

Starting soon...

# 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

An illustration of an iceberg floating in a blue ocean. The tip of the iceberg is above the water surface, while the much larger, jagged base is submerged below. The background is a gradient of blue, representing the sky and the water.

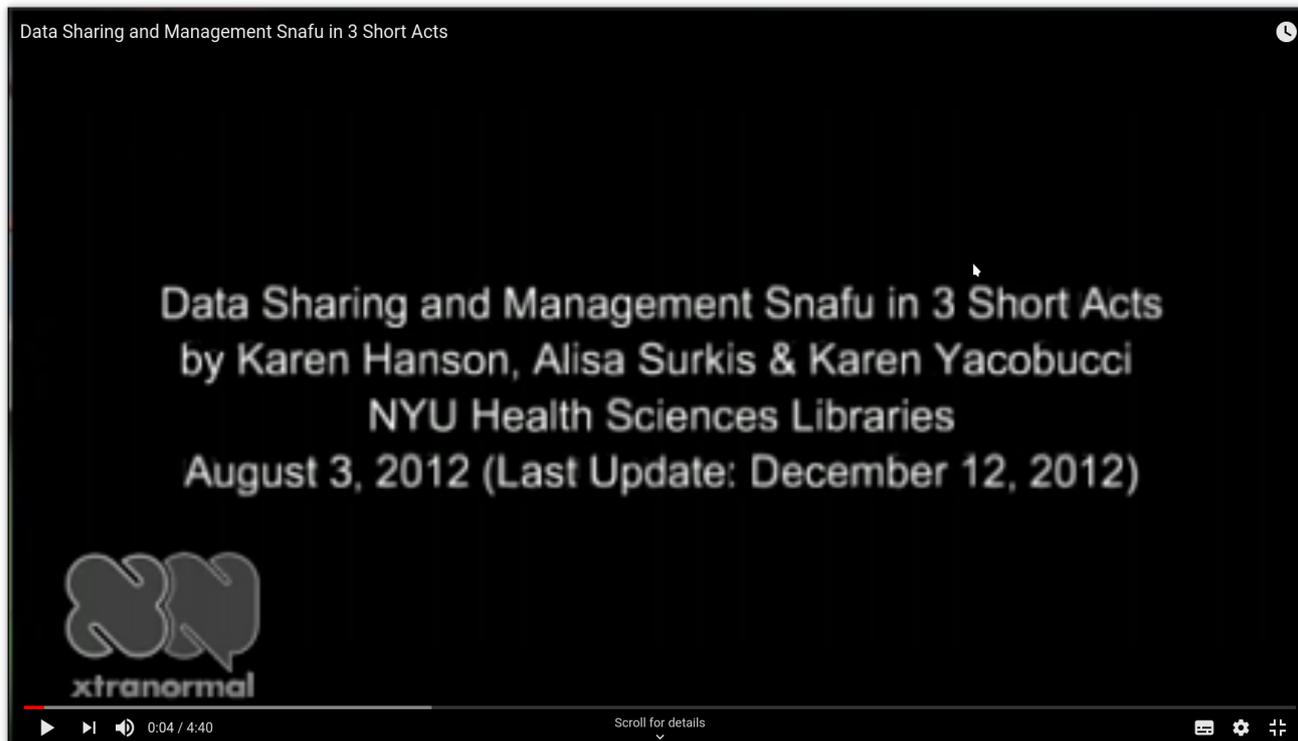
# Motivation: Why do we need data management?

Øystein Godøy

# 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.

# Data Sharing and Management Snafu in 3 Short Acts



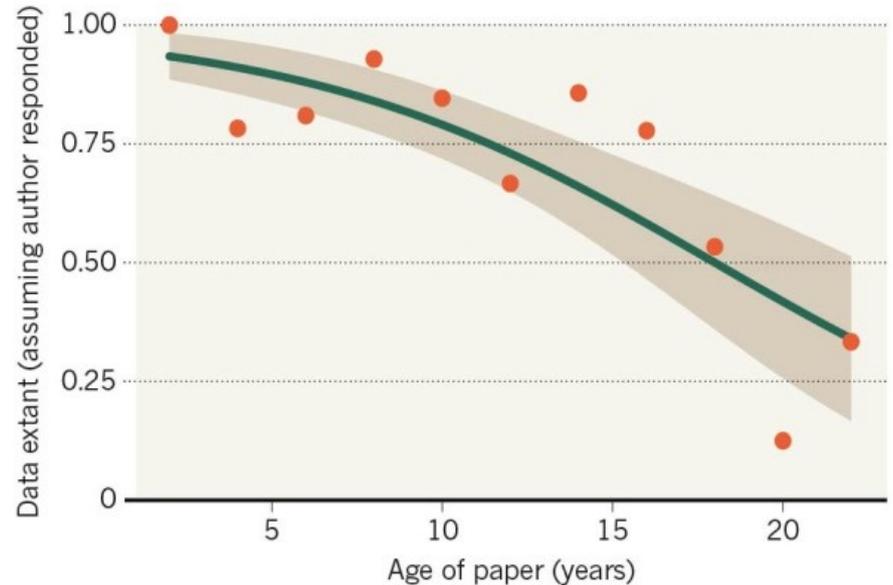
<https://www.youtube.com/watch?v=N2zK3sAtr-4>

# Why do we need data management?

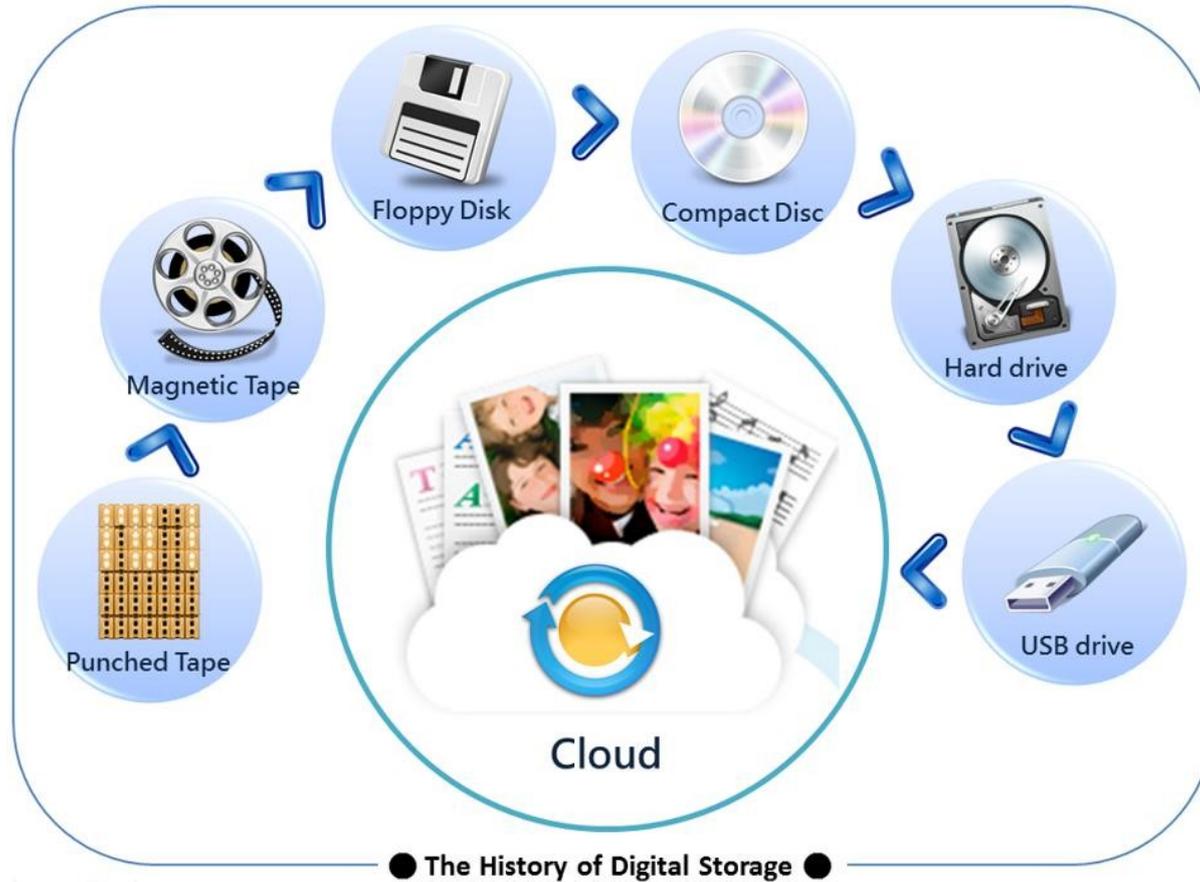
- Loosing scientific data
  - Decline can mean 80% of data are unavailable after 20 years.
    - Gibney and Van Noorden (2013), Nature

## MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.







# 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

# 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



CC image by SLU Madrid Campus  
on Flickr

# Making Your Research Easier and Cheaper

## **The 5 P's matter!**

**Prior**

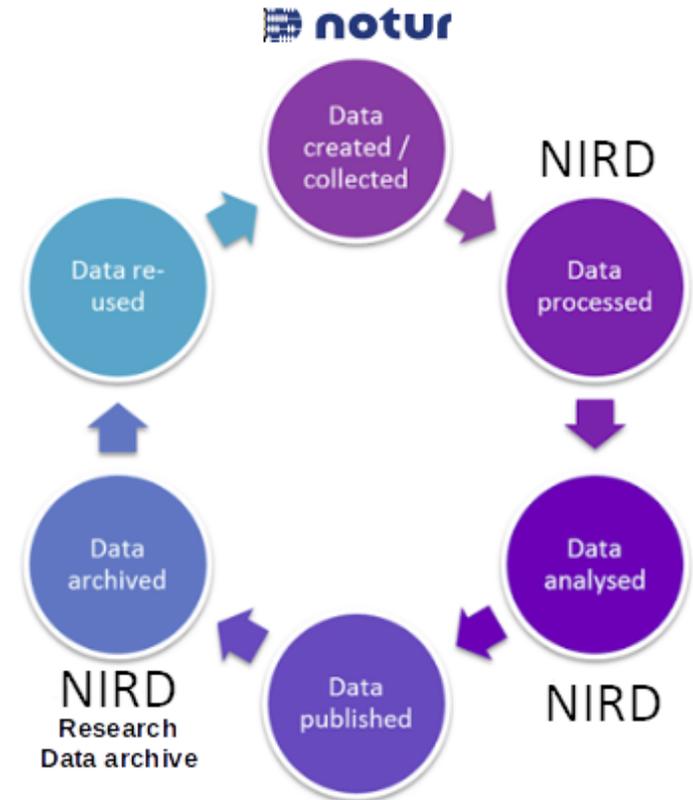
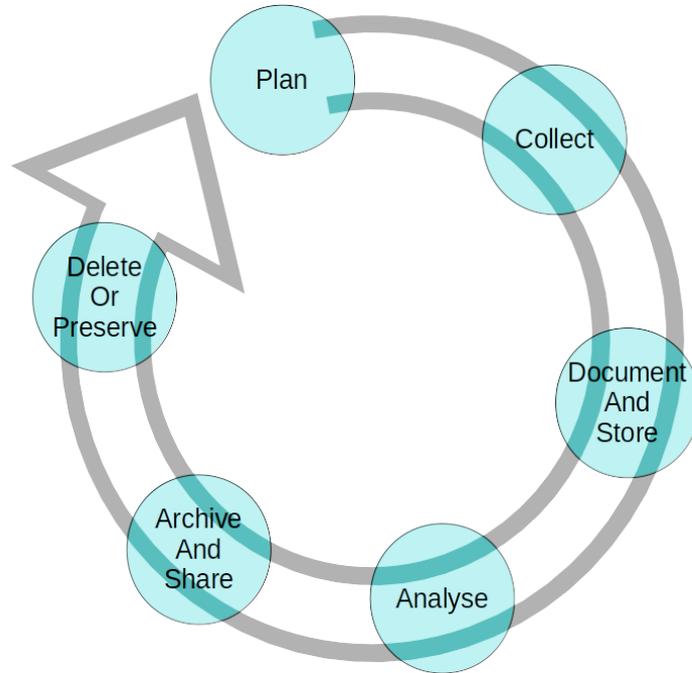
**Planning**

**Prevents**

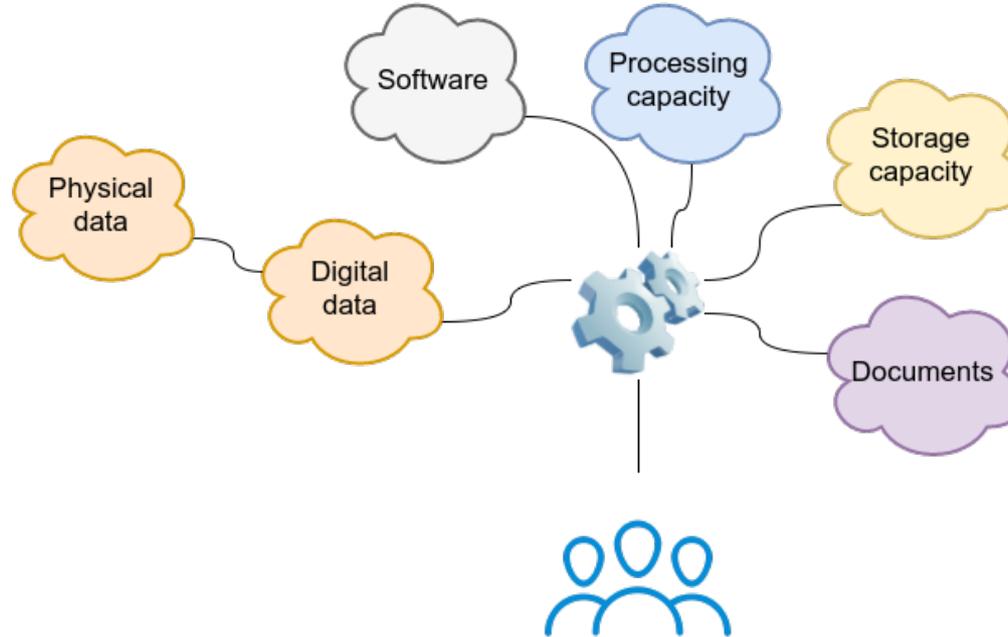
**Poor**

**Performance!**

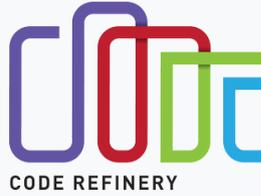
# Science life cycle/Data life cycle



<https://codemeta.github.io/>



# https://coderefinery.org/



## Training and e-Infrastructure for Research Software Development

We are working with students, researchers, Research Software Engineers from all disciplines and national e-infrastructure partners to advance FAIRness of Software management and development practices so that research groups can collaboratively develop, review, discuss, test, share and reuse their codes.

We are also an open-source project developing the lesson materials behind this. They are generic, anyone may use them or join us in developing them.

### Check out our [upcoming events and workshops](#)



#### Training opportunities

We offer training opportunities to researchers from Nordic research groups and projects to learn basic-to-advanced research computing skills and become confident in using state-of-the-art tools and practices from modern collaborative software engineering.

[More >](#)



#### Lesson materials

We develop and maintain training material on software best practices for researchers that already write code. Our material addresses all academic disciplines and tries to be as programming language-independent as possible.

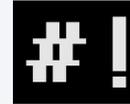
[More >](#)



#### Code Repository Hosting

Our code repository hosting service is open and free for all researchers based in Nordic universities and research institutes. Please contact us if you would like to use these services.

[More >](#)



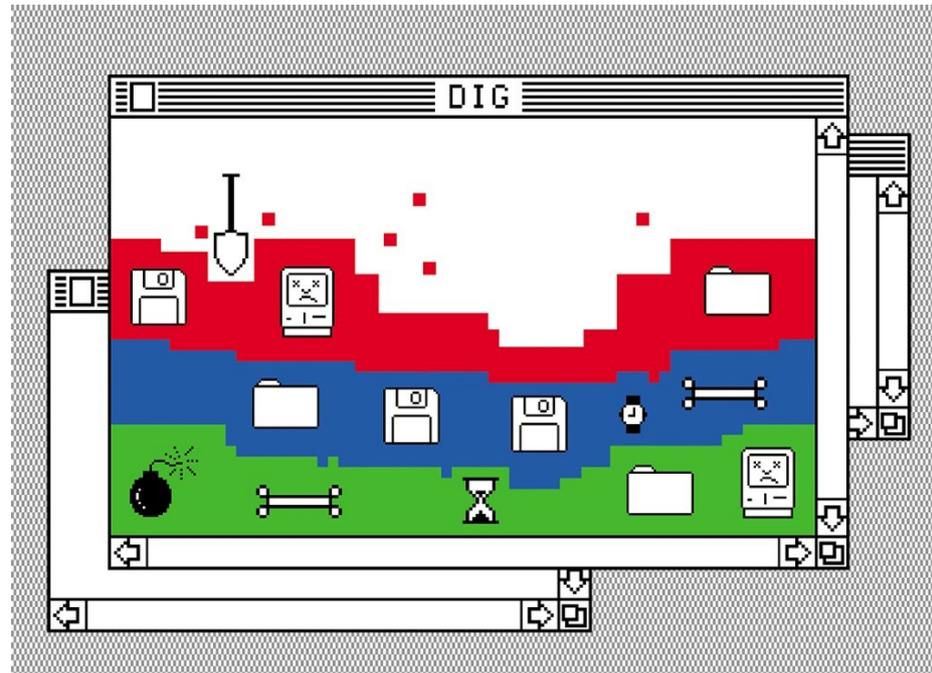
#### Research Software Hour

Research Software Hour is an online stream/show about scientific computing and research software. It is designed to provide the skills typically picked up via informal networks: each ~2 weeks, we do some combination of exploring new tools, analyzing and improving someone's research code, and discussion.

[More >](#)

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

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



# Reproducibility checklist (1)

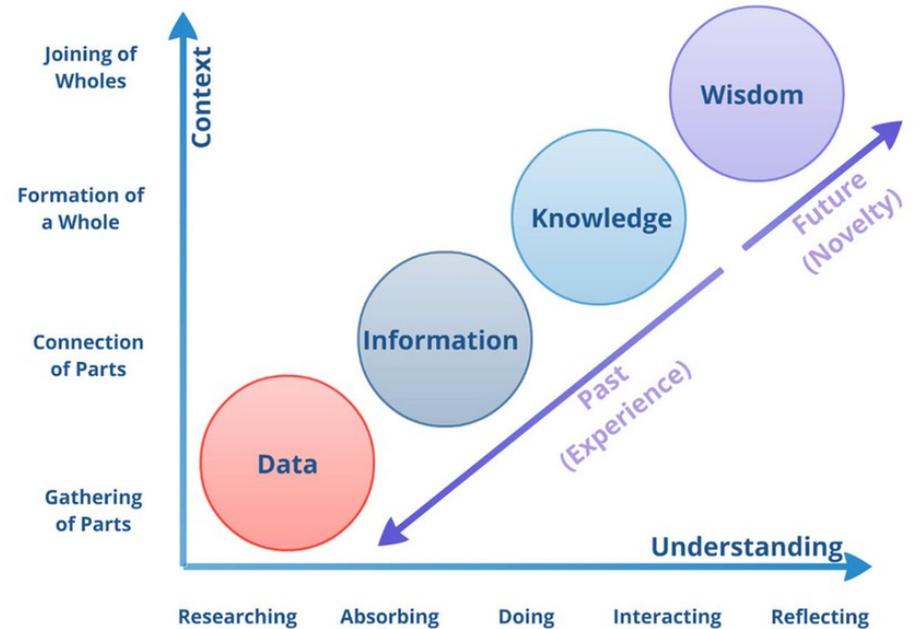
- **Code** Workflows based on point-and-click interfaces, such as Excel, are not reproducible. Enshrine your computations and data manipulation in code.
- **Document** Use comments, computational notebooks and README files to explain how your code works, and to define the expected parameters and the 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.

# 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.

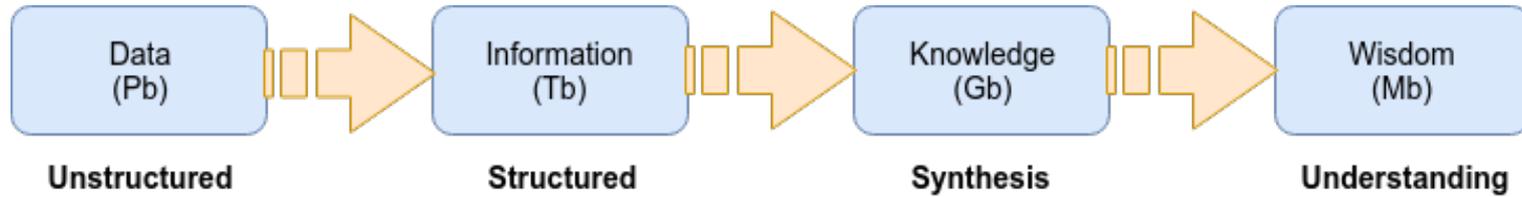
# DIKW chain

- 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/>

# DIKW chain



# The reality today

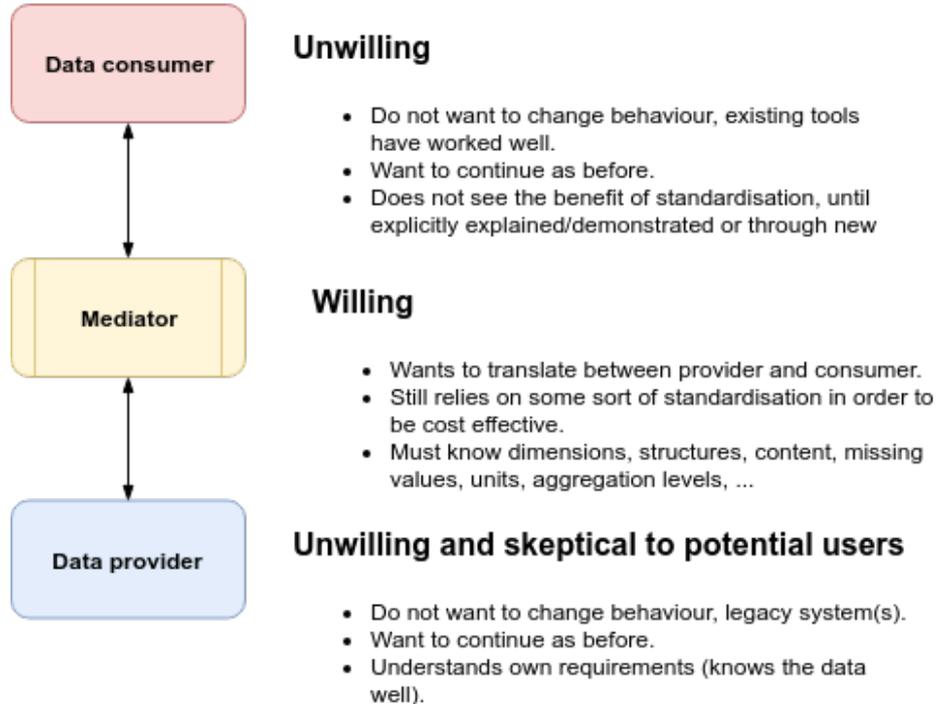


Recreated from Klump et al. 2006

# How to change data sharing culture?



# Bridging between data management actors



- 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

# 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*

# 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

The screenshot shows the Science journal website interface. The main article is titled "People rely on daily weather services to decide what to wear, make transport choices, prepare for rain, and more. Many societal decisions, however, need information not on time scales of days, but on climate time scales of months, years, or decades. New initiatives such as that of Copernicus in Europe provide a wealth of climate data, which are integral to climate services. However, data are only one aspect of climate services, which also involve translation and use of relevant information with the aim to help society manage the risks and opportunities of climate variability and change (1-5). To be successful, any climate service must have a clear problem focus, build on good-quality observations, and consider climate across different time scales." The article is by Miriam Leiss, Philipp R. Barton, et al. Below the article, there are sections for "What's Your Problem?", "Mounting evidence shows, however, that climate information can improve behaviors and outcomes when appropriately incorporated within the decision context.", and "In another application of climate services, this time to the health sector, a comprehensive surveillance and early warning system for dengue was set up in Ecuador over several years." The right sidebar features a "Science" magazine cover with the headline "CRIMINAL INJUSTICE" and a "TABLE OF CONTENTS" section with a PDF icon. At the bottom, there is a "LATEST NEWS" section and a cookie consent banner.

synop

Start Date End Date  
 dd / mm / yyyy 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):

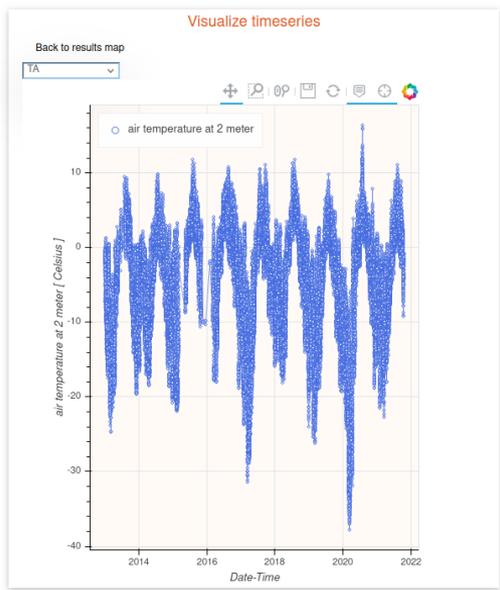
- ADC is the full collection of this service
- CC is the CryoClim<sup>†</sup> collection
- GCW is the collection for the WMO Global Cryosphere Watch
- NMAP is the NORMAP<sup>†</sup> e-Infrastructure collection
- NSDN is the data collection for the Norwegian Scientific Data Network<sup>†</sup>
- SIOS is the data collection for the Svalbard Integrated Arctic Earth Observing System<sup>†</sup>
- 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<sup>†</sup>
- ACCESS is the data collection for the EU FP7 project ACCESS<sup>†</sup>
- APPL is the data collection for the EU H2020 project APPLICATE<sup>†</sup>
- AP is the data collection for the EU H2020 project Arctic Passion<sup>†</sup>

- Project**
- APPLICATE (450)
  - SIOS (450)
  - YOPP (450)
- Organisation**
- Norwegian Meteorological Institute (450)
- Publisher**
- Data Center**
- WMO (450)



**SYNOP** data from station VERLEGENHUKEN  
 WMO Year of Polar Prediction, Svalbard Integrated Arctic Earth Observing System (YOPP, APPLICATE, SIOS)

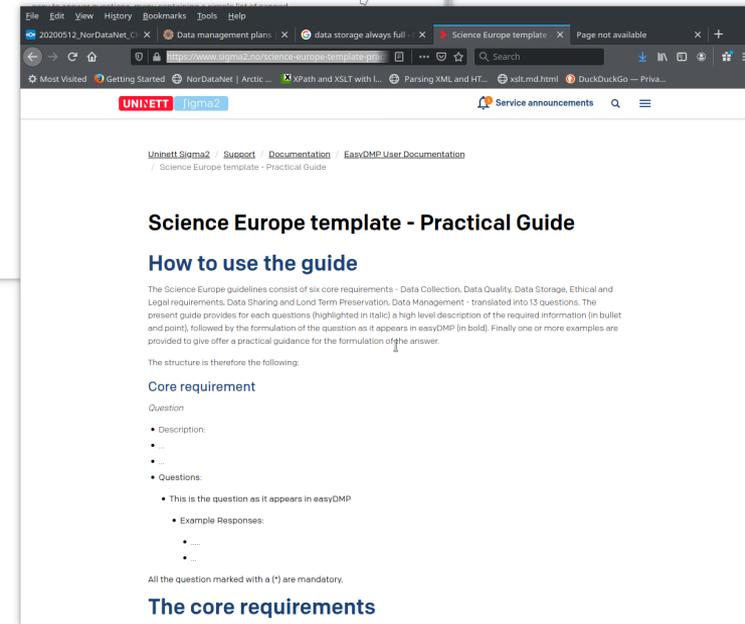
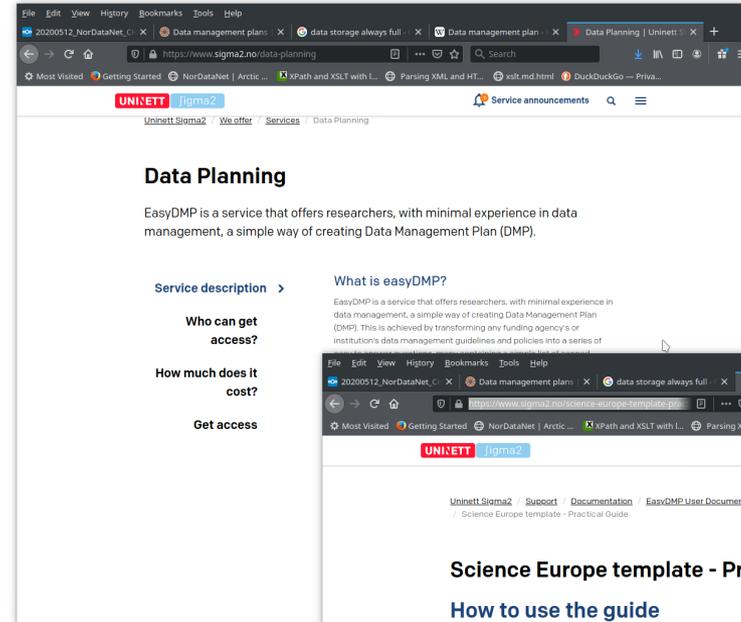
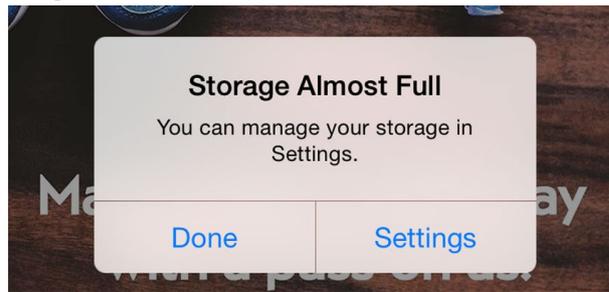


# Data management plans

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.

# Data management plans

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



File Edit View History Bookmarks Tools Help  
NILU skylagring x 20200512\_NorDataNet\_Ci x FIRST NAME's Plan x +  
https://dmponline.dcc.ac.uk/plans/62809  
Most Visited Getting Started NorDataNet | Arctic ... XPath and XSLT with L... Parsing XML and HT... xslt.md.html DuckDuckGo — Priva...  
DMPONLINE My Dashboard Create plans Reference Help Language FIRST NAME LAST NAME

Notice: Successfully created the plan.  
This plan is based on the default template.

## FIRST NAME's Plan

Project Details Plan overview Write Plan Share Download

### Project title

FIRST NAME's Plan

mock project for testing, practice, or educational purposes

### Funder

### Grant number

e.g. 123456

### Project abstract

### ID

62809

### Principal Investigator

Name

## Select Guidance

To help you write your plan, DMPonline provides you guidance from a variety of organisations.

Select up to 6 organisations to see the relevant guidance.

- Digital Curation Centre
- FAIRsFAIR - Fostering Fair Data Practices in Europe

Find guidance from additional organisations below

[See the full list](#)

Save

File Edit View History Bookmarks Tools Help  
NILU skylagring x 20200512\_NorDataNet\_Ci x FIRST NAME's Plan x +  
https://dmponline.dcc.ac.uk/plans/62809/overview  
Most Visited Getting Started NorDataNet | Arctic ... XPath and XSLT with L... Parsing XML and HT... xslt.md.html DuckDuckGo — Priva...  
DMPONLINE My Dashboard Create plans Reference Help Language FIRST NAME LAST NAME

## FIRST NAME's Plan

Project Details Plan overview Write Plan Share Download

## DCC Template

This plan is based on the "DCC Template" template provided by Digital Curation Centre.

The default DCC template

### Template version 0, published on 15 June 2020

Instructions

The DCC default template

Write plan

#### Data Collection

- What data will you collect or create?
- How will the data be collected or created?

#### Documentation and Metadata

- What documentation and metadata will accompany the data?

#### Ethics and Legal Compliance

- How will you manage any ethical issues?
- How will you manage copyright and Intellectual Property Rights (IPR) issues?

#### Storage and Backup

- How will the data be stored and backed up during the research?
- How will you manage access and security?

#### Selection and Preservation

- Which data are of long-term value and should be retained, shared, and/or preserved?
- What is the long-term preservation plan for the dataset?

#### Data Sharing

- How will you share the data?

# FIRST NAME's Plan

Project Details Plan overview **Write Plan** Share Download

expand all | collapse all 0/13

## Data Collection (0 / 2)

### What data will you collect or create?

**B** *I* [List] [List] [Link] [Table]

Save

Guidance Comments

### DCC

#### Questions to consider:

- What type, format and volume of data?
- Do your chosen formats and software enable sharing and long-term access to the data?
- Are there any existing data that you can reuse?

#### Guidance:

Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.

expand all | collapse all

File Edit View History Bookmarks Tools Help

NILU skylagring 20200512\_NorDataNet\_C EasyDMP

https://easydmp.sigma2.no/plan/start/

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

**easy.DMP** Your plans Help o.godoy@met.no Log out

Create data management plans

# Choose a template

Show 10 entries Search:

Template	Version	Description	
Horizon 2020	1	Simplified template based on Horizon 2020 guidelines.	<a href="#">Use</a>
Horizon 2020 Expert	1	A shorter template based on Horizon 2020 that assumes knowledge of data management.	<a href="#">Use</a>
Science Europe	2	Template for data management plans based on the Science Europe guidelines.	<a href="#">Use</a>

Showing 1 to 3 of 3 entries Previous 1 Next

    [User Guide](#) [About](#) [Support](#) [Terms of use](#) [Privacy policy](#)