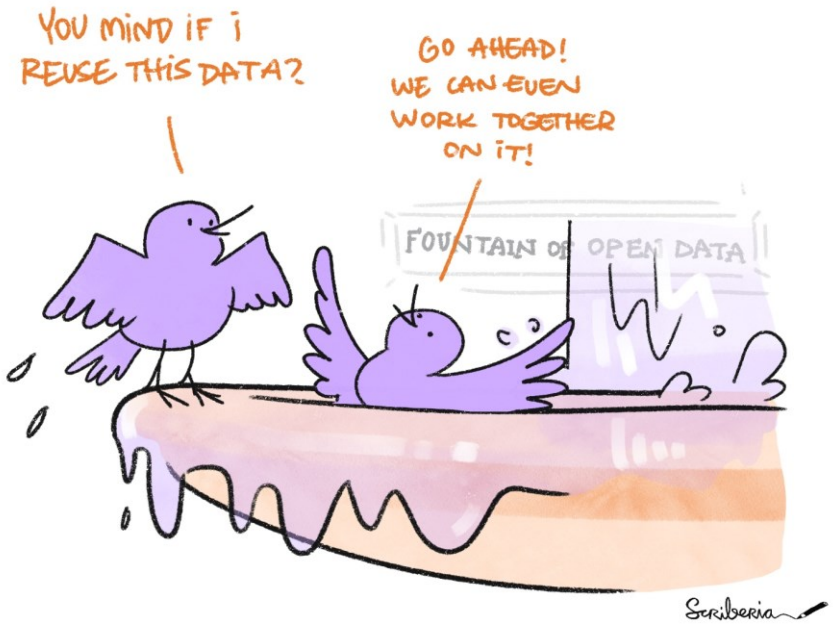
"..**a systematic human error in coding the name of the files** had been made during the extraction of the EEG template topographic maps best differentiating the two experimental conditions at the single subject level."

http://retractionwatch.com/2014/01/07/doing-the-right-thing-authors-retract-brain-paper-with-systematic-human-error-in-coding/

Why?



Heidi Seibold, CC-BY 4.0, https://twitter.com/HeidiBaya/status/1579385587865649153



Scriberia, CC-BY 4.0 DOI: 10.5281/zenodo.3332807.

Avoid data loss

Make sure your data are **understandable**
even several years from now

The most important elements:

- Data storage
- Structuring of files and folders
- Names on data files and folders
- Documentation of data: ReadMe and metadata
- The file formats

DATA FOR FUTURE GENERATIONS

Plan the data structure early in the project – before you start collecting data.

How it started …….. How it’s going

Images created with DALL*E

Folder naming and organization

Folders can be useful to organize and structure your data

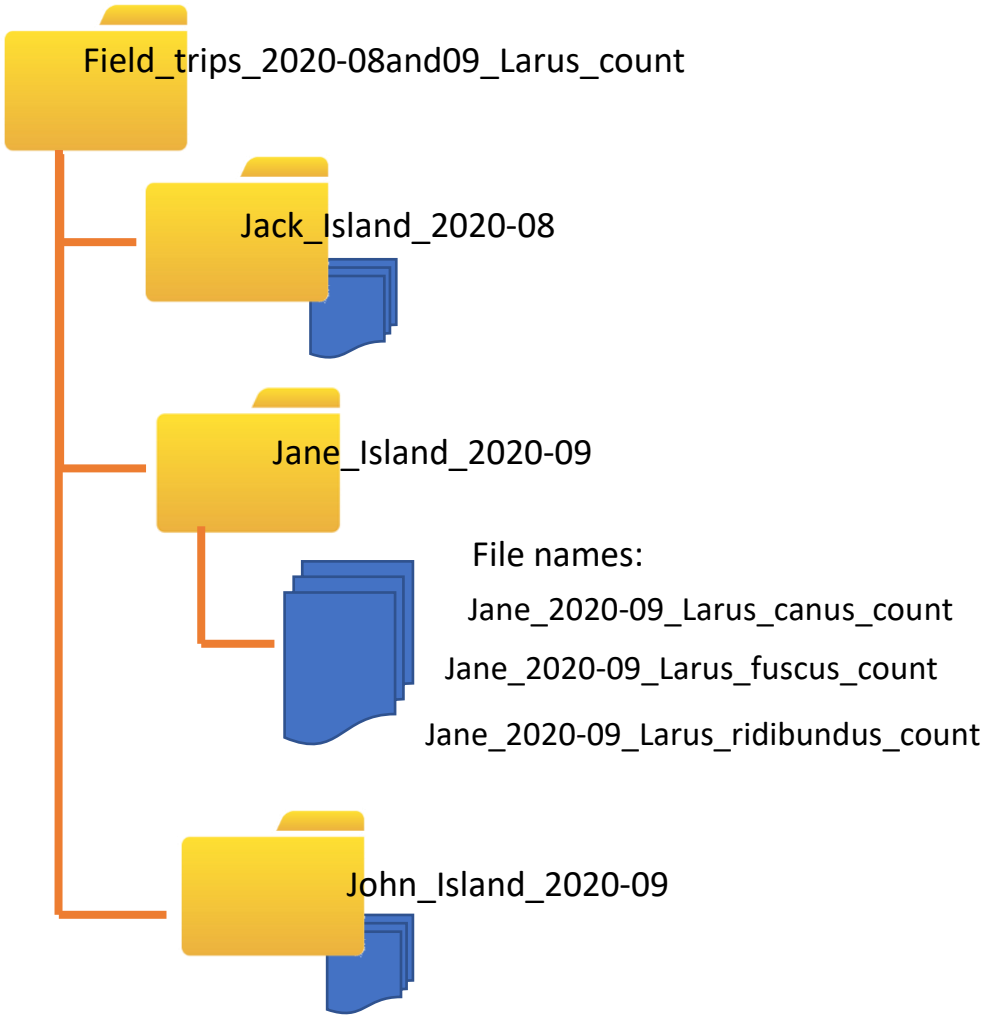
- especially when you have many files

Use a consistent strategy.

Main structure of data should be visible in the **file names**

=> also useful for archiving afterwards

Document file structure and naming convention (in a ReadMe file)



Names for files and folders

- Machine readable
  - Use underscores (_) instead of spaces
  - Avoid special characters
    " / \ : * . ? ' < > [ ] ( ) & $ æÆ øØ åÅ
- Human readable
  - Files should be named consistently
  - Descriptive, but short (< 25 signs)

- Use the international dating convention
  YYYY-MM-DD

✗

Joe's notes from today.txt

Tromsø&Ålesund.txt

Mynotes_versjon1.txt

Mine notater ny.txt

figure 1.png

Thesis_DONTdelete_new_draft_final*_last_v2.txt

☑

Joes_filenames_are_getting_better.txt

Tromsoe_og_aalesund.txt

2023-02-10_notes.txt

2023-02-11_notes.txt

Fig01_length-vs-interest.png

PhD_thesis_2023_finalversion.txt

Inspired by Jenny Brian's slides (link)

Use file names for sorting files by names

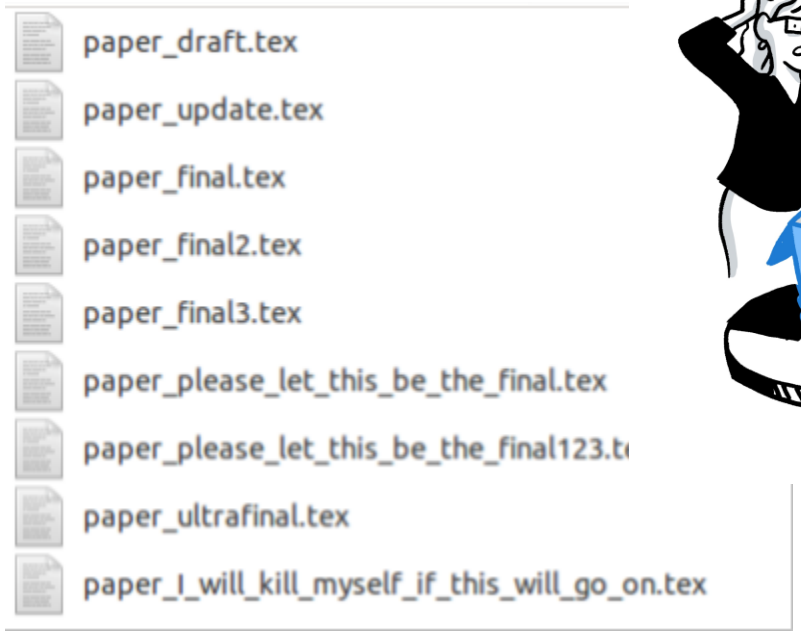
Sorted by date:
2020-08-01_notes_John.pdf
2020-08-31_observations_John.txt
2020-09-01_notes_Jane.pdf
2020-09-30_observations_Jane.pdf

Sorted by author:
Jane_notes_2020-09-01.pdf
Jane_observations_2020-09-30.txt
John_notes_2020-08-01.pdf
John_observations_2020-08-31.txt

Sorted on data type:
Notes_Jane_2020-09-01.pdf
Notes_John_2020-08-01.pdf
Observations_Jane_2020-09-30.txt
Observations_John_2020-08-31.txt

Forced numbering:
01_Notes_John_2020-08-01.pdf
02_Notes_Jane_2020-09-01.pdf
03_Observations_John_2020-08-31.txt
04_Observations_Jane_2020-09-30.txt

Use file names for versioning



The Turing Way Community, & Scriberia. (2020). Illustration from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.3695300

Bruk heller:

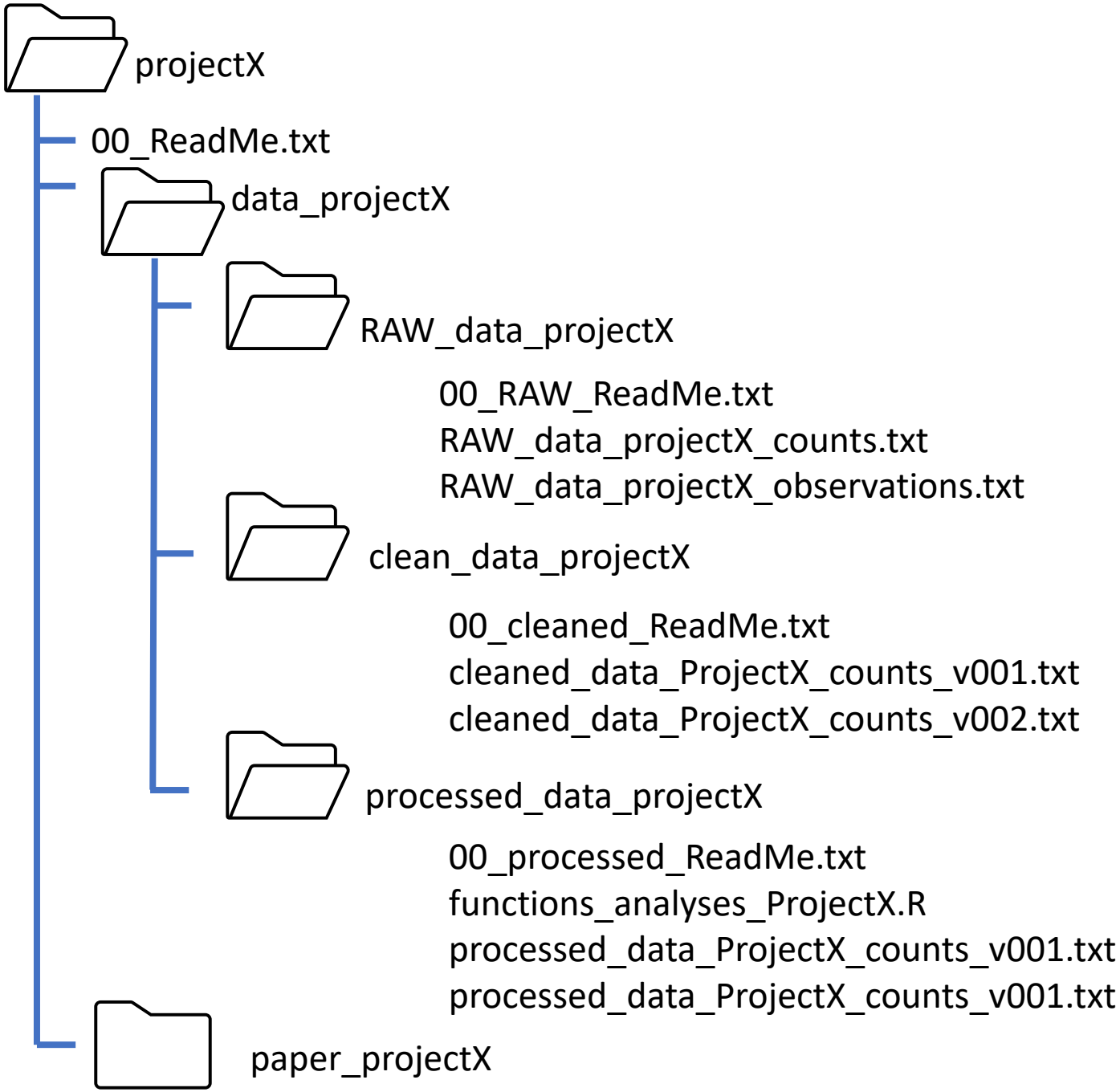
Larus_canus_counts_RAW

Larus_canus_counts_v001

Larus_canus_counts_v002

etc

Remember to document your changes in a ReadMe file


projectX
00_ReadMe.txt
data_projectX
RAW_data_projectX
00_RAW_ReadMe.txt
RAW_data_projectX_counts.txt
RAW_data_projectX_observations.txt
clean_data_projectX
00_cleaned_ReadMe.txt
cleaned_data_ProjectX_counts_v001.txt
cleaned_data_ProjectX_counts_v002.txt
processed_data_projectX
00_processed_ReadMe.txt
functions_analyses_ProjectX.R
processed_data_ProjectX_counts_v001.txt
processed_data_ProjectX_counts_v001.txt
paper_projectX

Documentation

Why?

To find, understand and re-use your data

For yourself - Also many years from now

For others – to understand and re-use your data correctly

Documentation

How?

ReadMe file = manual for your data

Human readable – important to understand and re-use your data correctly

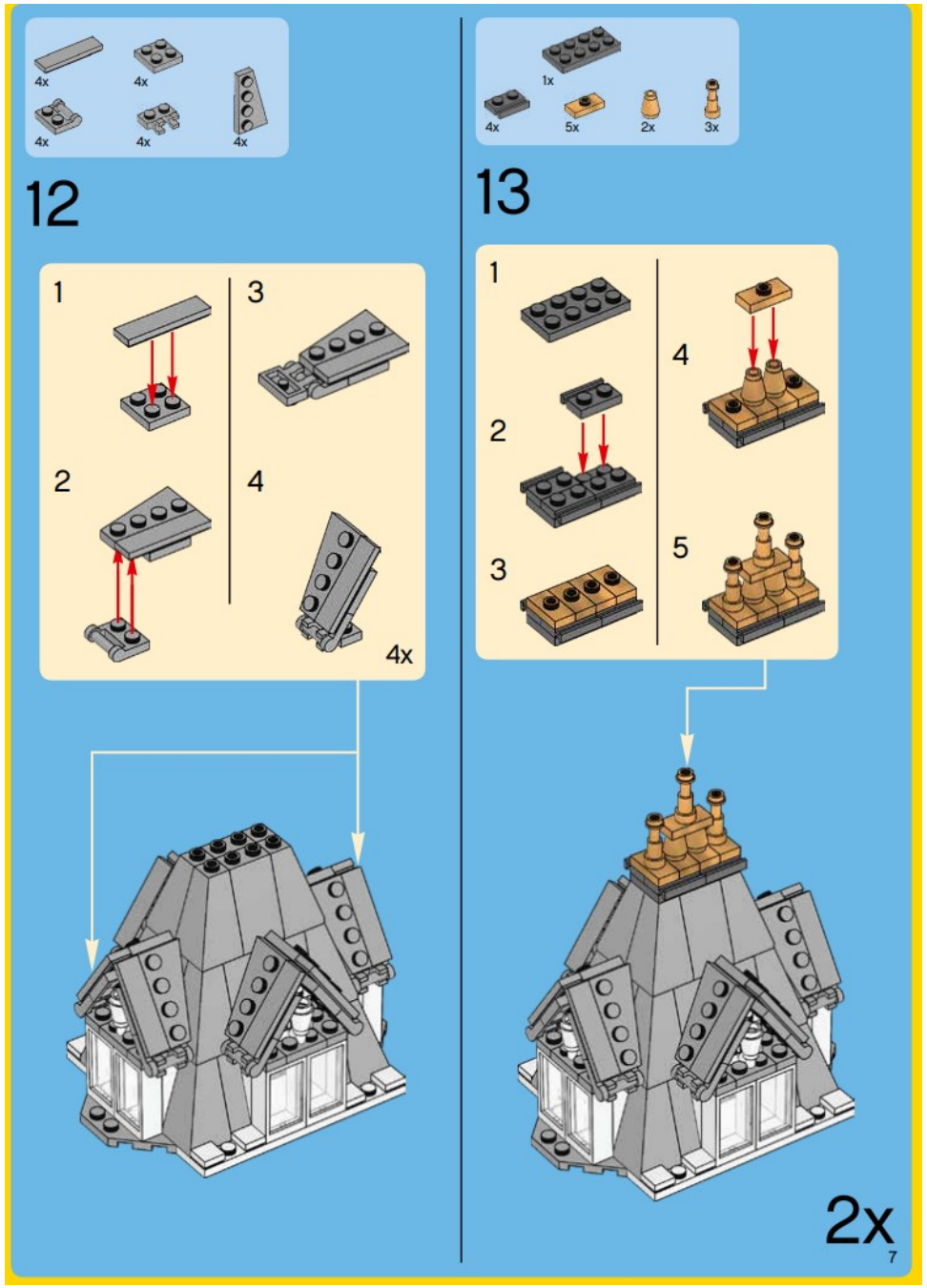
Metadata = data about your data

Computer readable – important for search and discovery of your data

When?

During the entire lifecycle

keep updated methodologies, code descriptions, lab notebooks, experimental protocols, provenance information for data, etc.



ReadMe-file: manual for your data

Enough information to understand and re-use your data

A roadmap for future users that informs how the data was generated, modified, processed and how to use it in the future.

Start documenting early, update continuously, in an open format (e.g. .txt file)

Have a look at ReadMe files in the repository where you are planning to archive your data

DataverseNO ReadMe-template (link)

ReadMe-file: manual for your data

General information about the project / dataset:

Title, project period, description, funding sources, project participants, contact information.

Methodology:

Data collection or generation data processing, data quality etc.

Data and file overview:

File overview, file formats, relation between files, version of the dataset.

Data-specific information:

Column headings, abbreviations, units of measure, contextual information

About the dataset:

Terms of reuse, related datasets, data sources.

0_README.txt

```
This README file was generated on 2021-01-25 by Thomas Karsten Kilvær.
Last updated: 2022-05-20.


-------------------
GENERAL INFORMATION
-------------------
// Title of Dataset: Replication Data for: A Pragmatic Machine Learning Approach to Quantify Tumor Infiltrating Lymphocytes in Whole Slide Images
// DOI: https://doi.org/10.18710/4YN9SZ
// Contact Information
<The person to be contacted for questions about the dataset>
     // Name: Kilvaer, Thomas K,
     // Institution: UiT The Arctic University of Norway
     // Email: thomas.k.kilvar@uit.no
     // ORCID: https://orcid.org/0000-0003-1669-0117

// Contributors: See metadata field Contributor.
// Kind of data: See metadata field Kind of Data.
// Date of data collection/generation: See metadata field Date of Collection.

// Description of dataset:
This dataset can be used to replicate the findings in "A Pragmatic Machine Learning Approach to Quantify Tumor Infiltrating Lymphocytes in Whole Slide Images".
The motivation for this paper is that increased levels of tumor infiltrating lymphocytes (TILs) indicate favorable outcomes in many types of cancer. Our aim is to leverage computational pathology to automatically
quantify TILs in standard diagnostic whole-tissue hematoxylin and eosin stained section slides (H&E slides). Our approach is to transfer an open source machine learning method for segmentation and classification of
nuclei in H&E slides trained on public data to TIL quantification without manual labeling of our data.
Our results show that additional augmentation improves model transferability when training on few samples/limited tissue types. Models trained with sufficient samples/tissue types do not benefit from our additional
augmentation policy. Further, the resulting TIL quantification correlates to patient prognosis and compares favorably to the current state-of-the-art method for immune cell detection in non-small lung cancer
(current standard CD8 cells in DAB stained TMAs HR 0.34 95% CI 0.17-0.68 vs TILs in HE WSIs: HoVer-Net PanNuke Aug Model HR 0.30 95% CI 0.15-0.60, HoVer-Net MoNuSAC Aug model HR 0.27 95% CI 0.14-0.53). Moreover, we
implemented a cloud based system to train, deploy and visually inspect machine learning based annotation for H&E slides. Our pragmatic approach bridges the gap between machine learning research, translational
clinical research and clinical implementation. However, validation in prospective studies is needed to assert that the method works in a clinical setting.
The dataset is comprised of three parts: 1) Twenty image patches with and without overlays used by pathologists to manually evaluate the output of the deep learning models, 2) The models trained and subsequently
used for inference in the paper, 3) the patient dataset with corresponding image patches used to clinically validate the output of the deep learning models.

--------------------------
METHODOLOGICAL INFORMATION
--------------------------
// Description of sources and methods used for collection/generation of data:


// Methods for processing the data:
* Models were aquired by running the following scripts from the HoVer-Net pipeline: <extract_patches.py>, <train.py>, <export.py>.
Configuration values used for generating config.yml via running ```sh generate.sh``` inside hover docker container
- consep_aug_linear_2-1.0: https://gist.github.com/nsh23/5e31ee910ca55fcb8c0076973374a717
- consep_standard-1.1: https://gist.github.com/nsh23/676a7f7d0d429bf845ac4afa59f6db5f
- pannuke_aug_p_linear_2-1.0: https://gist.github.com/nsh23/3b22307d981760761158c894308025c1
- pannuke_standard-1.0: https://gist.github.com/nsh23/633d4a45523c8c63dbff7b20a8d6ad9b
- monusac_standard-1.0: https://gist.github.com/nsh23/34ebaf35d6350145b2809fbb8844eecc
- monuasc_aug_p_linear_2-1.0: https://gist.github.com/nsh23/2c0aa35afcc5908742d844d28522595a
In order to use them copy config file for the target experiment to ```hovernet-pipeline/src``` and rename it as <config.yml>.

* Manual validation images (UiT_TILs/manual_validation.tar/.../) were aquired via HoVer-net inference part by running the following scripts from the HoVer-Net pipeline: <infer.py>, <process.py>.
Configuration values used for generating config.yml via running ```sh generate.sh``` inside hover docker container
// Facility-, instrument- or software-specific information needed to interpret the data:

* Transforming patch-level logs to patient-level that you get from HoVer-Net pipeline could be done via running 2 scripts:
First, convert image-level names and select counts for specific cell type with <counts.py> script (hovernet-pipeline/src/metrics/counts.py).
Second, aggregate quantification logs from image-level to patient-level (counts per 1000^2) and get (min, max, median, avg) numbers for each patient with <summarize.py> script (hovernet-pipeline/src/metrics/
summarize.py).

-------------------
DATA & FILE OVERVIEW
-------------------
// File List:
```

[ReadMe file example from a dataset in DataverseNO (link)]

Metadata: data about data

Metadata = description of data

Examples of metadata:

Author, **title**, **description**, ...

Keywords

Geographical information

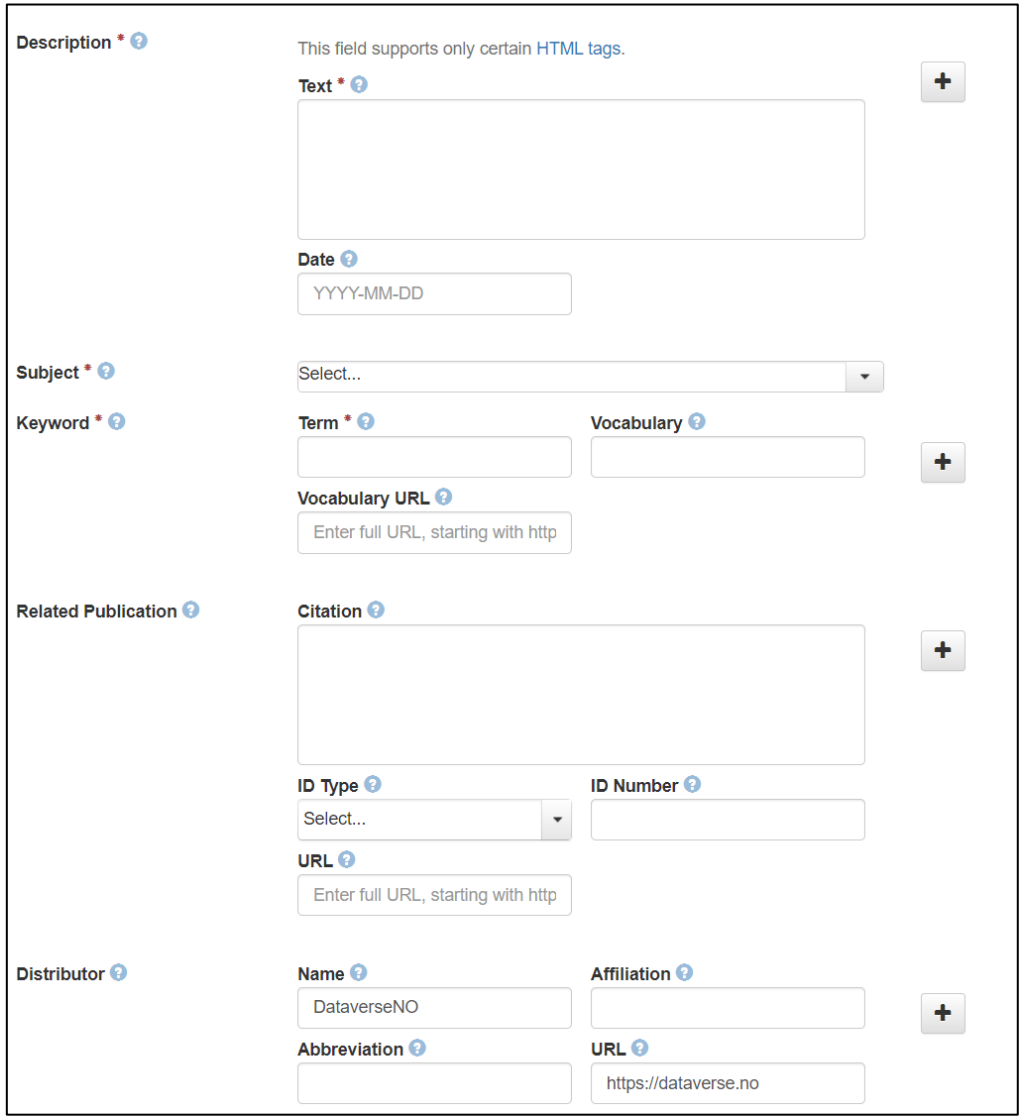
Standardized general metadata

e.g. international date format (e.g. ISO-8601): YYYY-MM-DD (2019-12-09)

Dublin Core (link), Data Documentation Initiative (link)

Domain-specific standardized metadata

e.g. Darwin Core(link) = a standard for description of data on biological diversity.



Metadata : data about data

Overview over different standards:

[Research Data Alliance (link)]

[FAIRSharing.org (link)]

[Digital Curation Centre (link)]

Tip: Have a look at metadata in the repository where you are planning on archiving your data. To see how they do it.

DataverseNO Certified by: CORE TRUST SEAL

Search Deposit Guide Support Kathleen (Katie) Alexandra Smart

UiT Open Research Data

DataverseNO > UiT Open Research Data >

UiT_TILs - Replication Data for "A Pragmatic Machine Learning Approach to Quantify Tumor Infiltrating Lymphocytes in Whole Slide Images"

Version 2.0

Kilvaer, Thomas K, 2021, "UiT_TILs - Replication Data for "A Pragmatic Machine Learning Approach to Quantify Tumor Infiltrating Lymphocytes in Whole Slide Images"", https://doi.org/10.18710/4YN9SZ, DataverseNO, V2

Cite Dataset Learn about Data Citation Standards.

Access Dataset
Edit Dataset
Link Dataset
Contact Owner Share

Dataset Metrics
117 Downloads

Description

This dataset can be used to replicate the findings in "A Pragmatic Machine Learning Approach to Quantify Tumor Infiltrating Lymphocytes in Whole Slide Images". The motivation for this paper is that increased levels of tumor infiltrating lymphocytes (TILs) indicate favorable outcomes in many types of cancer. Our aim is to leverage computational pathology to automatically quantify TILs in standard diagnostic whole-tissue hematoxylin and eosin stained section slides (H&E slides). Our approach is to transfer an open source machine learning method for segmentation and classification of nuclei in H&E slides trained on public data to TIL quantification without manual labeling of our data. Our results show that improved data augmentation improves immune cell detection in H&E WSIs. Moreover, the resulting TIL quantification correlates to patient prognosis and compares favorably to

Read full Description [+]

Subject Medicine, Health and Life Sciences

Keyword machine learning, ML, deep learning, DL, non-small cell lung cancer, NSCLC, immune cell, tissue infiltrating lymphycytes, TIL

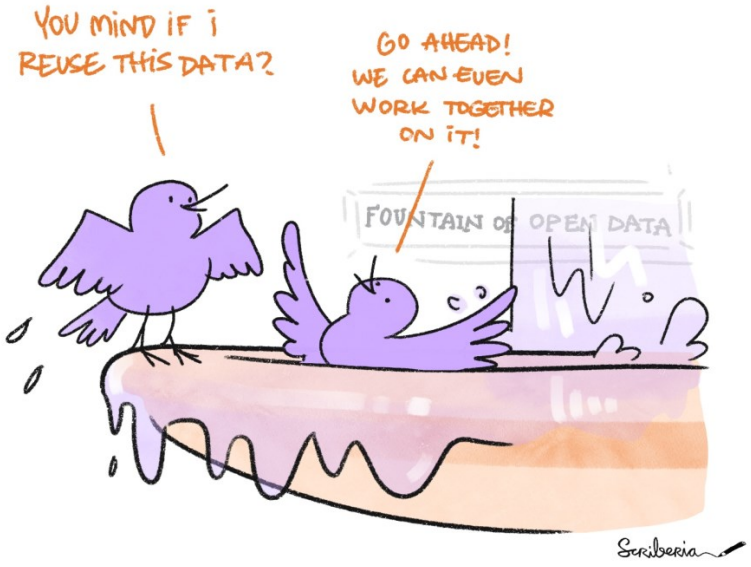
Related Publication submitted for review

Files Metadata Terms Versions

Metadata example from DataverseNO (link)

Preparing for archiving

- Selection
  - Data necessary to understand and replicate the study
  - Do not exclude negative data – meaning data that do not support the tested hypothesis
  - Include raw version of the data and processed version(s)
- Anonymization and/or aggregation?
  - Anonymize personal data
- Provide your data in original AND preferred (persistent) file formats
  - ensure the long-term use of your files
- Note: Webinars tomorrow & Wednesday (17 & 18 April 2023): How to archive



Preferred file formats

Characteristics of preferred file formats:

- non-proprietary
- open, with documented international standards
- in common usage by the research community
- using standard character encodings (e.g. ASCII, UTF-8)
- uncompressed (space permitting)

Archiving: Persistent file formats

File type	Preferred file formats (examples)	Non-preferred file formats (examples)
Audio	→ Uncompressed and lossless Wav or AIFF (.wav/.aiff) → Compressed and lossless FLAC (.flac) → Compressed and lossy Mp3 (.mp3)	→ AAC (.m4a) → Monkey's Audio (.ape) → Ogg Vorbis (.ogg) → Windows Media Audio (.wma)
Container file	Container files are automatically unpacked when uploaded and should only be used to keep the folder structure in your dataset; see more in section Upload data files.	In case container files need to be archived as container files, use .zip. Note! In this case, files must be packed twice. That way, the inner container will be preserved when uploaded.
Image	→ Uncompressed TIFF (.tif or .tiff) → Compressed and lossless PNG (.png) → Compressed and lossy JPEG (.jpg)	→ Adobe Photoshop (.psd) → Apple Picture File (.pct) → Graphics Interchange Format (.gif) → Raw Image Data File (.raw) → Windows Bitmap (.bmp)
Text (slides, illustrations)	→ PDF/A (.pdf) combined with original file	→ PowerPoint (.pptx)
Text (tables)	→ Tab separated Unicode plain text (.txt)	→ Excel (.xlsx)
	→ Plain text (.txt)	See the User Guide of DataverseNO for more information (link)

DataverseNO

Search User Guide Support Log In

DataverseNO

(UiT The Arctic University of Norway)

User guide

Metrics 401,523 Downloads

Contact Share

Western Norway University of Applied Sciences
HVL Open Research Data

Inland Norway University of Applied Sciences
INN Open Research Data

NORWEGIAN SCHOOL OF THEOLOGY, RELIGION AND SOCIETY
MF Norwegian School of Theology, Religion and Society

NMBU Open Research Data
NMBU Open Research Data

Search this dataverse... Advanced Search

- [x] Dataverses (24)
- [x] Datasets (1,363)
- [ ] Files (102,340)

Dataverse Category

Organization or Institution (13)
Research Project (5)
Research Group (3)
Department (1)
Researcher (1)

Publication Year

2023 (48)
2022 (171)
2021 (298)
2020 (185)
2019 (375)

1 to 10 of 1,387 Results

Sort

Replication Data for: Spatial changes in gas transport and sediment stiffness influenced by regional stress: observations from piezometer data along the Vestnesa Ridge, eastern Fram Strait

Apr 11, 2023 - UiT Open Research Data

Plaza-Faverola, Andreia; Sultan, Nabil, 2023, "Replication Data for: Spatial changes in gas transport and sediment stiffness influenced by regional stress: observations from piezometer data along the Vestnesa Ridge, eastern Fram Strait", https://doi.org/10.18710/GUX2O8, DataverseNO, V2

This database comprises piezometer pore-pressure and temperature data, analyses from Calypso sediment cores (grain sizes and logs) and results from geotechnical tests (index tests and oedometer tests), used for the study of seepage systems along the Vestnesa Ridge, in the eastern...

Adaptive management and social-ecological recovery in Athabasca oil sands mine reclamation

Apr 11, 2023 - NMBU Open Research Data

Gouin, Clayton, 2023, "Adaptive management and social-ecological recovery in Athabasca oil sands mine reclamation", https://doi.org/10.18710/8YJFTX, DataverseNO, V1

Digital survey and interview data for Ph.D. research project on social-ecological reclamation of oil sands mine sites and adaptive management in site and regional reclamation. Data collected as requirement of Ph.D. project. Selected participants data submitted based on consent fo...

Two introductory videos

[The research data management life cycle (link to Vimeo)]

[The FAIR data principles (link to Vimeo)]

More information and help

UiT Research Data Portal (link)

- Tips
- Overview of webinars/courses

DataverseNO deposit guidelines (link)

- Tips on how to prepare and describe data

Email us!

researchdata@hjelp.uit.no

Are you a Data Steward?

Join our Data Steward Network!

Sign up!

Join the Kick-off seminar 3 May!

More information on [Tavla (link)]

Evaluation

We are constantly working to improve the content of our webinars. Feedback from you will be of great help to us.

Please answer our 2 minute questionaire (link)

Date: 16.04.20223

Course code: Research data

researchdata@hjelp.uit.no

Noortje Haugstvedt – Adrian Verhoef