Starting soon...

NorDataNet
Norwegian Scientific Data Network

Practical information

- 15 minutes breaks between sessions
- Remember to mute when listening
- Use chat actively during sessions
- Use “raise hand” functionality during discussions or to indicate wish to ask questions
- No recording of sessions are planned
- Working groups are organised using the breakout session functionality in Zoom
  - Participants are allocated to sessions by host
- Supposed to be an interactive course!
- Picture of all participants wanted, will start with this

NorDataNet
Norwegian Scientific Data Network

Motivation:
Why do we need data management?

Øystein Godøy

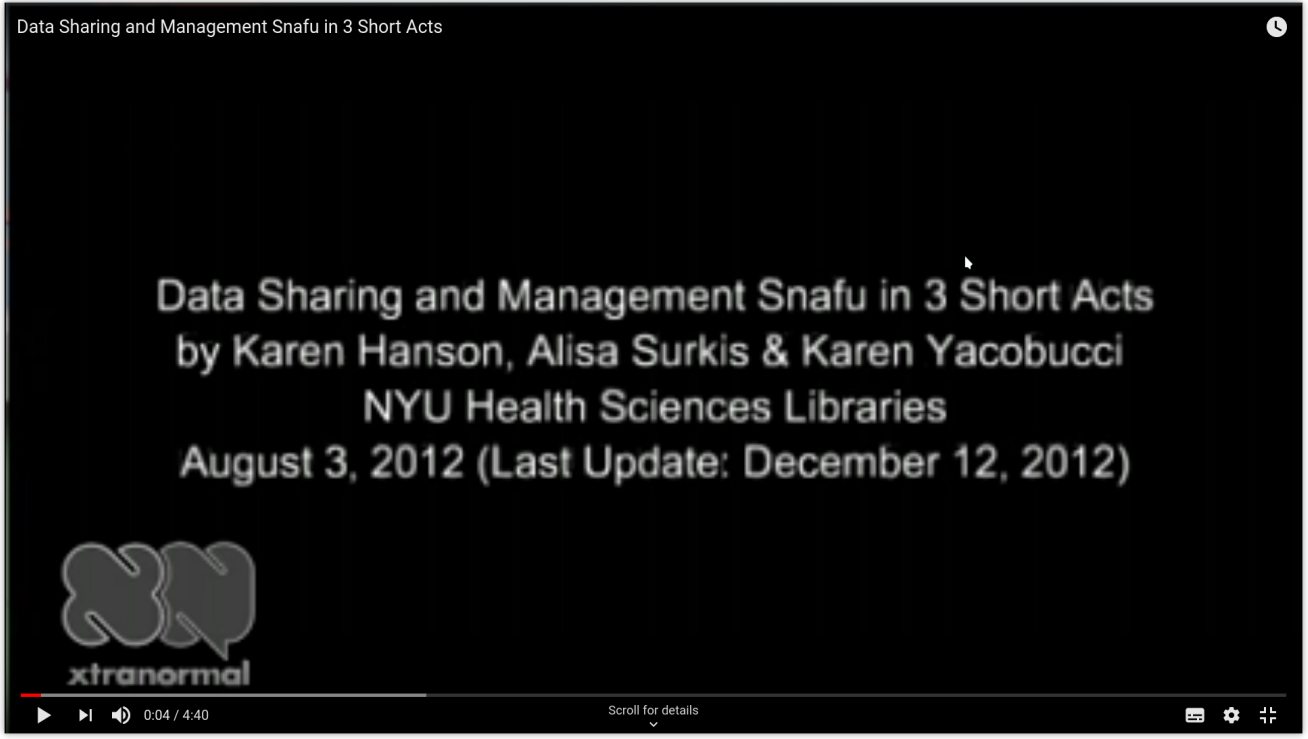
NorDataNet
Norwegian Scientific Data Network

Outline

- Data Sharing and Management Snafu in 3 Short Acts
  - https://www.youtube.com/watch?v=N2zK3sAtr-4
- Why do we need data management?
- Science life cycle/Data life cycle
- How to change data sharing culture.
- What are the FAIR data principles?
  - How do they help with good data management?
- External boundary conditions by funding agencies and publishers.
- Scientific data as service.
- Data management plan.

NorDataNet
Norwegian Scientific Data Network

Data Sharing and Management Snafu in 3 Short Acts

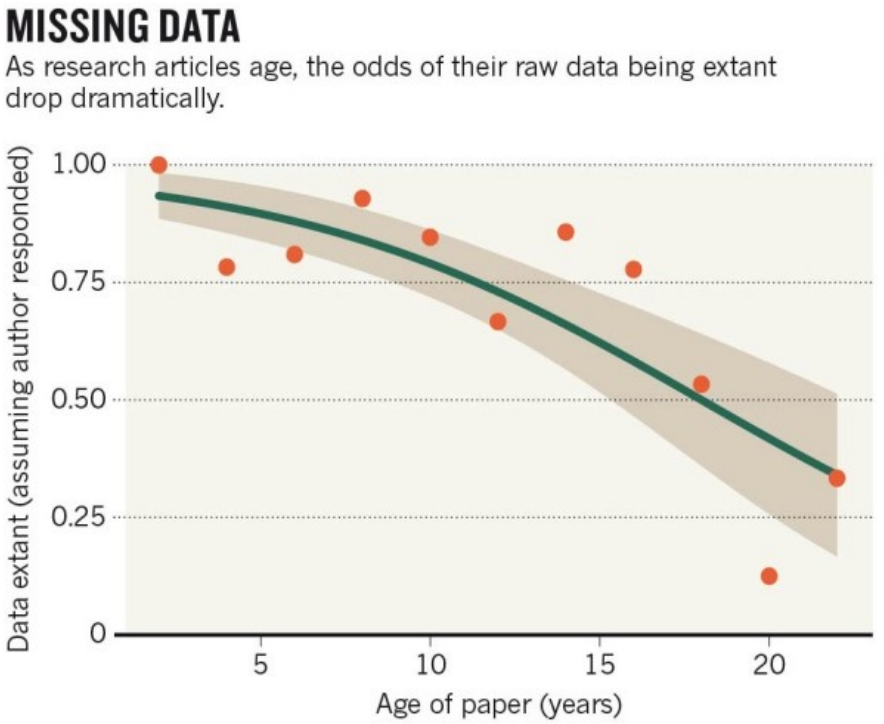


https://www.youtube.com/watch?v=N2zK3sAtr-4

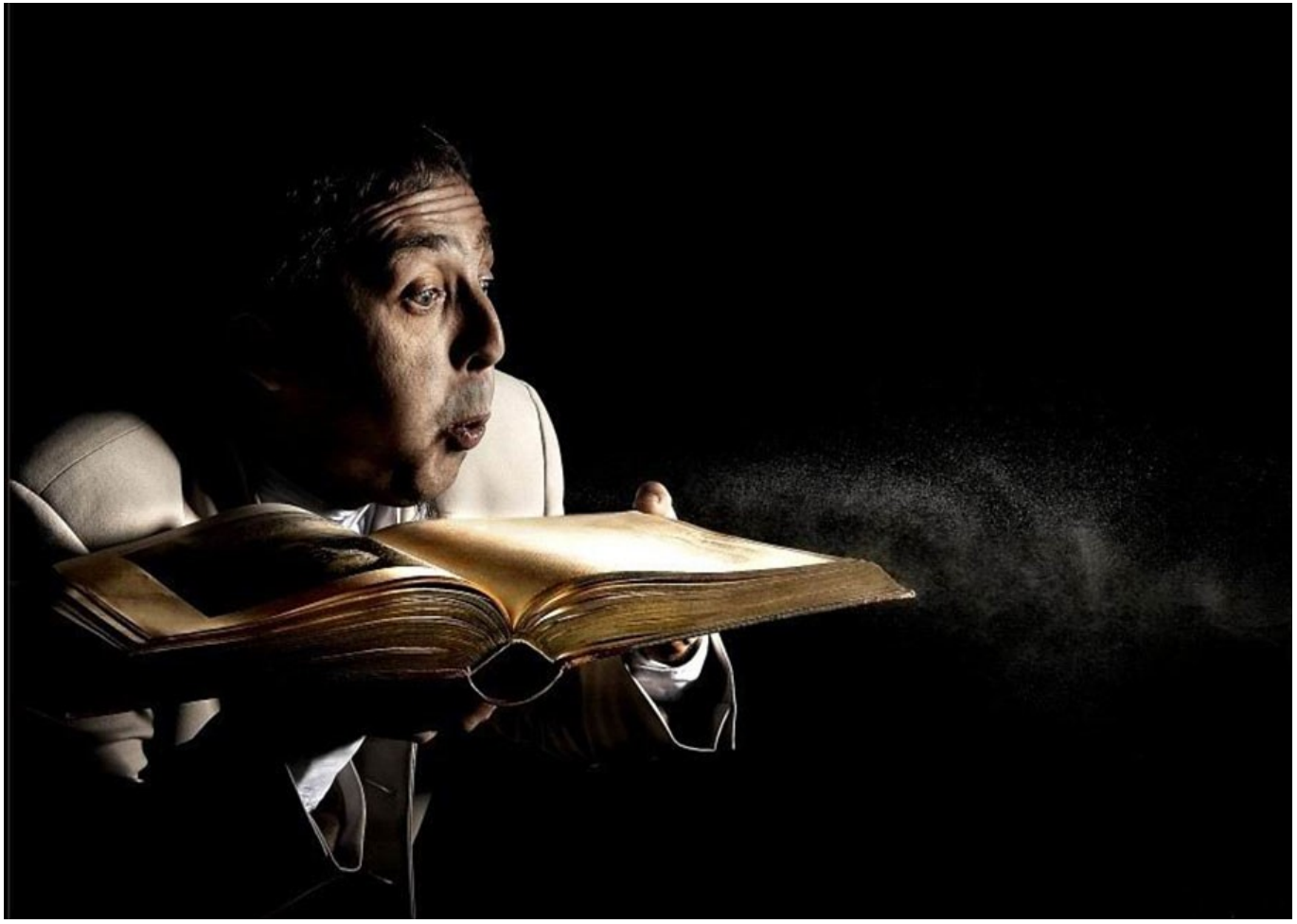
NorDataNet
Norwegian Scientific Data Network

Why do we need data management?

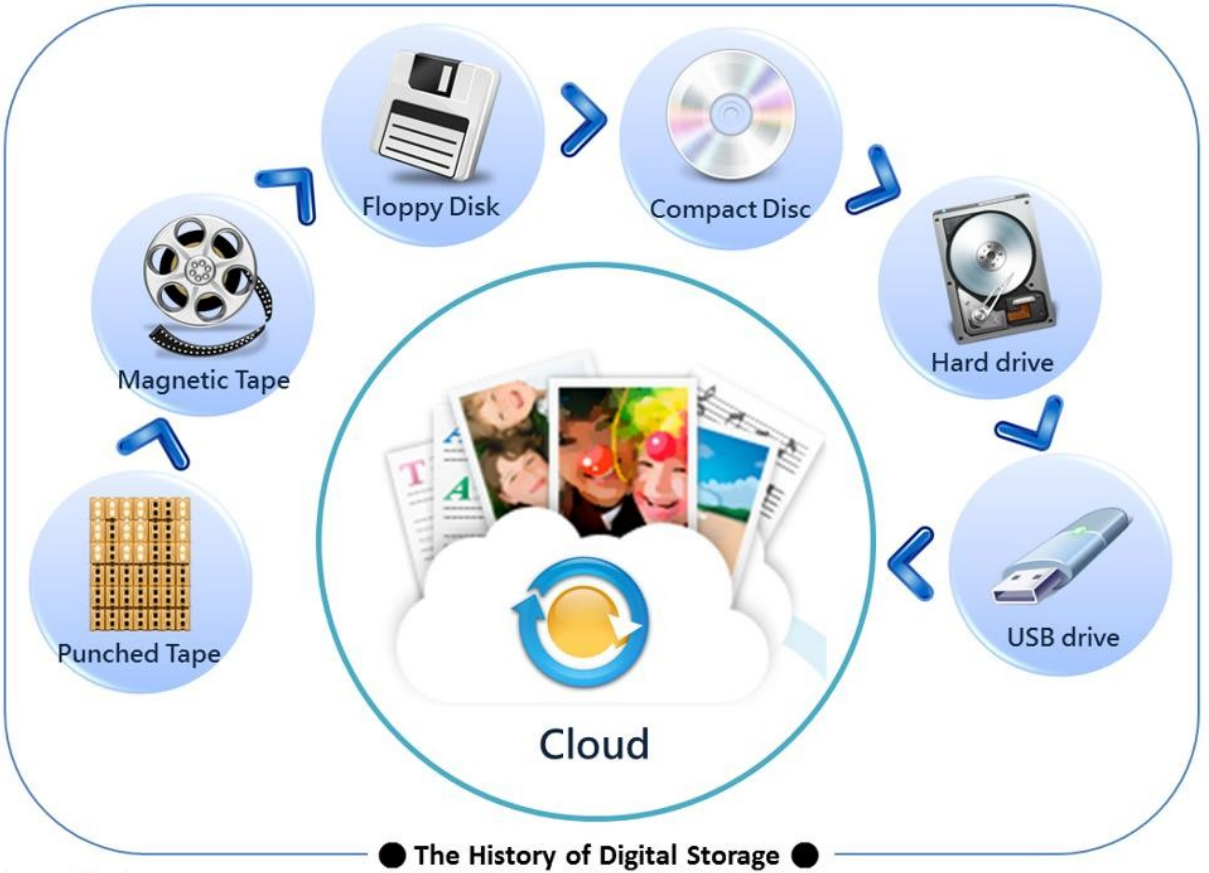
- Loosing scientific data
  - Decline can mean 80% of data are unavailable after 20 years.
    - Gibney and Van Noorden (2013), Nature



NorDataNet
Norwegian Scientific Data Network



NorDataNet
Norwegian Scientific Data Network



NorDataNet
Norwegian Scientific Data Network

Why bother with structured data management?

- Benefits
  - Maximise public investment in data collection and production
  - Promote scientific collaboration
  - Promote interdisciplinary science
  - Promote scientific transparency
  - Leave a legacy
- Science paradigms
  - according to Jim Gray
  - empirical science
  - theoretical science
  - computational science
  - data exploration science

NorDataNet
Norwegian Scientific Data Network

Why share data?

- Research sponsor require it
  - recognition as an authoritative source and wise investment
- Quality control
  - improved data quality due to expanded use, field checks, and feedback
- Improved visibility
  - improved connections to scientific network, peers, and potential collaborators
- Journals require it
  - Reproducible research
- Far upstream sponsors require it

NorDataNet
Norwegian Scientific Data Network

Making Your Research Easier and Cheaper

The 5 P's matter!

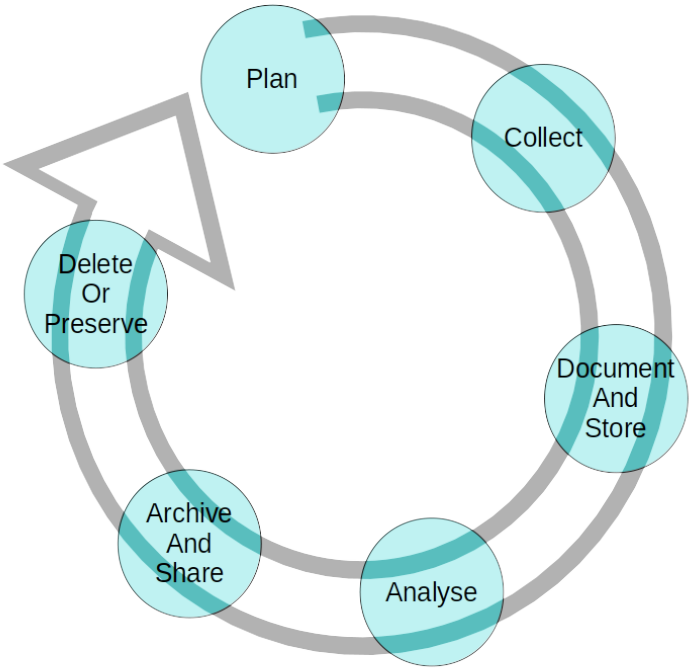
Prior
Planning
Prevents
Poor
Performance!

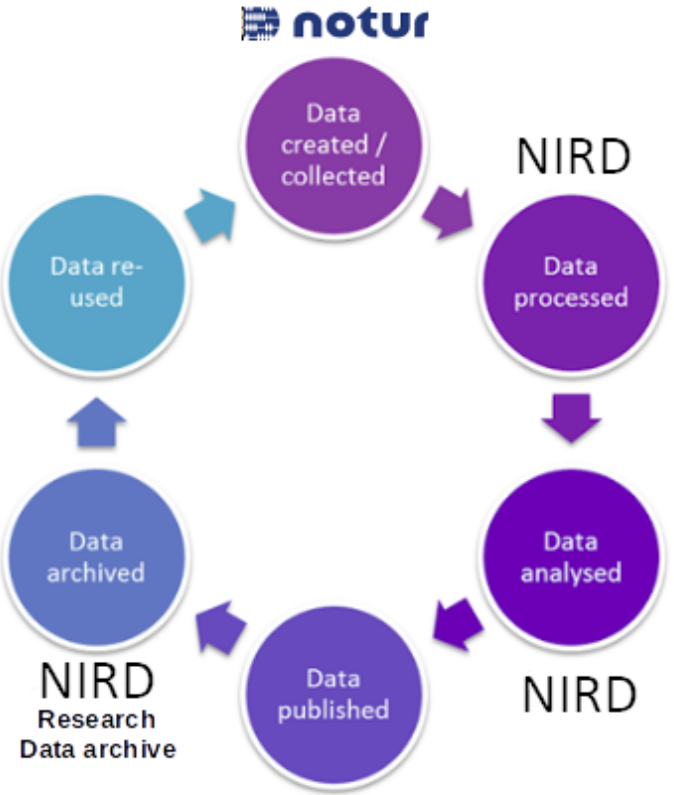
NorDataNet
Norwegian Scientific Data Network

Ruth Duerr, NSIDC, 2013 ESIP Material

ESIP

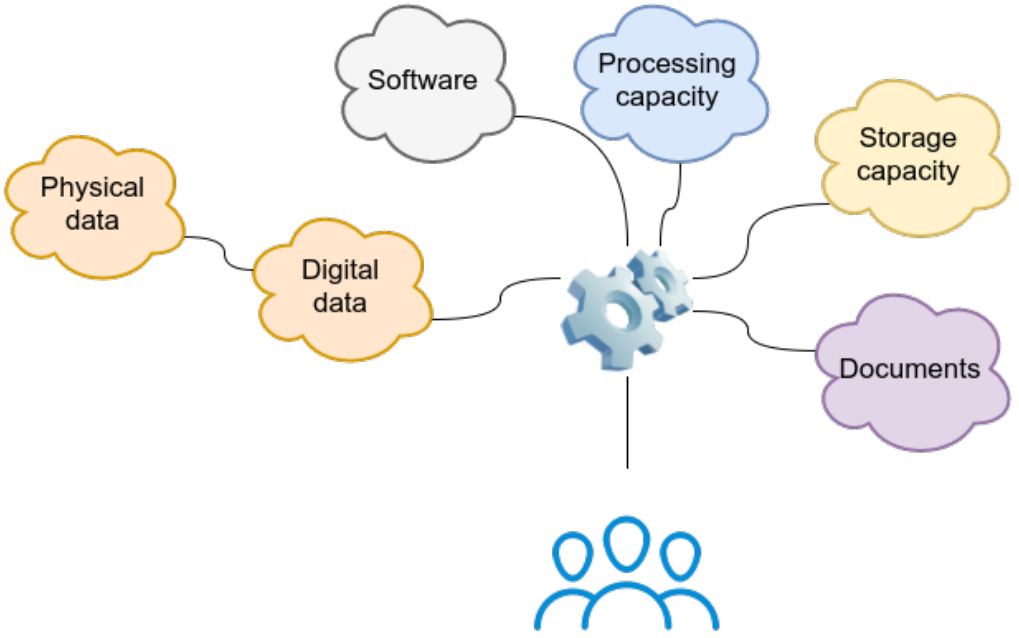
Science life cycle/Data life cycle





NorDataNet
Norwegian Scientific Data Network

https://codemeta.github.io/

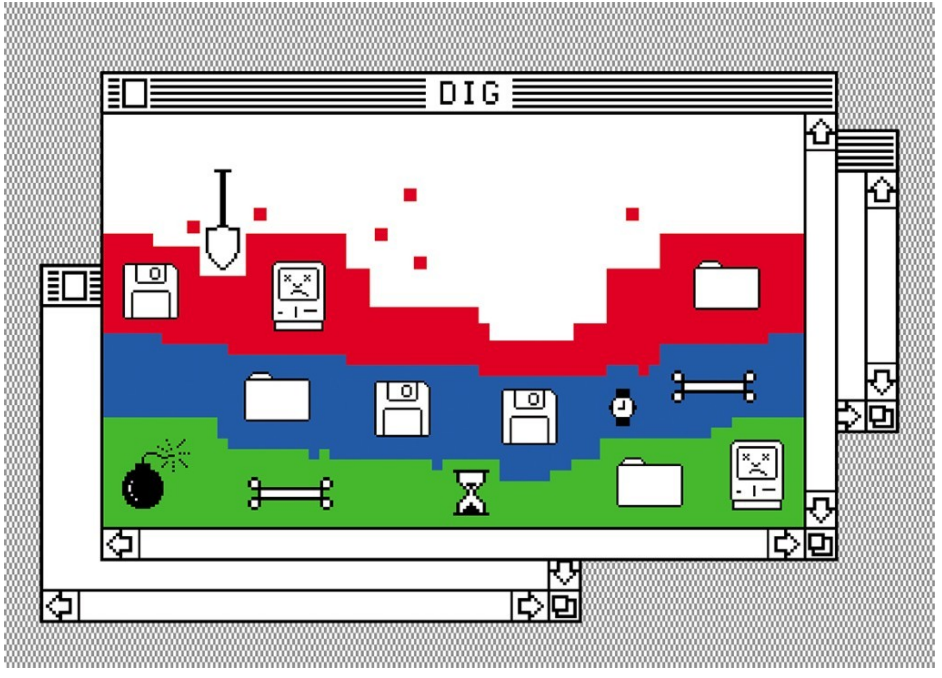


NorDataNet
Norwegian Scientific Data Network

NorDataNet
Norwegian Scientific Data Network

Challenge to scientists: does your ten-year-old code still run?

https://www.nature.com/articles/d41586-020-02462-7



NorDataNet
Norwegian Scientific Data Network

Reproducibility checklist (1)

- **Code** Workflows based on point-and-click interfaces, such as Excel, are not reproducible. Enshrine your computations and data manipulation in code.
- **Document** Use comments, computational notebooks and README files to explain how your code works, and to define the expected parameters and the computational environment required.
- **Record** Make a note of key parameters, such as the ‘seed’ values used to start a random-number generator. Such records allow you to reproduce runs, track down bugs and follow up on unexpected results.
- **Test** Create a suite of test functions. Use positive and negative control data sets to ensure you get the expected results, and run those tests throughout development to squash bugs as they arise.
- **Guide** Create a master script (for example, a ‘run.sh’ file) that downloads required data sets and variables, executes your workflow and provides an obvious entry point to the code.
- **Archive** GitHub is a popular but impermanent online repository. Archiving services such as Zenodo, Figshare and Software Heritage promise long-term stability.
- **Track** Use version-control tools such as Git to record your project’s history. Note which version you used to create each result.
- **Package** Create ready-to-use computational environments using containerization tools (for example, Docker, Singularity), web services (Code Ocean, Gigantum, Binder) or virtual-environment managers (Conda).
- **Automate** Use continuous-integration services (for example, Travis CI) to automatically test your code over time, and in various computational environments.
- **Simplify** Avoid niche or hard-to-install third-party code libraries that can complicate reuse.
- **Verify** Check your code’s portability by running it in a range of computing environments.

NorDataNet
Norwegian Scientific Data Network

https://www.nature.com/articles/d41586-020-02462-7

Reproducibility checklist (2)

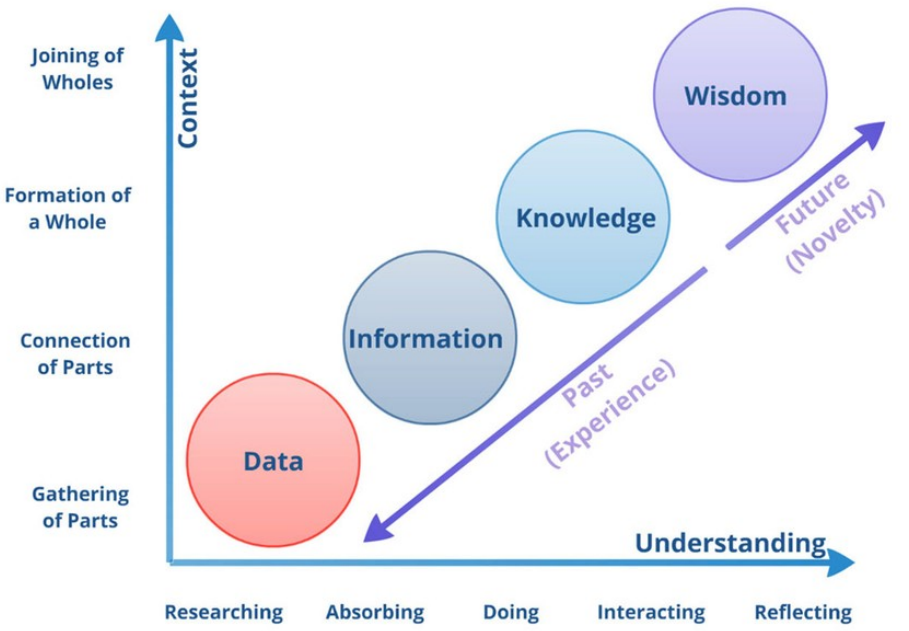
- **Code** Workflows based on point-and-click interfaces, such as Excel, are not reproducible. Enshrine your computations and data manipulation in code.
- **Document** Use comments, computational notebooks and README files to explain how your code works, and to define the expected parameters and the computational environment required.
- **Record** Make a note of key parameters, such as the 'seed' values used to start a random-number generator. Such records allow you to reproduce runs, track down bugs and follow up on unexpected results.
- **Guide** Create a master script (for example, a 'run.sh' file) that downloads required data sets and variables, executes your workflow and provides an obvious entry point to the code.
- **Archive** GitHub is a popular but impermanent online repository. Archiving services such as Zenodo, Figshare and Software Heritage promise long-term stability.
- **Track** Use version-control tools such as Git to record your project's history. Note which version you used to create each result.
- **Simplify** Avoid niche or hard-to-install third-party code libraries that can complicate reuse.
- **Verify** Check your code's portability by running it in a range of computing environments.
- **Test** Create a suite of test functions. Use positive and negative control data sets to ensure you get the expected results, and run those tests throughout development to squash bugs as they arise.
- **Package** Create ready-to-use computational environments using containerization tools (for example, Docker, Singularity), web services (Code Ocean, Gigantum, Binder) or virtual-environment managers (Conda).
- **Automate** Use continuous-integration services (for example, Travis CI) to automatically test your code over time, and in various computational environments.

NorDataNet
Norwegian Scientific Data Network

https://www.nature.com/articles/d41586-020-02462-7

DIKW chain

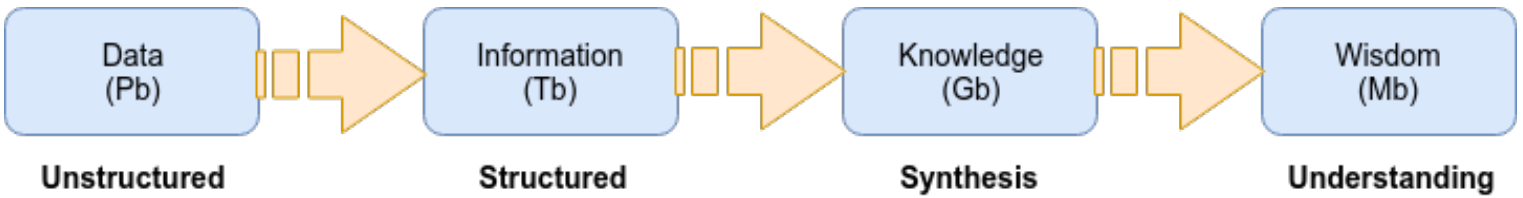
- How to transition from data to knowledge and understanding...
  - The illustration is a common redrawing of Russ Ackoff “From Data to Wisdom”
    - Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9
- DIKW is necessary to
  - Take care of data for the future
  - Ensure data is the basis for knowledge
    - Now and in the future
  - Knowledge based management depends on national, regional and global interaction



http://www.easterbrook.ca/steve/2012/09/what-is-climate-informatics/

NorDataNet
Norwegian Scientific Data Network

DIKW chain



NorDataNet
Norwegian Scientific Data Network

The reality today

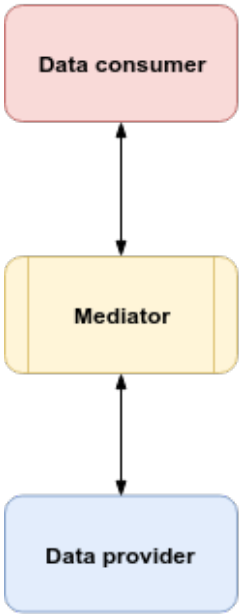
Recreated from Klump et al. 2006

NorDataNet
Norwegian Scientific Data Network

How to change data sharing culture?

NorDataNet
Norwegian Scientific Data Network

Bridging between data management actors



Unwilling

- Do not want to change behaviour, existing tools have worked well.
- Want to continue as before.
- Does not see the benefit of standardisation, until explicitly explained/demonstrated or through new

Willing

- Wants to translate between provider and consumer.
- Still relies on some sort of standardisation in order to be cost effective.
- Must know dimensions, structures, content, missing values, units, aggregation levels, ...

Unwilling and skeptical to potential users

- Do not want to change behaviour, legacy system(s).
- Want to continue as before.
- Understands own requirements (knows the data well).

- A prerequisite for efficient data sharing across communities is application of proper metadata and standards
- Although standards exist, they are often not used by data providers who doesn't see the benefit
- Lacking understanding for the importance of use metadata
  - Enabling reuse across communities and generations
  - Lacking understanding for the importance of semantic standardisation
- Need a business model crediting all involved parties
  - Scientists, institutions, data centres, ….
- It is about leaving a legacy

NorDataNet
Norwegian Scientific Data Network

The FAIR Guiding Principles for scientific data management and stewardship

- To be **Findable**:
  - F1. (meta)data are assigned a *globally unique and persistent identifier*
  - F2. data are described with *rich metadata* (defined by R1 below)
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are *registered or indexed in a searchable resource*
- To be **Accessible**:
  - A1. (meta)data are retrievable by their identifier using a *standardized communications protocol*
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  - A2. metadata are accessible, even when the data are no longer available
- To be **Interoperable**:
  - I1. (meta)data *use a formal, accessible, shared, and broadly applicable language for knowledge representation*
  - I2. (meta)data *use vocabularies* that follow FAIR principles
  - I3. (meta)data *include qualified references* to other (meta)data
- To be **Reusable**:
  - R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible *data usage license*
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data *meet domain-relevant community standards*

NorDataNet
Norwegian Scientific Data Network

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Scientific data as service

- From science to service
  - Goddard, Science 23 Sep 2016: Vol. 353, Issue 6306, pp. 1366-1367 DOI: 10.1126/science.aag3087

NorDataNet
Norwegian Scientific Data Network

File Edit View History Bookmarks Tools Help

https://adc.met.no/metsis/search?fulltext=synop&start_date=&end_date=&is_p

Most Visited Getting Started NorDataNet | Arctic ... XPath and XSLT with l... Parsing XML and HT... xslt.md.html DuckDuckGo — Priva... Other Bookmarks

Home Find data Support About

synop

Start Date dd / mm / yyyy

End Date dd / mm / yyyy

Has children

- Any -

Check whenever datasets is parent and have children

Search Reset

Dataset Level

Parent (450)

Collection

ADC (450)
APPL (450)
GCW (450)
SIOS (450)
YOPP (450)

Show more

Start date year

- 2013 (450)

End date

Iso Topic Category

- climatologyMeteorologyAtmosphere (450)

Keywords

- air_pressure (450)
- air_pressure_at_sea_level (450)
- air_temperature (450)
- Atmosphere > Atmospheric Pressure > Surface Pressure (450)
- Atmosphere > Atmospheric Temperature > Surface Air Temperature (450)
- Atmosphere > Atmospheric Water Vapor > Humidity (450)
- Atmosphere > Atmospheric Winds > Surface Winds (450)
- Atmosphere > Clouds > Cloud Amount/Frequency (450)
- Atmosphere > Clouds > Cloud Types (450)
- Atmosphere > Precipitation > Precipitation Amount (450)

Show more

Collections allows the user to search in subsets of the existing catalogue. The collections are primarily data management projects that have been incorporated in the ADC catalogue. The collections currently served through ADC include (datasets may belong to multiple data collections):

- ADC is the full collection of this service
- CC is the CryoClim collection
- GCW is the collection for the WMO Global Cryosphere Watch
- NMAP is the NORMAP e-infrastructure collection
- NSDN is the data collection for the Norwegian Scientific Data Network
- SIOS is the data collection for the Svalbard Integrated Arctic Earth Observing System
- YOPP is the data collection used to serve the Year of Polar Prediction

In order to search a specific data collection select that collection. If no data collection is selected all collections are searched.

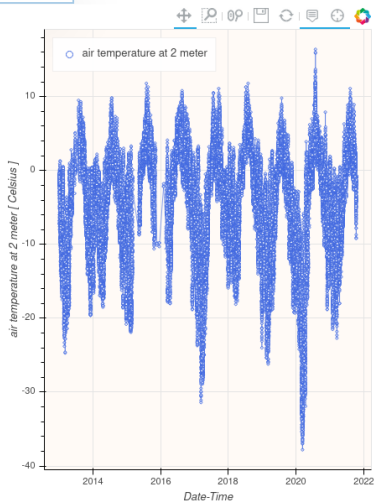
Some data collections are currently in the process of being added to the system. This includes:

- DOKI is the data collection for the Norwegian contributions to the International Polar Year.
- DAMOC is the data collection for the EU FP6 project DAMOCLES
- ACCESS is the data collection for the EU FP7 project ACCESS
- APPL is the data collection for the EU H2020 project APPLICATE
- AP is the data collection for the EU H2020 project Arctic Passion

Visualize timeseries

Back to results map

TA



Project

APPLICATE (450)
SIOS (450)
YOPP (450)

Organisation

- Norwegian Meteorological Institute (450)

Publisher

Data Center

- WMO (450)

450 datasets found. Showing datasets 1 - 15 on page 1 of 30 pages.

SYNOP data from station VERLEGENHUKEN

WMO Year of Polar Prediction, Svalbard Integrated Arctic Earth Observing System (YOPP, APPLICATE, SIOS)

NorDataNet
Norwegian Scientific Data Network

Data management plans

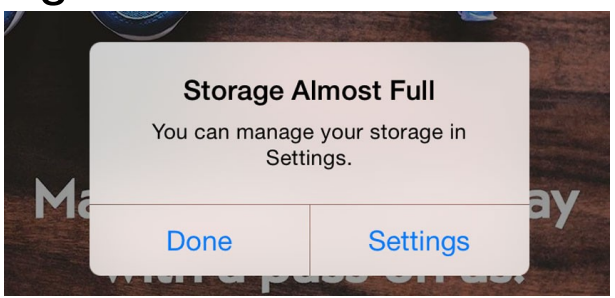
A data management plan (DMP) is a written document that describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

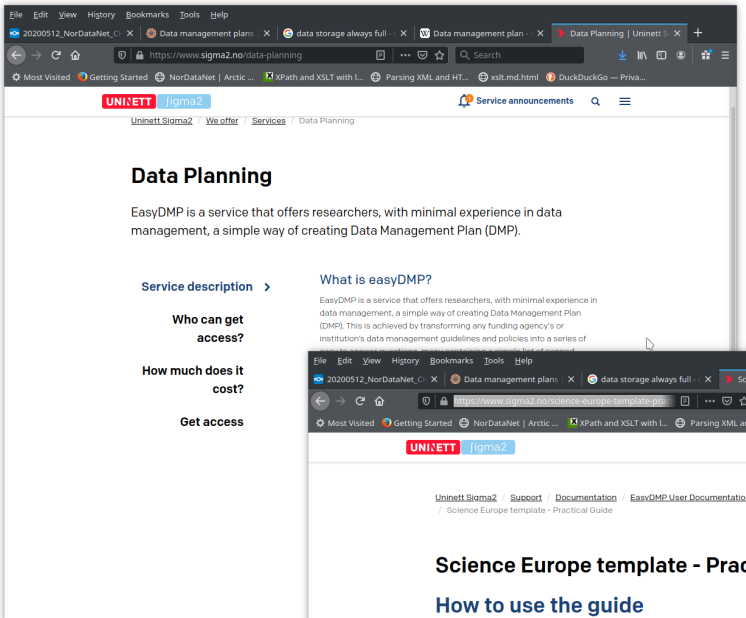
NorDataNet
Norwegian Scientific Data Network

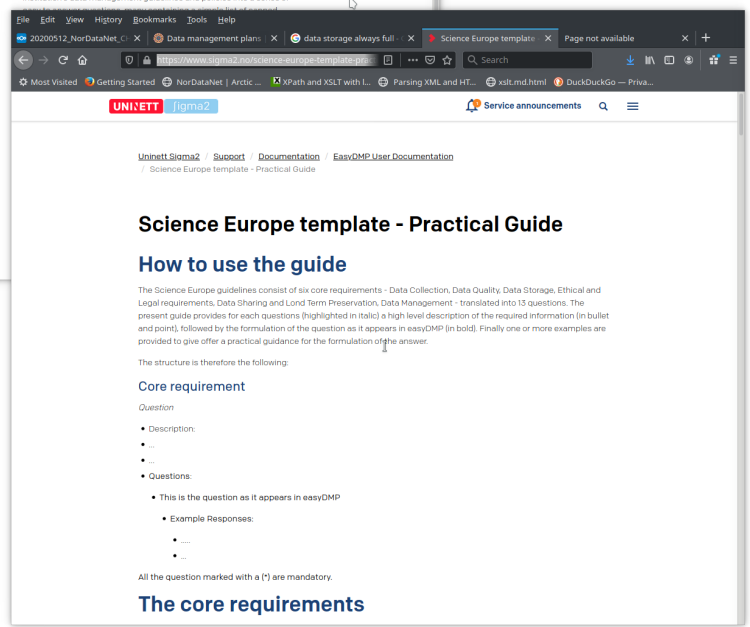
Stanford Libraries definition

Data management plans

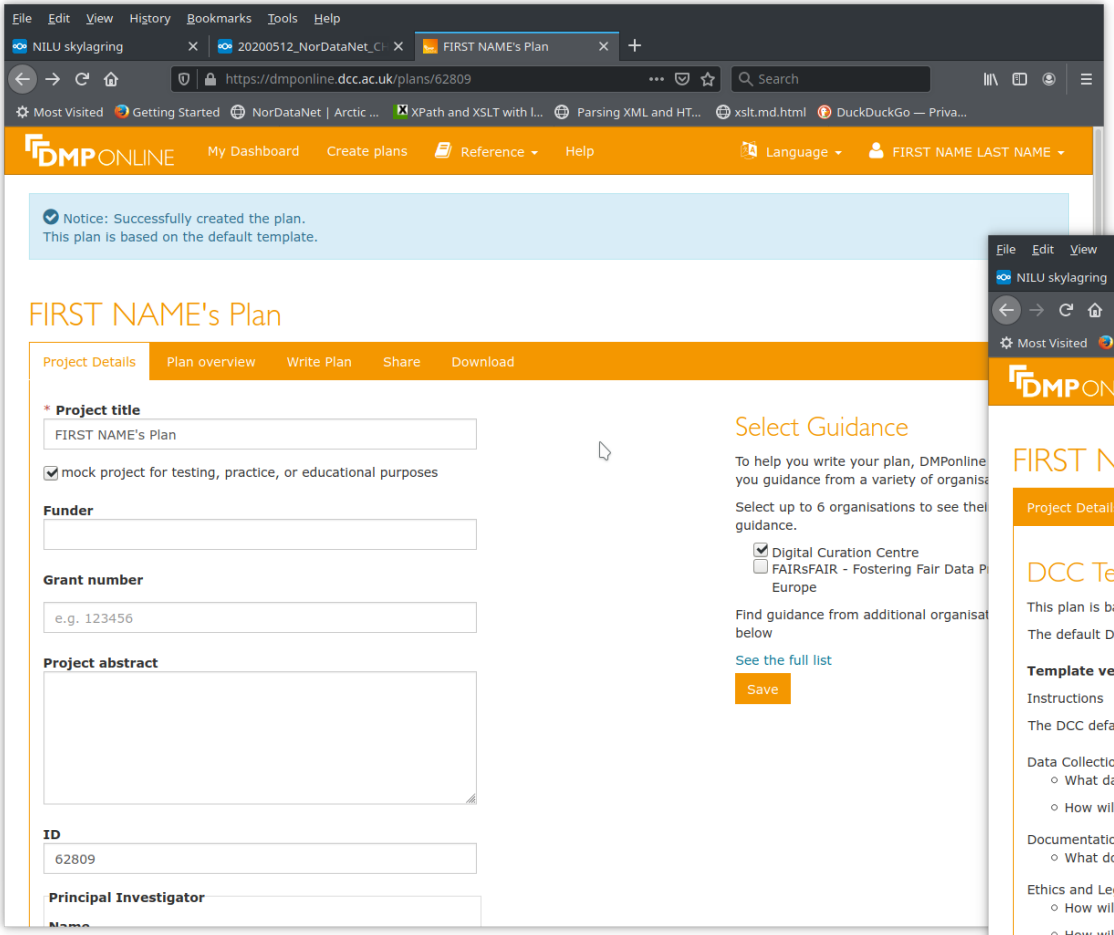
- Are increasingly being required by funding agencies
  - e.g. RCN, EU
- And by e-infrastructure providers
  - e.g. Sigma2

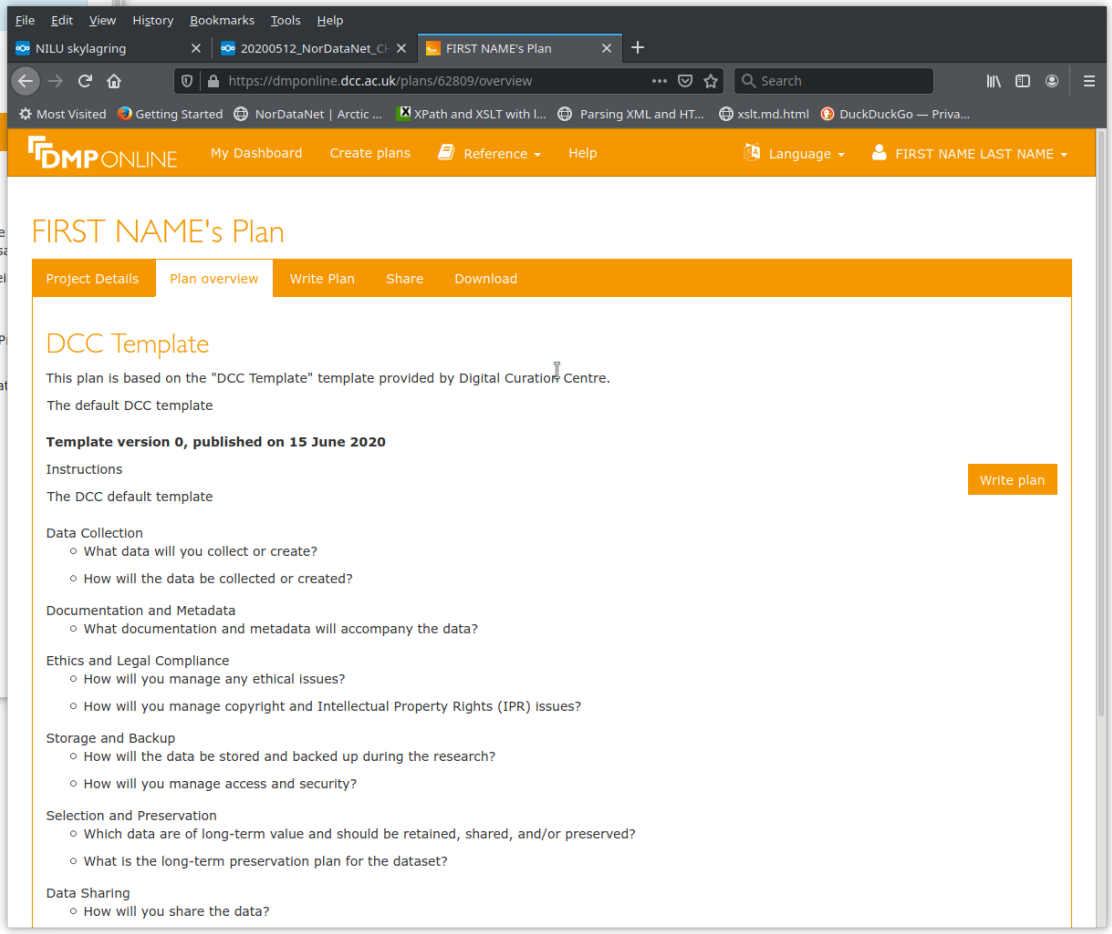




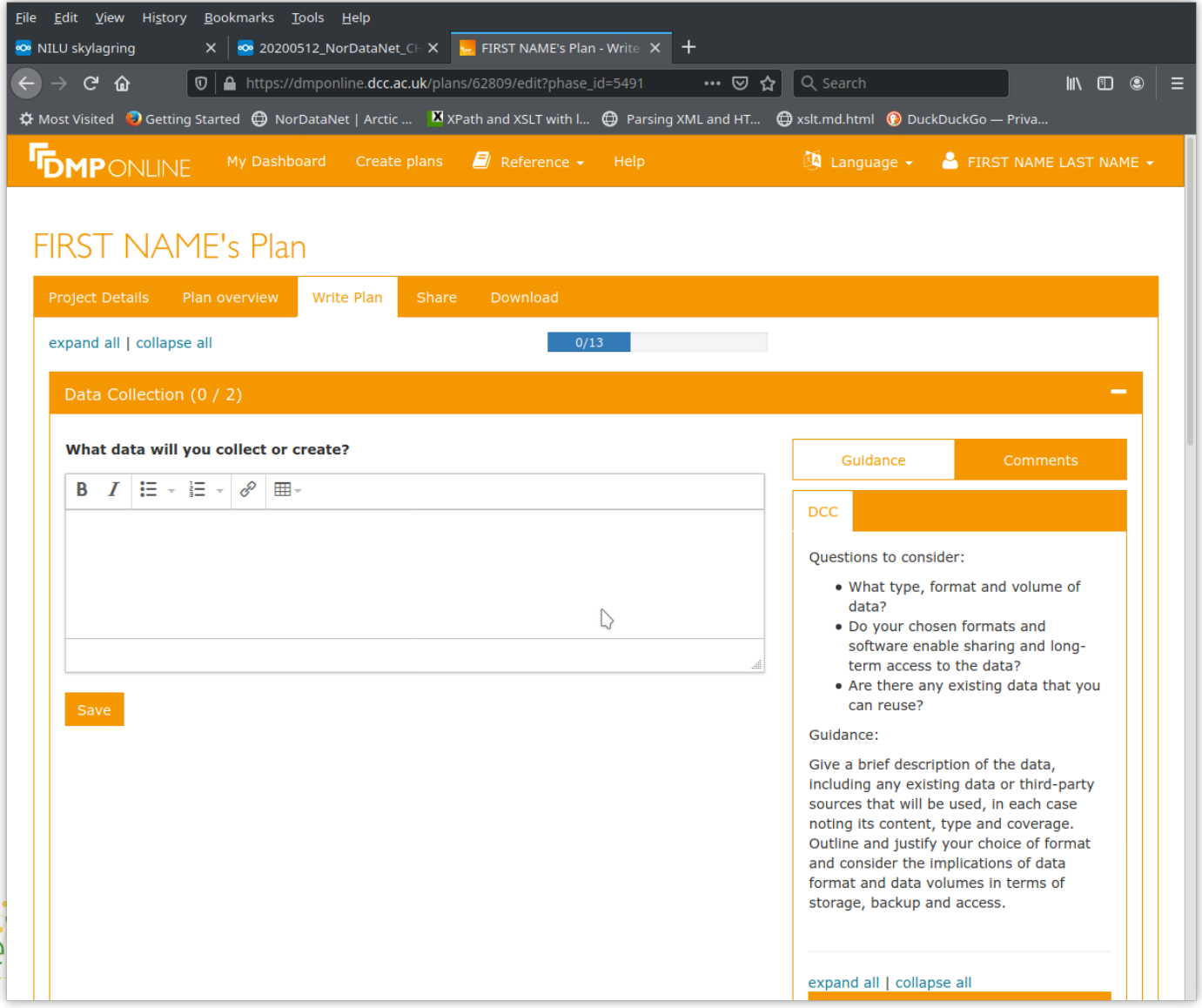


NorDataNet
Norwegian Scientific Data Network

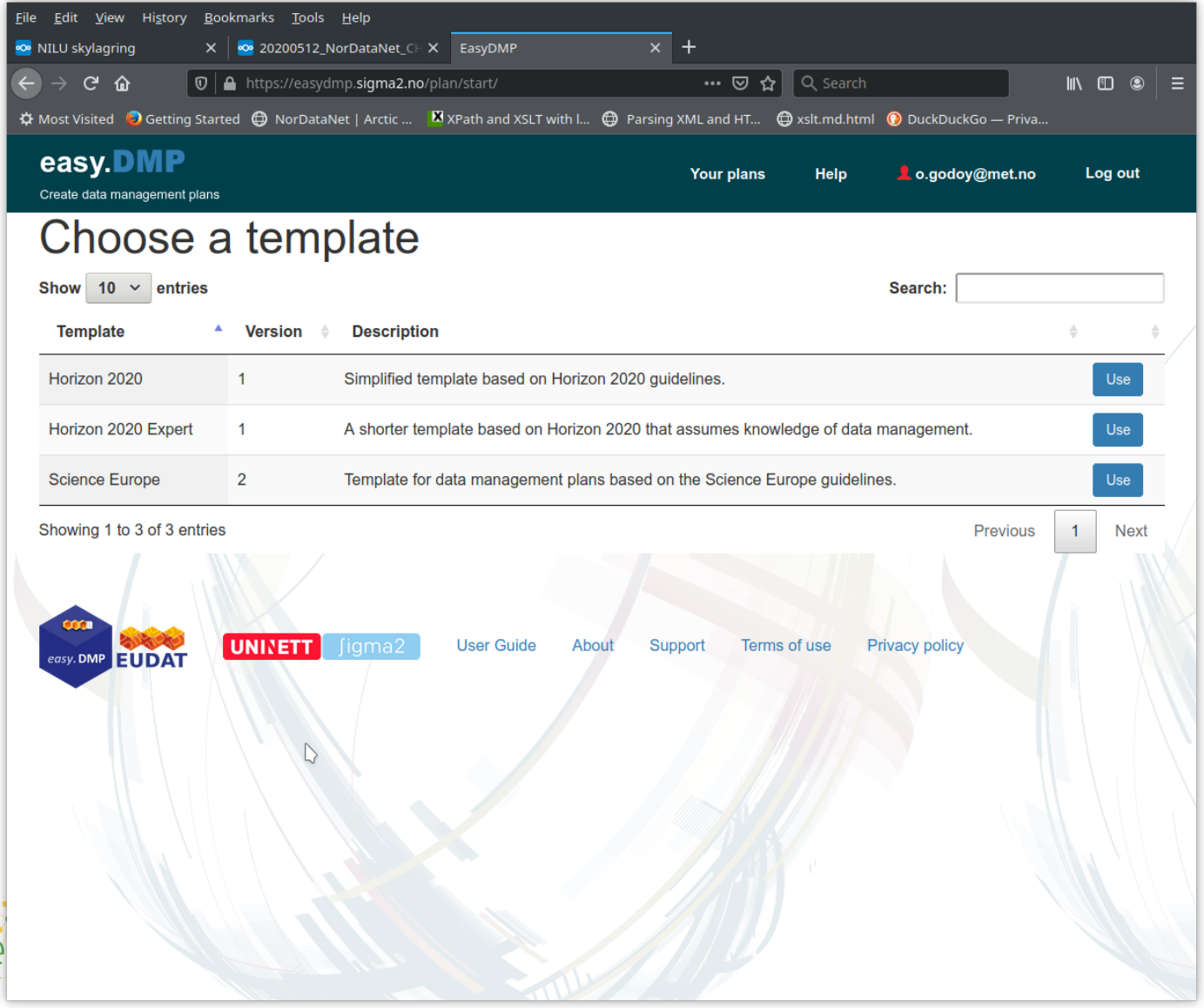




NorDataNet
Norwegian Scientific Data Network



NorDataNet
Norwegian Scientific Data Network



NorDataNet
Norwegian Scientific Data Network