UiT Norges arktiske universitet

Hvordan strukturere og dokumentere forskningsdata

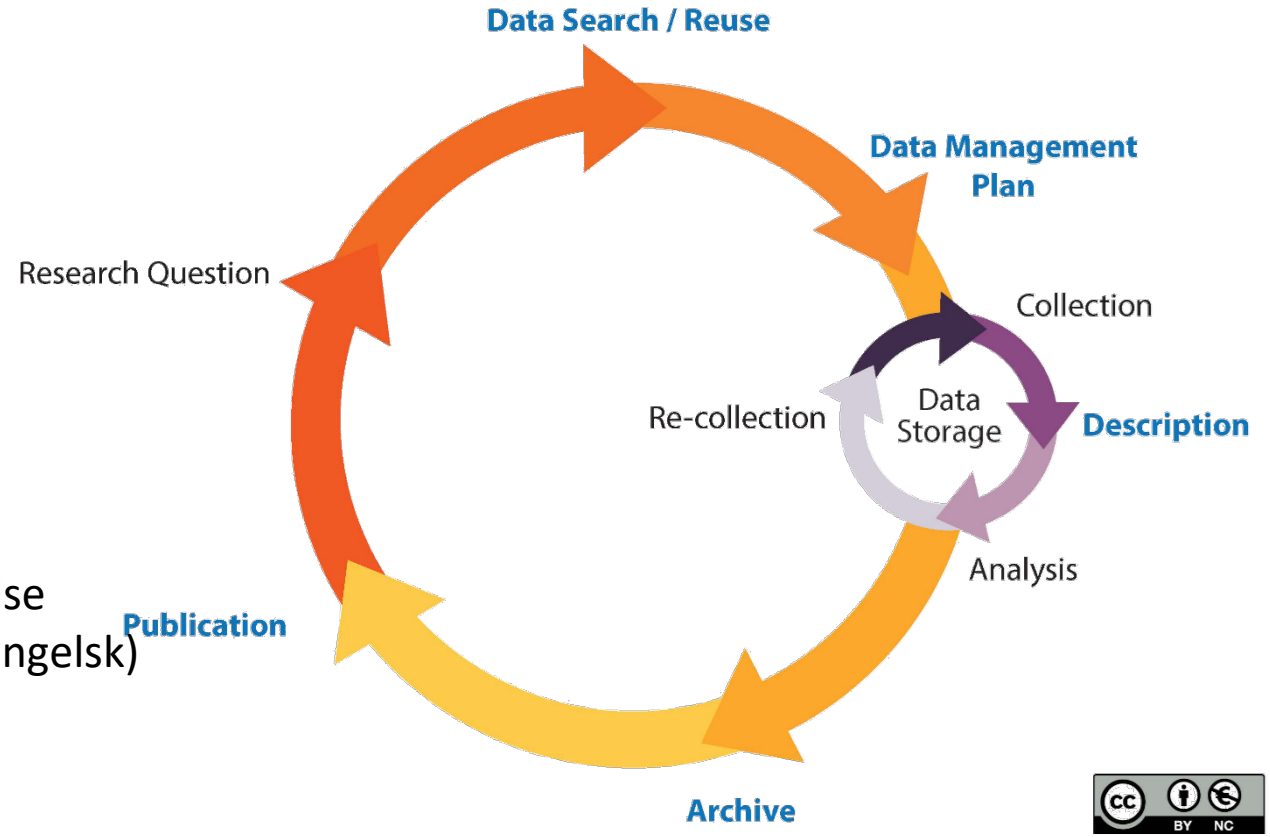
Noortje Haugstvedt

Leif Longva

25.09.2023 – UB, UiT

Forskningsdata og kurs på UiT

25. september: Datavask (på engelsk)
25. september: Lagring av forskningsdata
26. september: Hvordan arkivere forskningsdata
26. september: Hvordan arkivere data i DataverseNO
28. september: Forskningsdata: Rettigheter og lisenser
28. september: Hvordan skrive en datahåndteringsplan
29. september: Håndtering av data som trenger beskyttelse
17-19. og 23. oktober: Kurs i FAIR data visualisation (på engelsk)



Mer info på [Forskningsdataportalen](#) > arrangementer

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>

Læringsmål for denne modulen

- Forstå **hvorfor** er det viktig med god **strukturering** og **dokumentasjon** av forskningsdata
- Vite **hvordan** gjøre det på en **bevaringsverdig** måte
- Vite **hvor** du kan finne mer **informasjon** og få **hjelp**

Avbryt når du har spørsmål eller kommentarer!

Handout : RDM Training @ UiT Zenodo collection

Hvorfor?

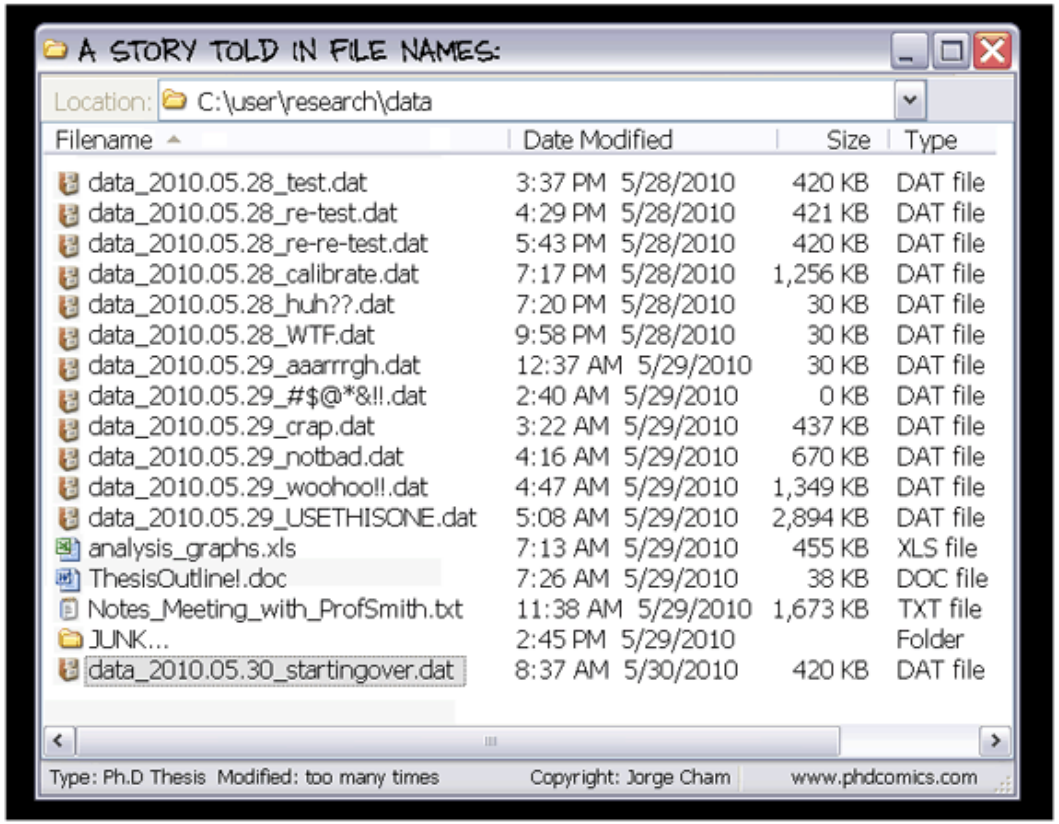
Morgan Edwards
@mangoedwards

Følg

I can't send you the original data because I don't remember what my excel file names mean anymore #overlyhonestmethods

RETWEETS 129 LIKER 80

09.11 - 8. jan. 2013



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

Hvorfor?

Retraction Watch

Tracking retractions

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

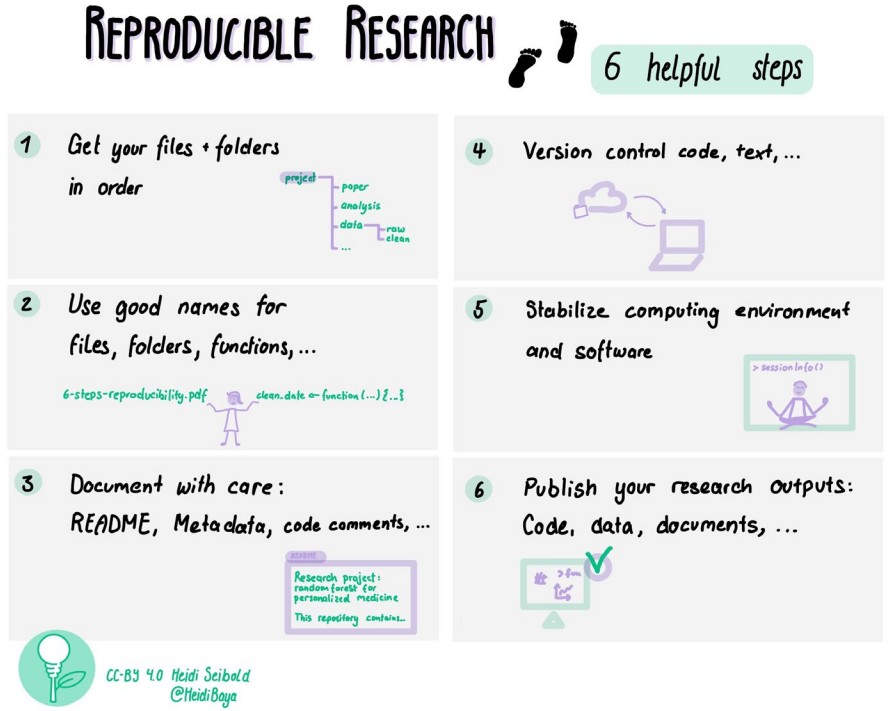
with one comment

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.

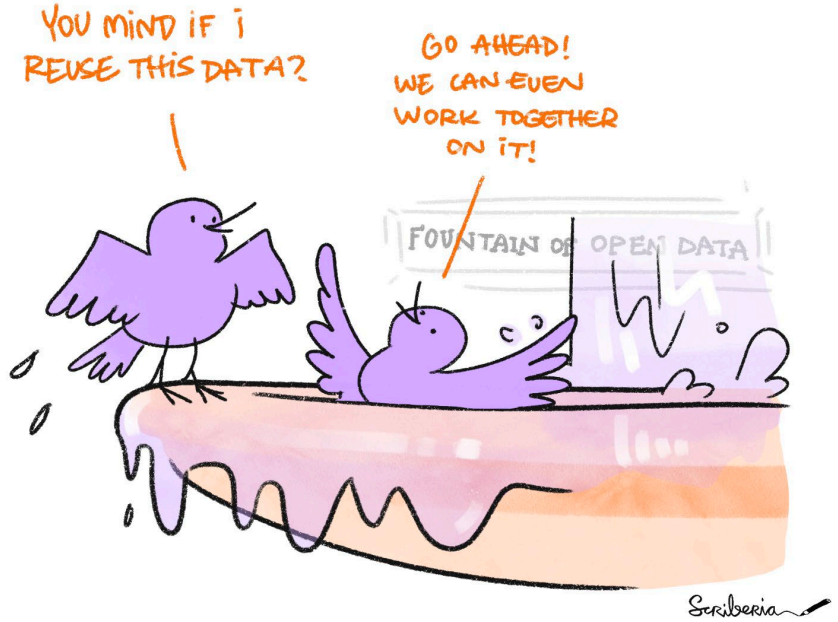
“..**a systematic human error in coding the name of the files** had been made during the extraction of the EEG template topographic maps best differentiating the two experimental conditions at the single subject level.”

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

Hvorfor?



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



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

Unngå at data blir borte

Forstå dine data, også mange år fra nå.

De viktigste elementene:

- Datalagring
- Strukturering av filer og mapper
- Navngivning av filer og mapper
- Dokumentasjon av data : ReadMe og metadata
- Filformat for arkivering

Tenk strukturering og dokumentasjon **tidlig**.

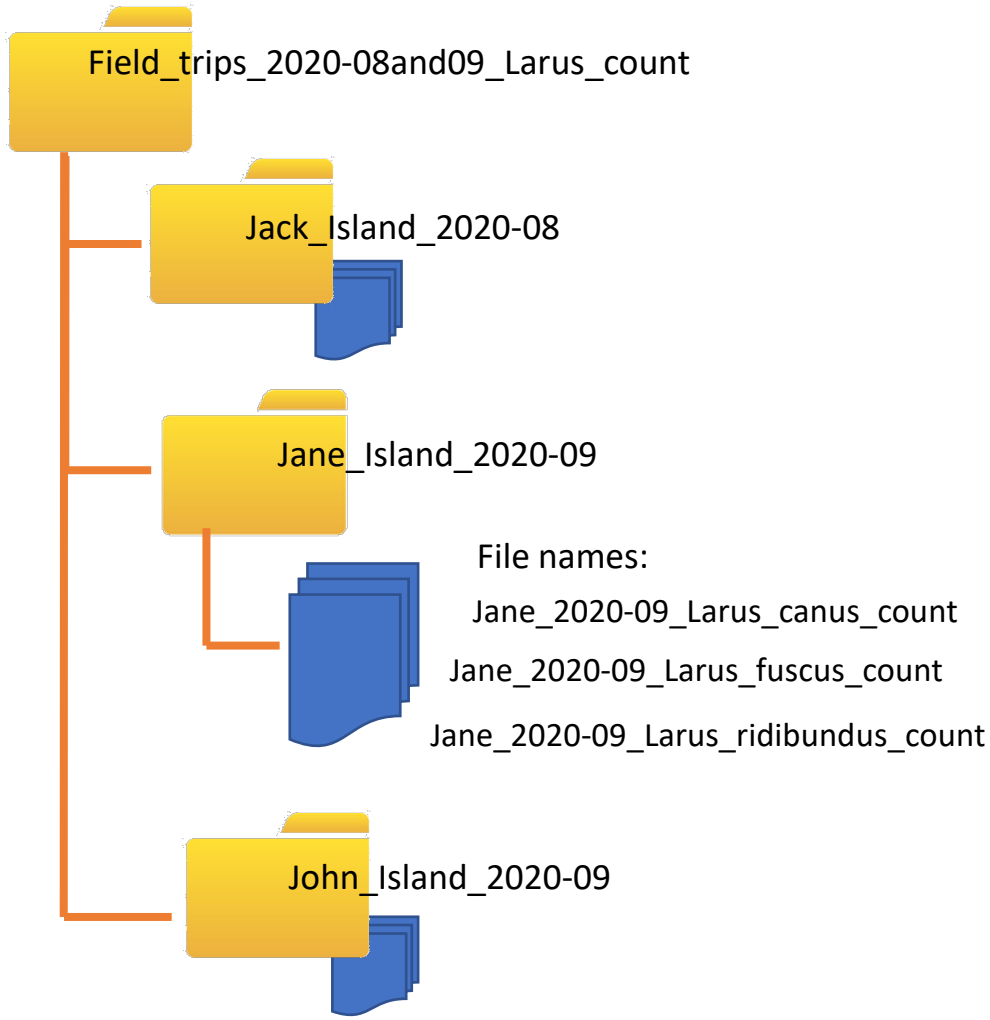
How it started …….. How it’s going

Images created with DALL*E

Mappe- og filnavn

- Mapper kan være nyttige for å holde orden og strukturere dataene
  - Især om du har mange filer
- Bruk samme navnestruktur på alle mappene
- La mappstrukturen gjenspeiles i filnavnene
  - Det reduserer faren for å forveksle filene
  - Det blir lettere å beholde orden når du senere skal arkivere dataene

- Husk å dokumentere struktur og navnesyntaks i ReadMe-filen



Navn på filer (og mapper)

- Maskinlesbar
  - Ikke bruk mellomrom
  - Unngå spesialtegn:
    “ / \ : * . ? ‘ < > [ ] ( ) & $ æÆ øØ åÅ
- Menneskelesbar
  - Bruk konsistente filnavn
  - Deskriptive, men korte (< 25 tegn)

- Internasjonalt format på datoer: ÅÅÅÅ-MM-DD

- Bruk filnavn til sortering

✗

Joe’s notes from today.txt

Tromsø&Ålesund.txt

Mynotes_versjon1.txt

Mine notater ny.txt

figure 1.png

Thesis_DONTdelete_new_draft_final*.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

Inspirert av Jenny Brian’s slides om filnavn

Eksempler på sortering med filnavn

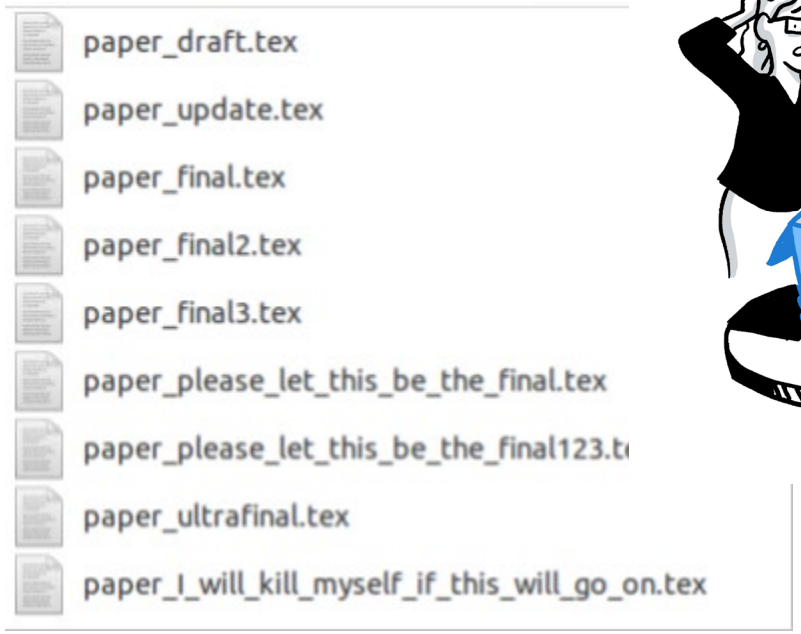
Sortert etter dato:
2020-08-01_notes_John.pdf
2020-08-31_observations_John.txt
2020-09-01_notes_Jane.pdf
2020-09-30_observations_Jane.pdf

Sortert etter tema:
Jane_notes_2020-09-01.pdf
Jane_observations_2020-09-30.txt
John_notes_2020-08-01.pdf
John_observations_2020-08-31.txt

Sortert etter 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

Tvungen sortering med nummerering:
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

Bruk navn som versjonskontroll



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

Bruk heller:

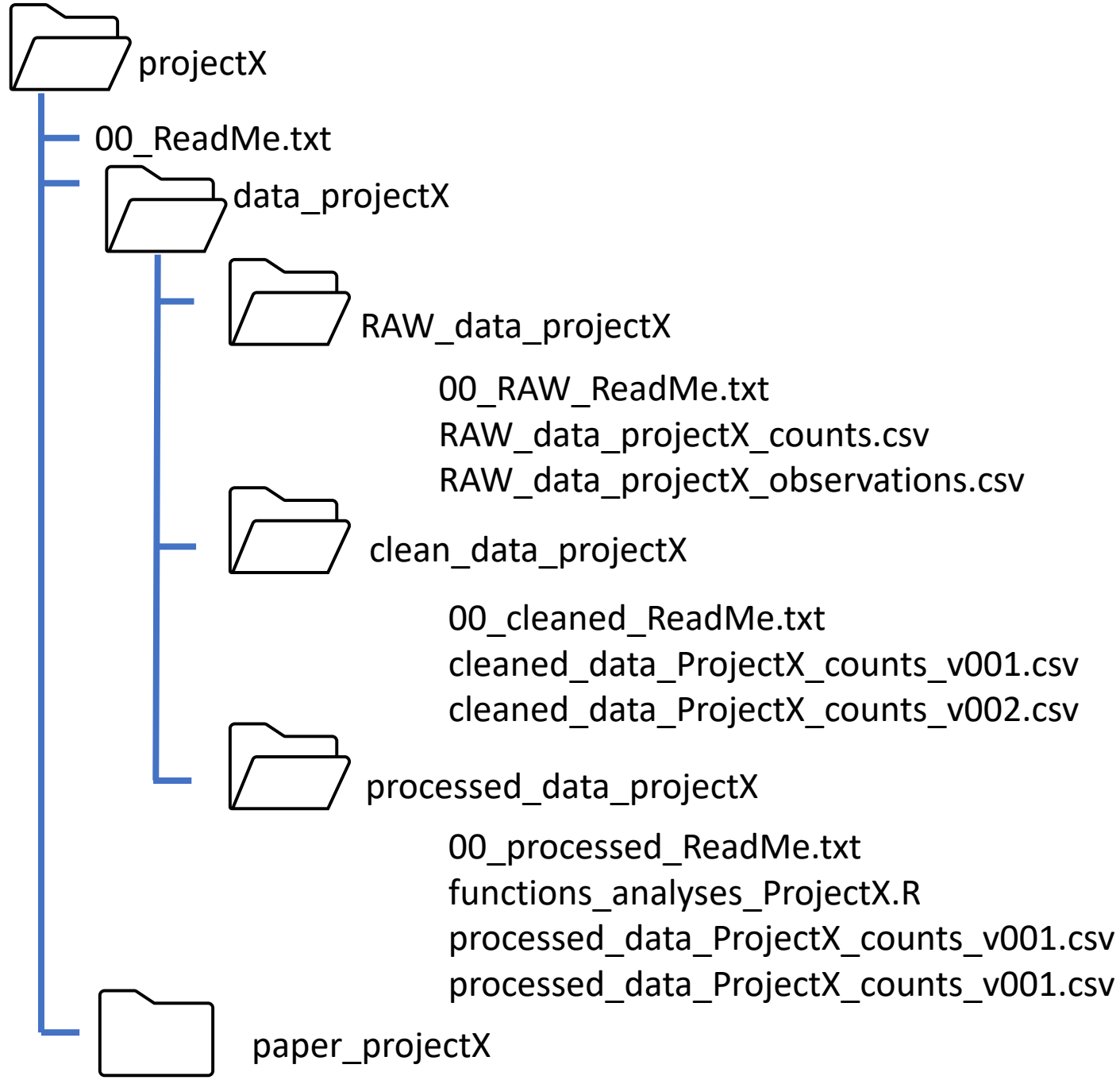
Larus_canus_counts_RAW

Larus_canus_counts_v001

Larus_canus_counts_v002

osv

Husk å dokumentere versjonsendringer i en ReadMe-fil


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

Dokumentere dataene

Hvorfor?

Gjør dataene gjenfinnbare og forståelige

- Dette gjelder deg selv – mange år inn i fremtiden
- Og andre som skal gjenbruke dataene dine

Dokumentere dataene

Hvordan?

ReadMe-fil = en beskrivende manual over dataene dine

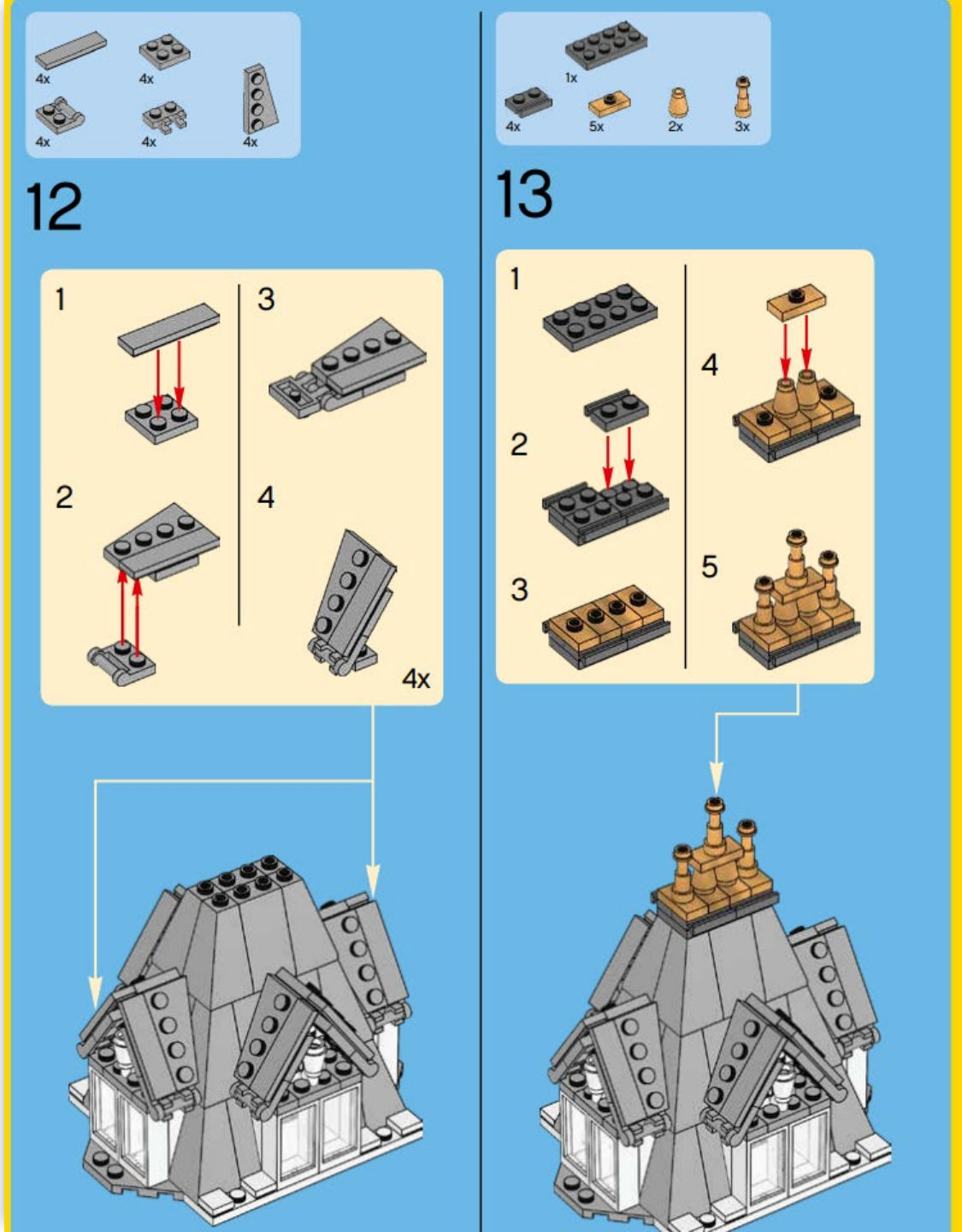
- Menneskelesbar – viktig for å sikre rett tolkning av dataene

Metadata = informasjon om dataene dine

- Strukturert og maskinlesbar informasjon – bidrar til presis gjenfinning av dataene

Når?

Gjennom hele livssyklussen, start tidlig



ReadMe-fil: En veiledning til dine data

- Tilstrekkelig informasjon for rett tolkning av datasettet
- Et veikart for fremtidige brukere som informerer om hvordan dataene ble generert, modifisert, prosessert og hvordan de kan gjenbrukes.
- Begynn å dokumentere tidlig, oppdater kontinuerlig i et åpent format (f.eks. .txt-fil)

DataverseNO sin mal: https://site.uit.no/dataverseno/deposit/prepare/#readmefile

ReadMe-fil: En veiledning til dine data

- Tilstrekkelig informasjon for rett tolkning av datasettet

Generell bakgrunnsinformasjon

tittel, DOI, kontaktinfo, dato, sted, eierskap, finansiør

Metodebeskrivelser

protokoller, instrumenter, programvare, prosessering, analysering, osv

Filoversikt

Oversikt over filene – hva finnes i de ulike filene, informasjon om oppdateringer (versjoner)

Filspesifikk informasjon

forklaring av kolonneoverskrifter, forkortelser, måleenheter, tolkning av datasettet

Referanse og vilkår for gjenbruk

lisens

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 om data

Eksempelvis:

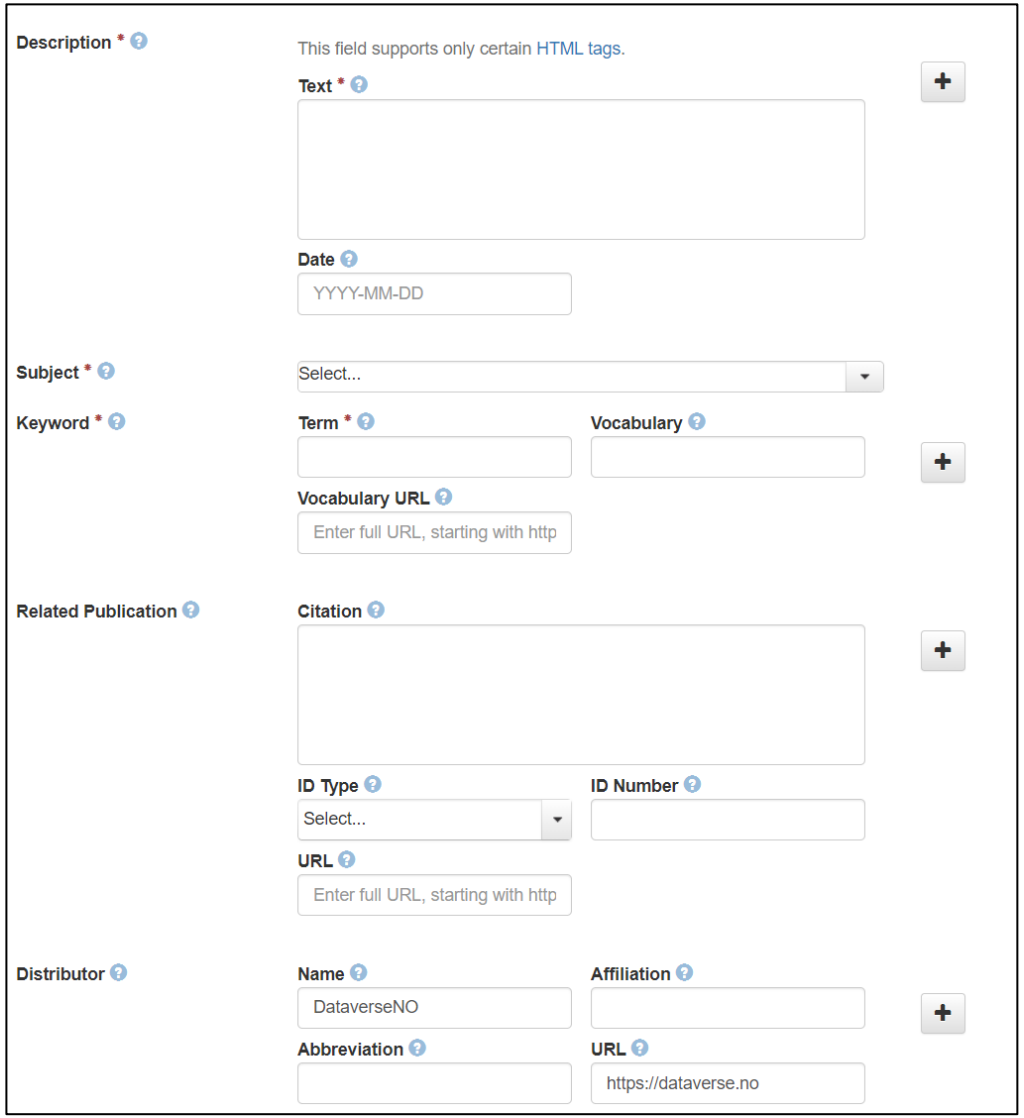
- Forfatter, tittel, beskrivelse, …
- Emneord
- Geografisk informasjon
- Innsamlingsdato
- Språk

Generiske standarder

- F.eks. internasjonalt datoformat (e.g. ISO-8601): YYYY-MM-DD (2021-09-11)
- Eksempler: Dublin Core, Data Documentation Initiative

Fagspesifikke standarder

- F.eks. Darwin Core = en standard for beskrivelse av biologisk diversitet.



Metadata : data om data

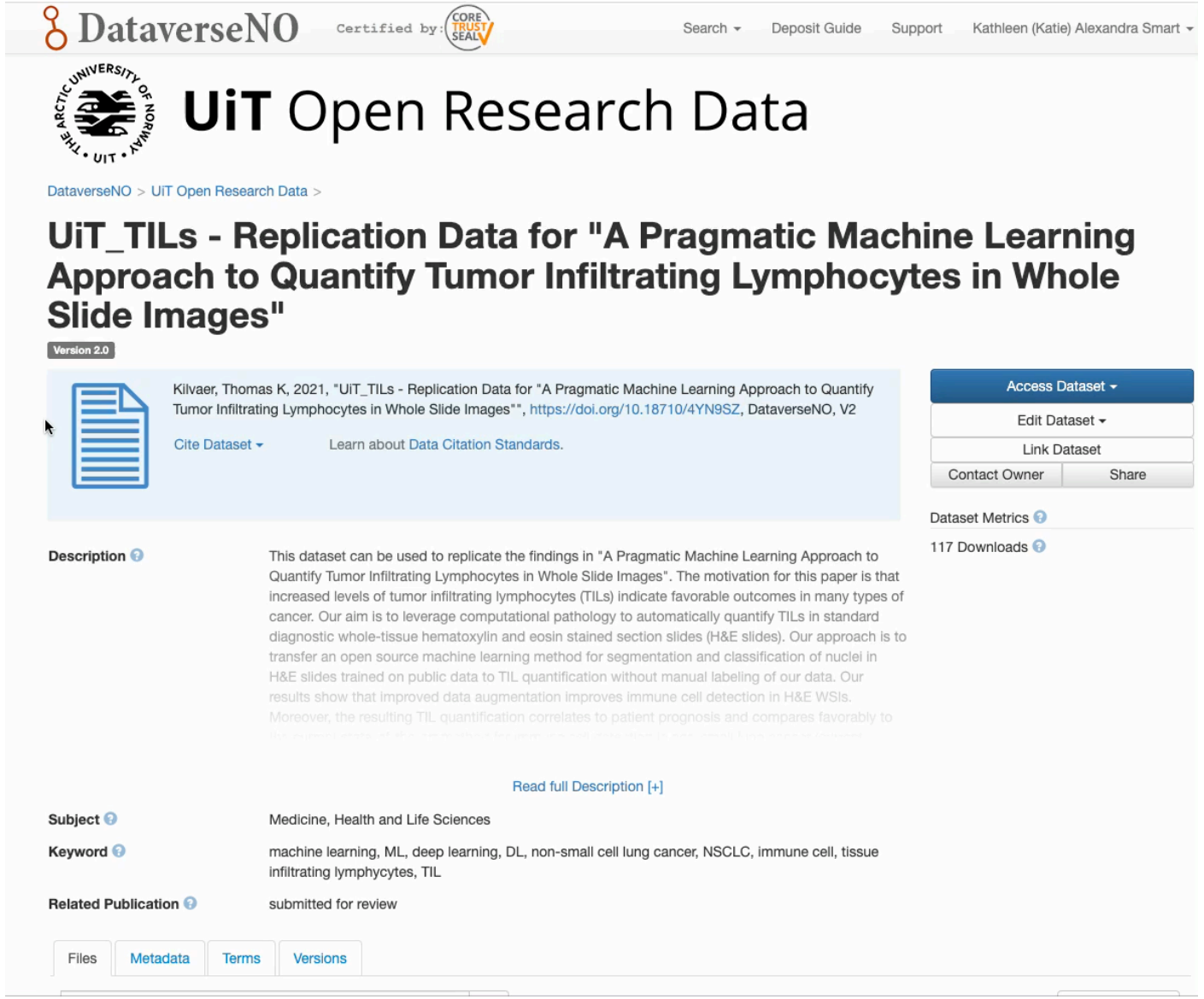
Oversikter over ulike standarder

[Research Data Alliance]

[FAIRSharing.org]

[Digital Curation Centre]

Tips: Ta en kikk på metadataskjemaet som brukes i arkivet du planlegger å bruke for dine data.



[Metadata example from DataverseNO (link)]

Dokumentasjon i den aktive fase: ELN

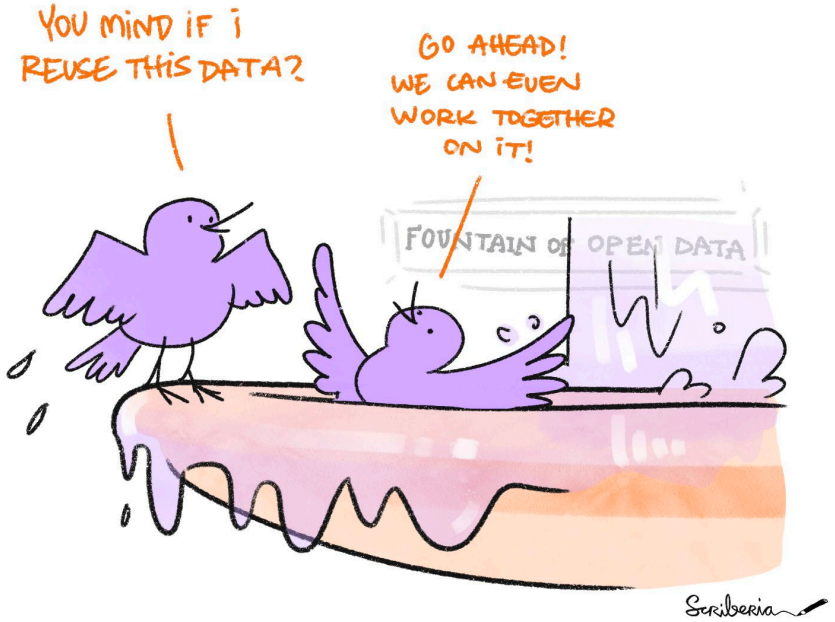
Elektronisk lab notatbok (ELN)

- Et elektronisk verktøy som kan erstatte papirbaserte labnotatbøker
  - Notater, protokoller, maler, office filer
  - Inventory: hold styr på prøver og labinventar
- Sikrere dataforvaltning og effektivisering av arbeidsflyt

- Mer info på Forskningsdataportalen > Behandling og lagring > Elektronisk labnotatbok

Klargjøring for arkivering

- Utvalg
  - Viktig: Ikke ta bort negative data – dvs. data som ikke støtter hypotesen man tester
  - Ta med råversjon og behandlede data
- Ev. anonymisering og/eller aggregering
  - spesielt for persondata / sensitive data
- Både originalfiler og i arkivverdig filformat
- Kurs i morgen - Arkivering
  - Et kurs om arkivering generelt
  - Et eget kurs spesifikt om UiTs arkiv: DataverseNO



Arkivverdige filformat

Kjennetegn på arkivverdige filformat:

- ikke-proprietære
- åpne, og følger dokumenterte, internasjonale standarder
- bruker standard tegnkoding (f.eks. ASCII, UTF-8)
- er ikke komprimerte

Arkivering: Arkivverdige filformat

Filtype	Føretrekte filformat (døme)	Ikkje-føretrekte filformat (døme)
Lyd	→ Ukomprimert utan tap: Wav or AIFF (.wav/.aiff) → Komprimert utan tap: FLAC (.flac) → Komprimert med tap: Mp3 (.mp3)	→ AAC (.m4a) → Monkey's Audio (.ape) → Ogg Vorbis (.ogg) → Windows Media Audio (.wma)
Tekst (lysbilete, illustrasjonar)	→ PDF/A (.pdf) saman med originalfil	→ PowerPoint (.pptx)
Tekst (tabellar)	→ Tabulatorseparert rein tekstfil i Unicode (.txt)	→ Excel (.xlsx)
Tekst (tekst)	→ Rein tekst (.txt) **Dersom formattering/struktur trengst:** → XML, PDF/A (.pdf) saman med originalfil	→ Word (.docx)
Markeringsspråk	→ XML (.xml) → HTML (.html) → Relaterte filer: .css, .xslt, .js, .es	→ SGML (.sgml) → Markdown (.md)

Mer informasjon i arkiveringsguiden til DataverseNO

UiT The Arctic University of Norway

UiT The Arctic University of Norway

Contact Share

Looking for **TROLLing**? Click here: **https://trolling.uit.no/**

NMDC Node UiT

PolarFront

The Stein Rokkan Research Group for Quantitative Social and Political Science

Tromsø Geophysical Observatory

Search this dataverse...

Advanced Search

- Dataverses (7)
- Datasets (906)
- Files (8,616)

Dataverse Category

Research Project (4)

Department (1)

1 to 10 of 913 Results

Sort

Data of the i-MASTER project: A novel initiative in maritime education and training experience

Sep 20, 2023

i-MASTER consortium, 2023, "Data of the i-MASTER project: A novel initiative in maritime education and training experience", https://doi.org/10.18710/T1VQLA, DataverseNO, V1

i-MASTER datasets. The i-MASTER project is an EU funded project under grant agreement No. 101060107. The project's objective is to study and

Mer info og hjelp

UiT Forskningsdataportalen :

- Tips om forskningsdatahåndtering
- Oversikt over kurs

Email:

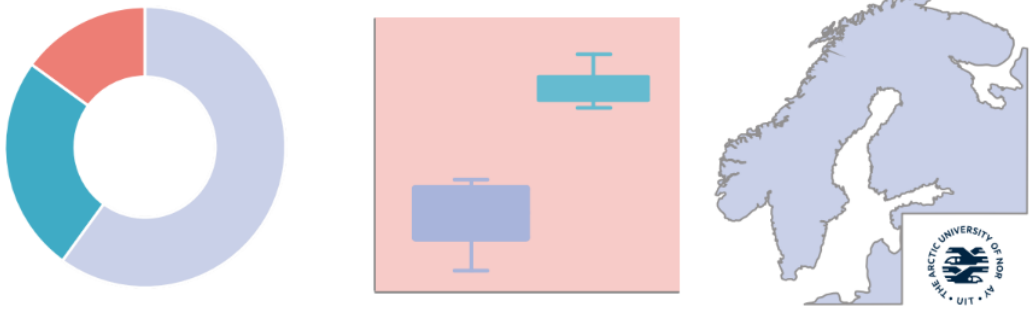
researchdata@hjelp.uit.no

Data visualisering

Mer informasjon og påmelding via Tavla

FAIR Data Visualisation @ *UiT*

Realise the potential of your research data



The University Library and the Research Software Engineering group are pleased to offer a 4-day workshop on data visualisation. Topics will include:

- FAIR data and Research Data Management
- Principles of data visualisation
- Charts, attributes, and colour
- Oral presentation skills
- Practical introduction to using Python and Jupyter notebooks

Participants will make and present visualisations of their own research data in a constructive feedback session. This workshop is intended for graduate students and early-career researchers who have quantitative data.

Register on Tavla or contact:

Katie Smart
The University Library (UB)
kathleen.a.smart@uit.no

Radovan Bast
Research Software Engineering (IT)
radovan.bast@uit.no

Practical information:
In-person at UB and Teorifagbygget. Optional joint afternoon working sessions.
October 17 09:00 - 12:00 lecture& practical
October 18 09:00 - 12:00 lecture& practical
October 19 09:00 - 14:00 lecture& practical
October 23 12:00 - 15:00 feedback session

Space is limited to 20 participants.

Datarøkter-nettverket ved UiT

(UiT Data Stewards Network)

Hva er en datarøkter?

Datarøkter iflg. Bing Image Creator

Dataansvarlig innenfor en forskningsgruppe, avdeling eller på instituttnivå.

Sikrer at data samles inn og håndteres i tråd med beste praksis.

Kan omfatte håndtering av fysiske prøver, arkiver eller samlinger.

Nettverket:

- Skape et aktivt fellesskap for nettverksbygging og faglig utvikling.
- Kunnskapsdeling og samarbeid på tvers av fagområder.

Mer info: researchdata@hjelp.uit.no

Evaluering

Vi jobber kontinuerlig med å forbedre innholdet i webinarene våre. En tilbakemelding fra deg vil være til god hjelp.

Her: skjema.uio.no/ubevalno

Dato: XX

Emnekode: Forskningsdata

researchdata@hjelp.uit.no

Noortje Haugstvedt

Leif Longva