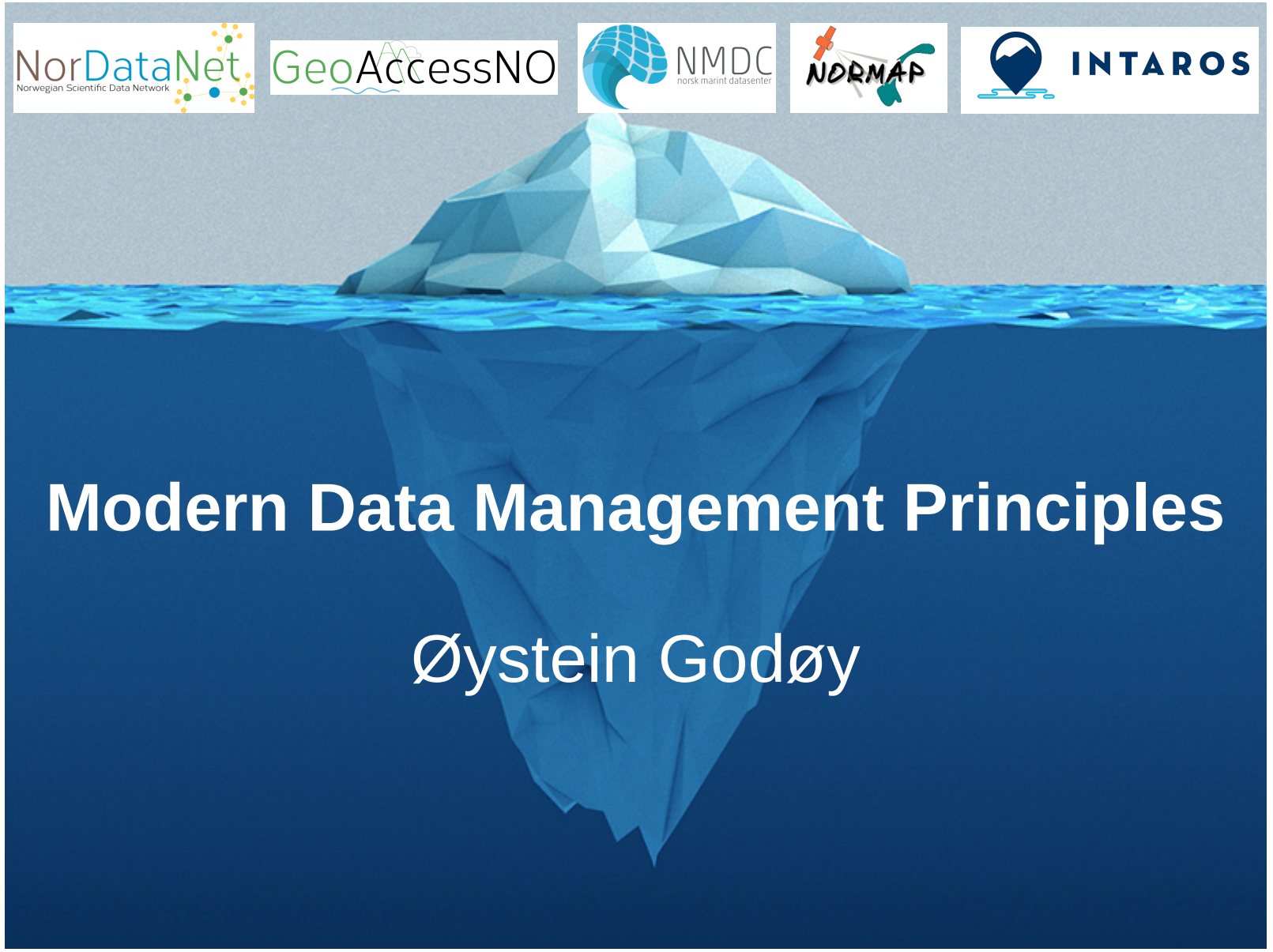

NorDataNet
Norwegian Scientific Data Network
GeoAccessNO
NMDC
norsk marint datasenter
NORMAP
INTAROS
Modern Data Management Principles
Øystein Godøy

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

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Making Your Research Easier and Cheaper

The 5 P's matter!

Prior

Planning

Prevents

Poor

Performance!

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

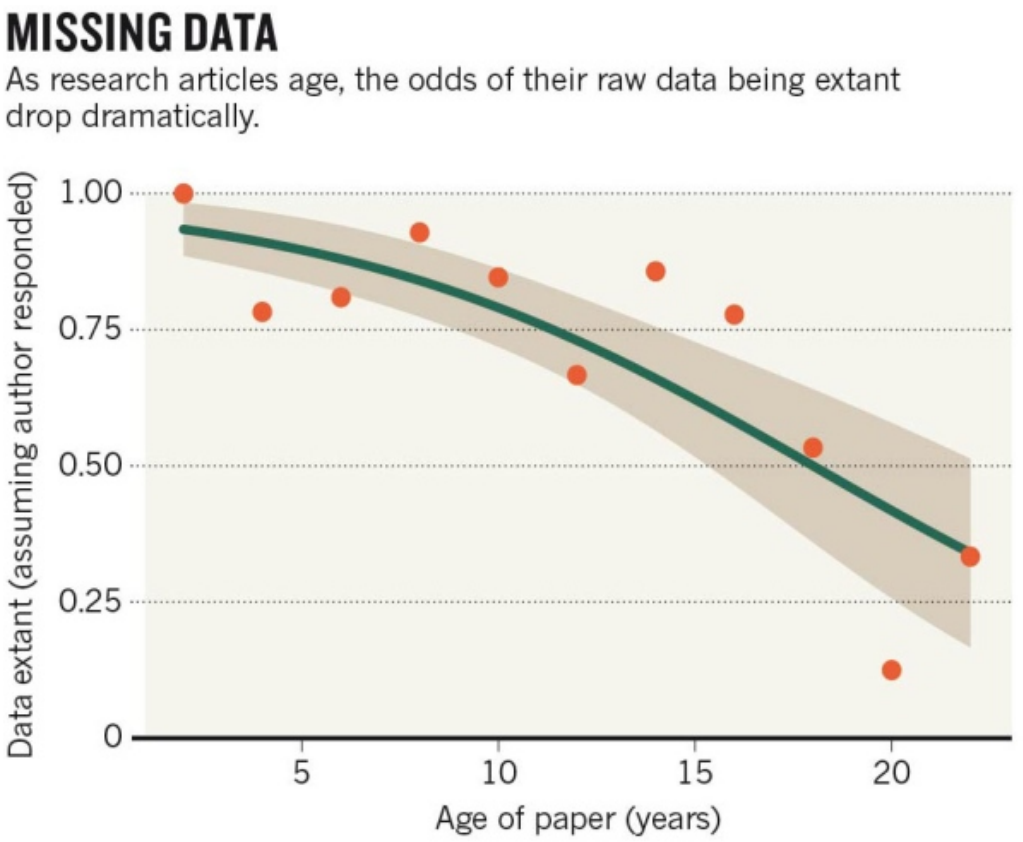
In other words

Learn form the mistakes of other...

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

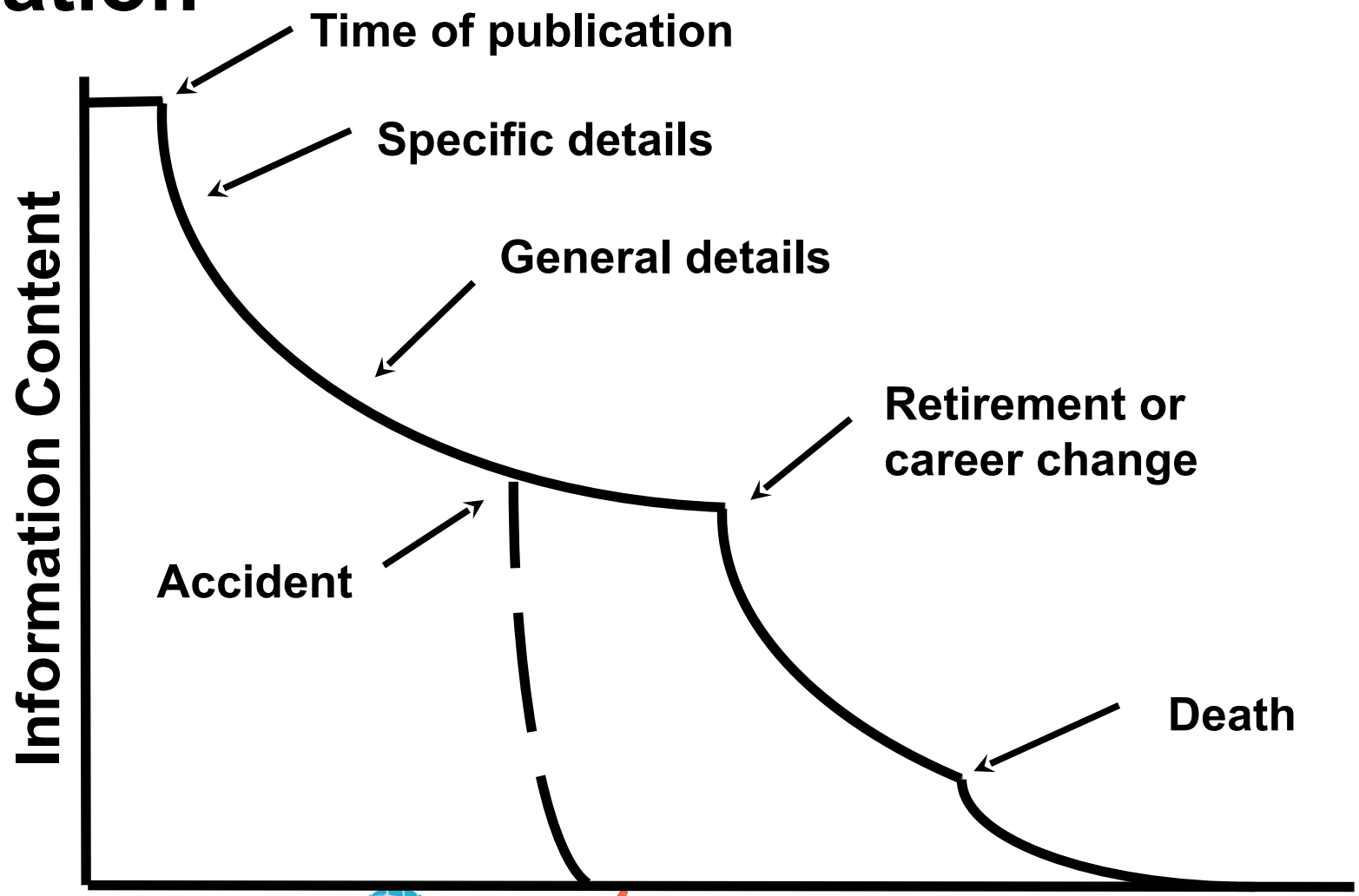
Loosing scientific data

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



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Poor data practice results in loss of information



(Michener et al. 1997)

Why bother with structured data management?

- Science paradigms
  - according to Jim Gray
  - empirical science
  - theoretical science
  - computational science
  - data exploration science

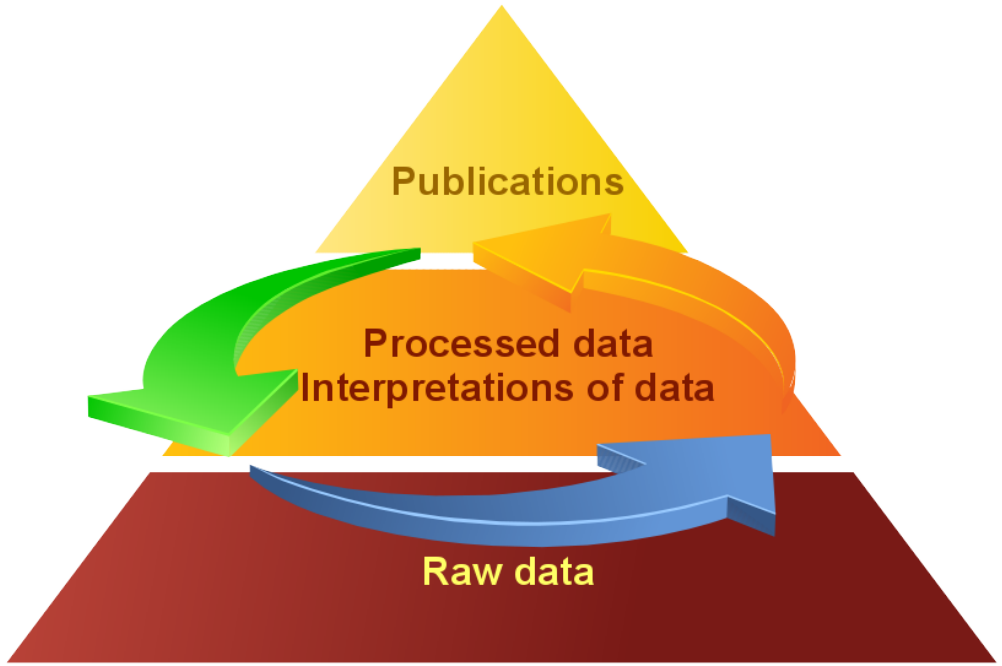
- Maximise public investment in data collection and production
- Promote scientific collaboration
- Promote interdisciplinary science
- Promote scientific transparency
- Leave a legacy

NorDataNet Geo
Norwegian Scientific Data Network

INTAROS

All scientific data online

Source: Jim Gray on eScience:
A Transformed Scientific Method

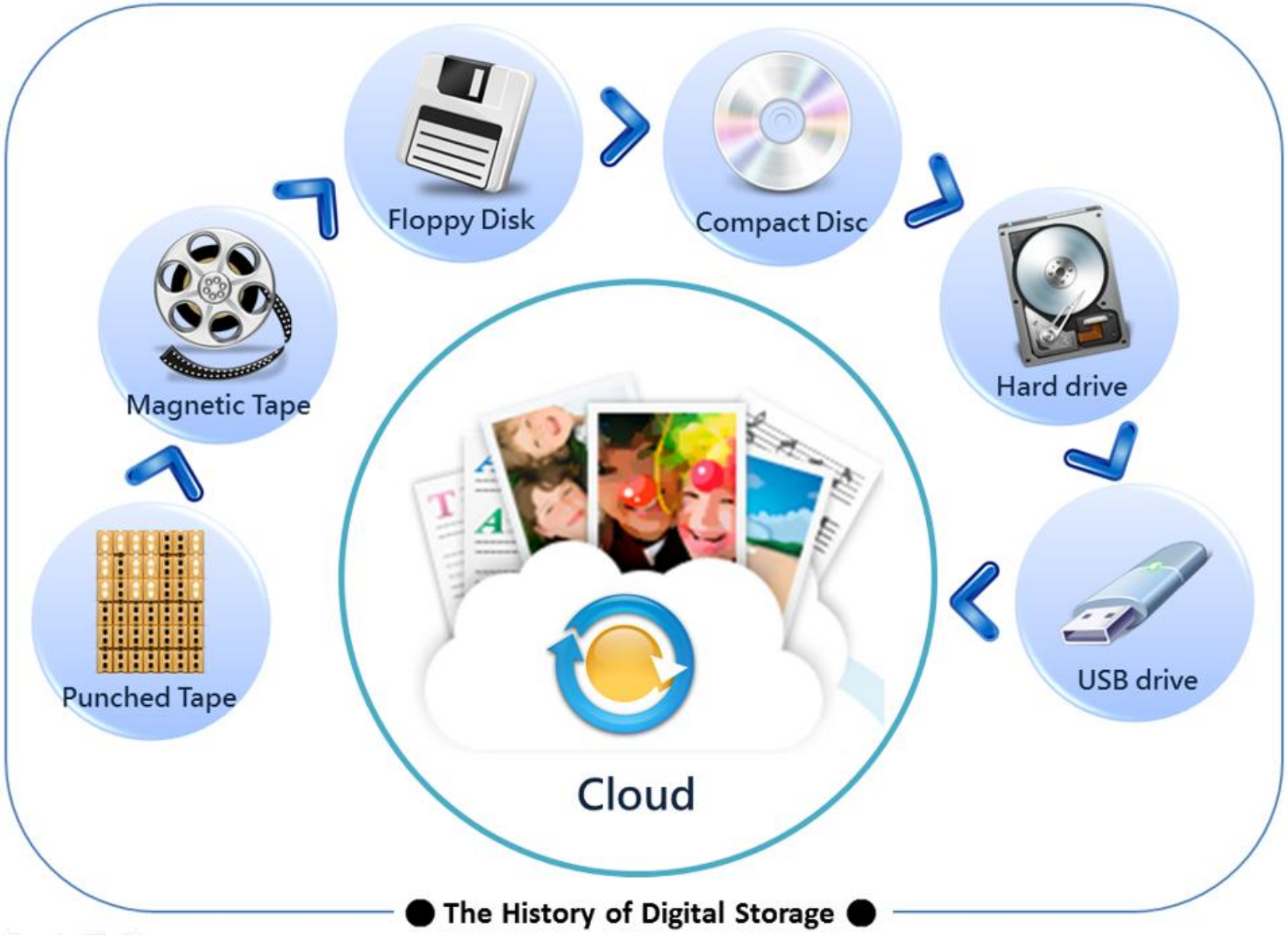


- Many disciplines overlap and use data from other sciences
- Science, government agencies and companies get a broader data background
- Internet can unify data, software and literature
- Go from literature to computation to data back to literature
- Information is at your fingertips for everyone and everywhere
- Potentially Increased Scientific Information Velocity
- Potentially Huge increase in Science Productivity

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

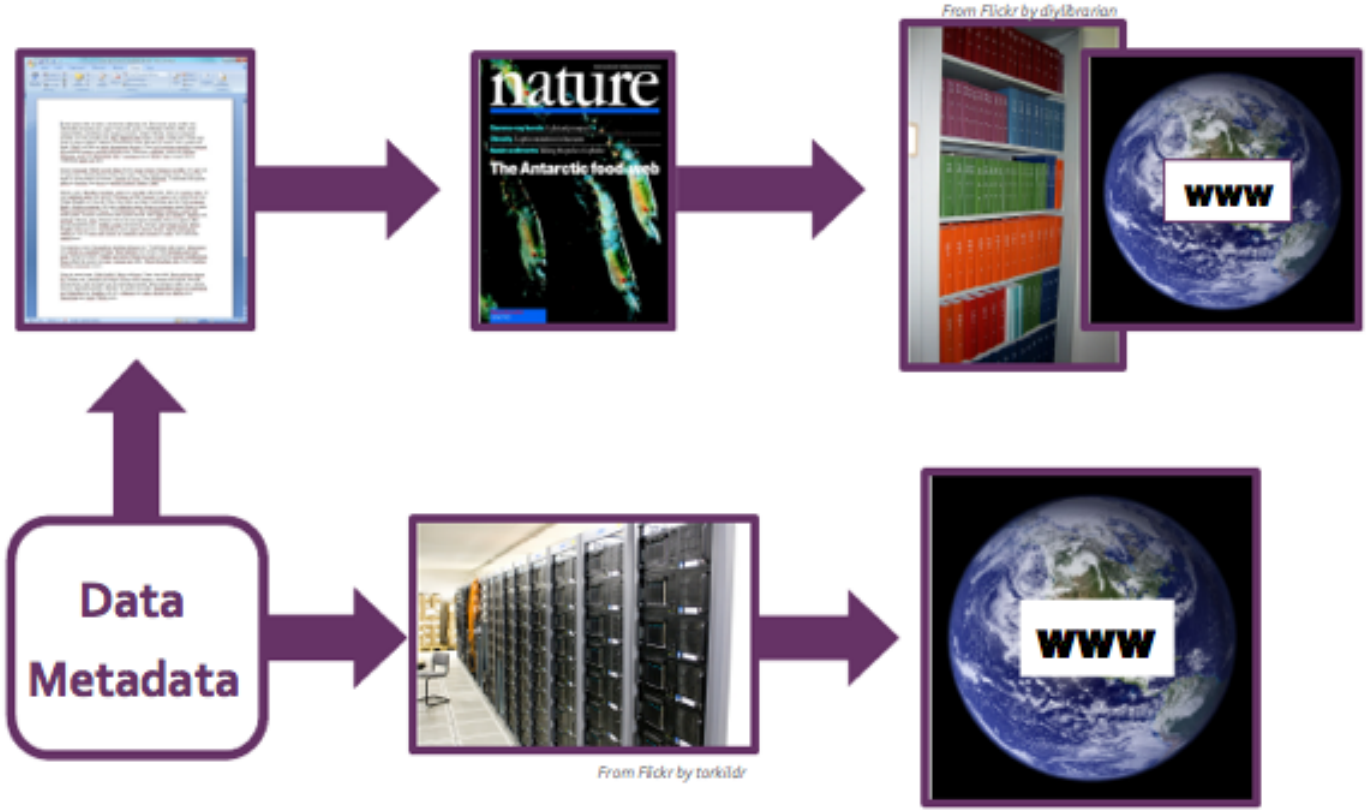
The reality today

Recreated from Klump et al. 2006



http://blog.asuswebstorage.com/tag/digital/

The vision for the future



Recreated from Klump et al. 2006

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS DataONE

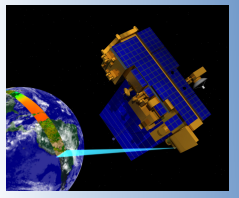
New science

eBird

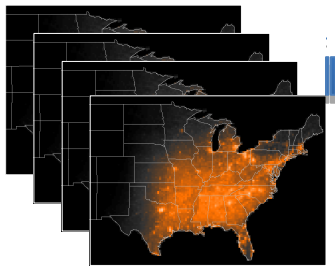
Land Cover

Meteorology

MODIS – Remote sensing data



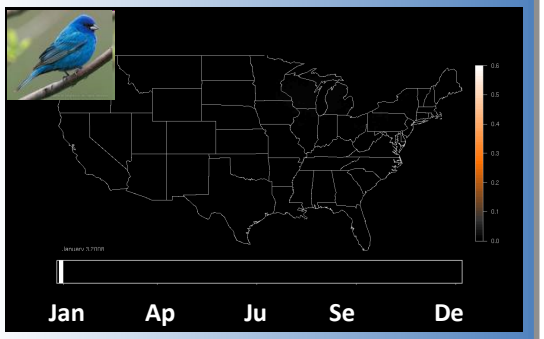
TeraGrid



$$F(X,s,t)=\frac{1}{n(s,t)}\sum_{i=1}^{m} f_i(X,s,t)I(s,t\in\theta_i)$$

Spatio-Temporal Exploratory Models predict the probability of occurrence of bird species across the United States at a 35 km x 35 km grid.

Model results

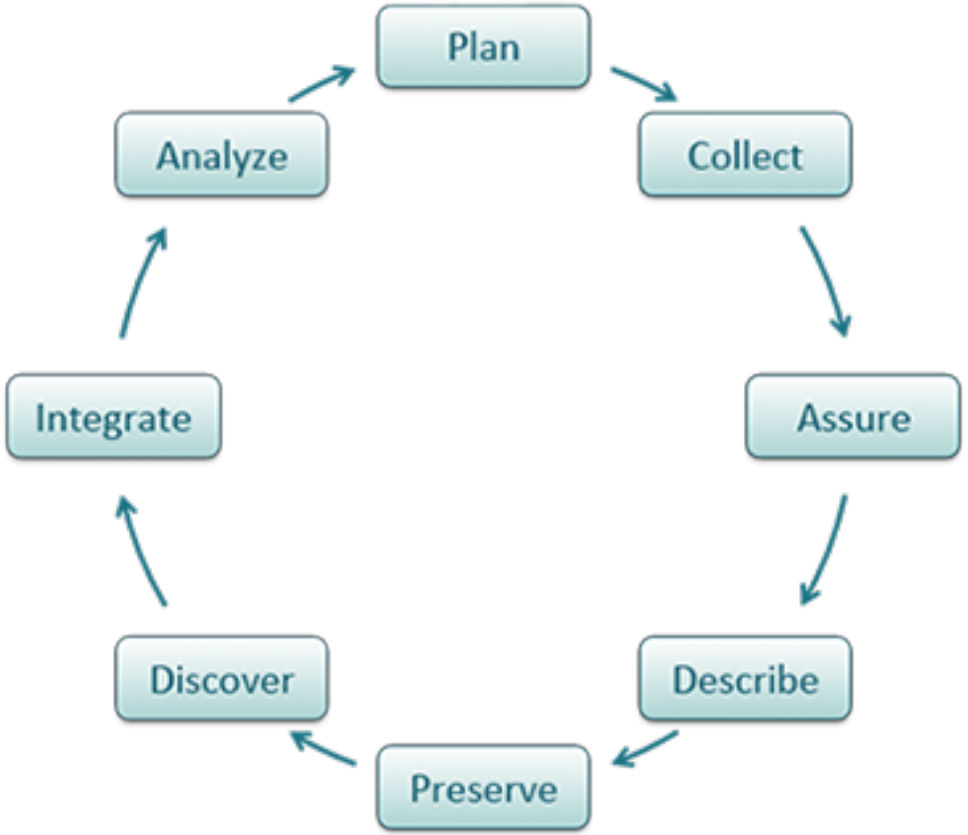


Potential Uses-

- Examine patterns of migration
- Infer impacts of climate change
- Measure patterns of habitat usage
- Measure population trends

By re-using data collected from a variety of sources – eBird database, land cover data, meteorology, and remotely sensed by NASA – this project was able to compile and process the data using supercomputering to determine bird migration routes for particular species.

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS DataONE



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

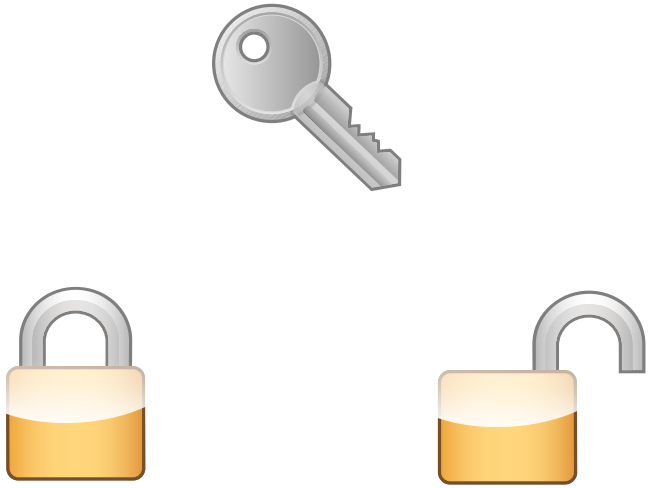
Why share data?

- 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
- Far upstream sponsors require it

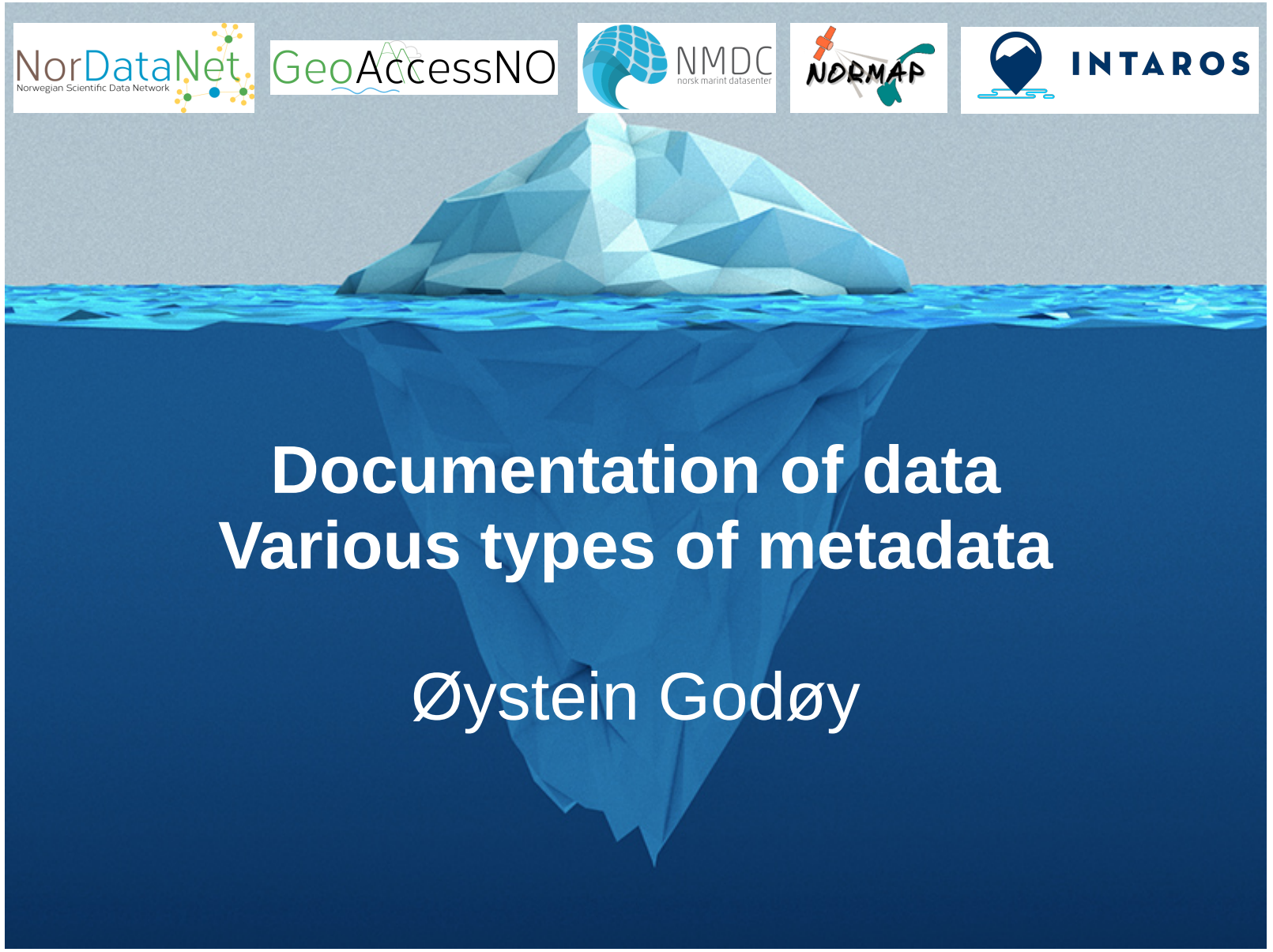
NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS DataONE

Ownership of data

- This depends on the contract between the funding agency and the institution affiliated with the scientist
- For RCN and H2020, ownership lies with the institution
  - Not the individual scientist



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS


NorDataNet
Norwegian Scientific Data Network
GeoAccessNO
NMDC
norsk marint datasenter
NORMAP
INTAROS
Documentation of data
Various types of metadata
Øystein Godøy

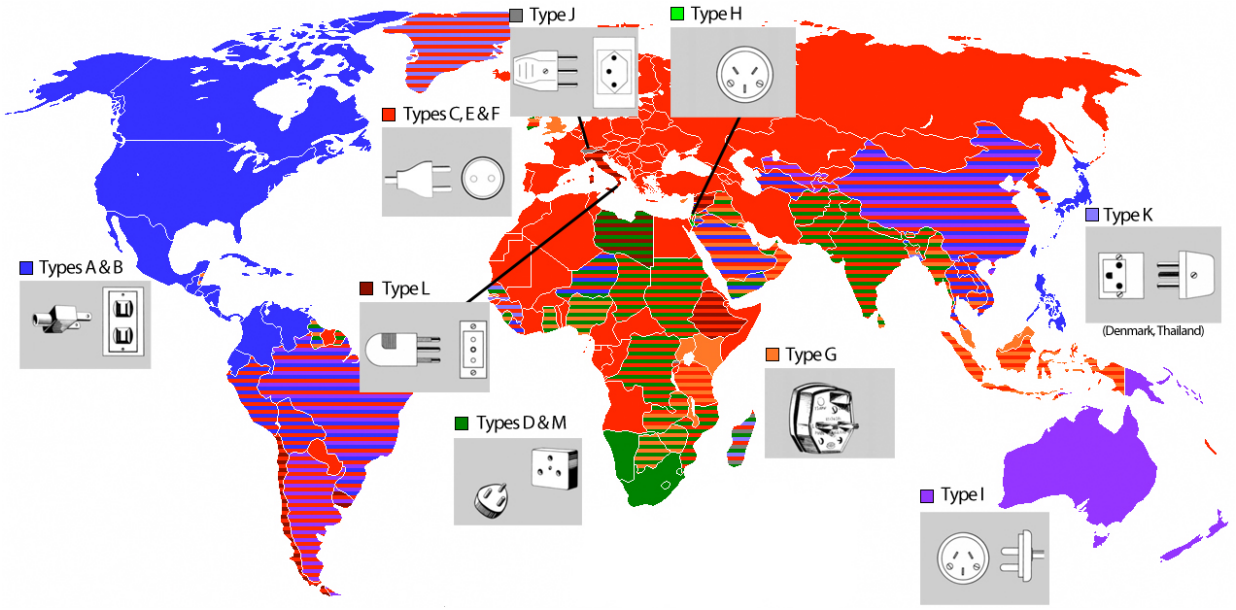
Interoperability

- At the discovery level
  - Documentation
  - Exchange of documentation
- At the data level
  - Documentation
  - Exchange of documentation

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Benefits of standardisation

- Makes life easier
  - Promotes reuse of efforts
- Promotes a common understanding of content
  - Improves performance
  - Reduce risk
  - Promotes sustainability
  - Encourage innovation
  - Reduce cost
  - Improve quality



From HobbyProjects

Benefits of standardised documentation

- Why not use the “Google” approach?
- Science is based on a shared terminology
  - There will never be only one proper way of documenting
  - There will always be a need for brokering
- Data and metadata must be connected
  - To find data
  - To use data
- Standardised documentation and formatting
  - enables the possibility to filter datasets
  - enables the possibility to link datasets
  - enables standardised applications to analyse data
  - enables users to use the data
- Need to be pragmatic…
  - And let computers do the boring part
  - But humans need to instruct computers

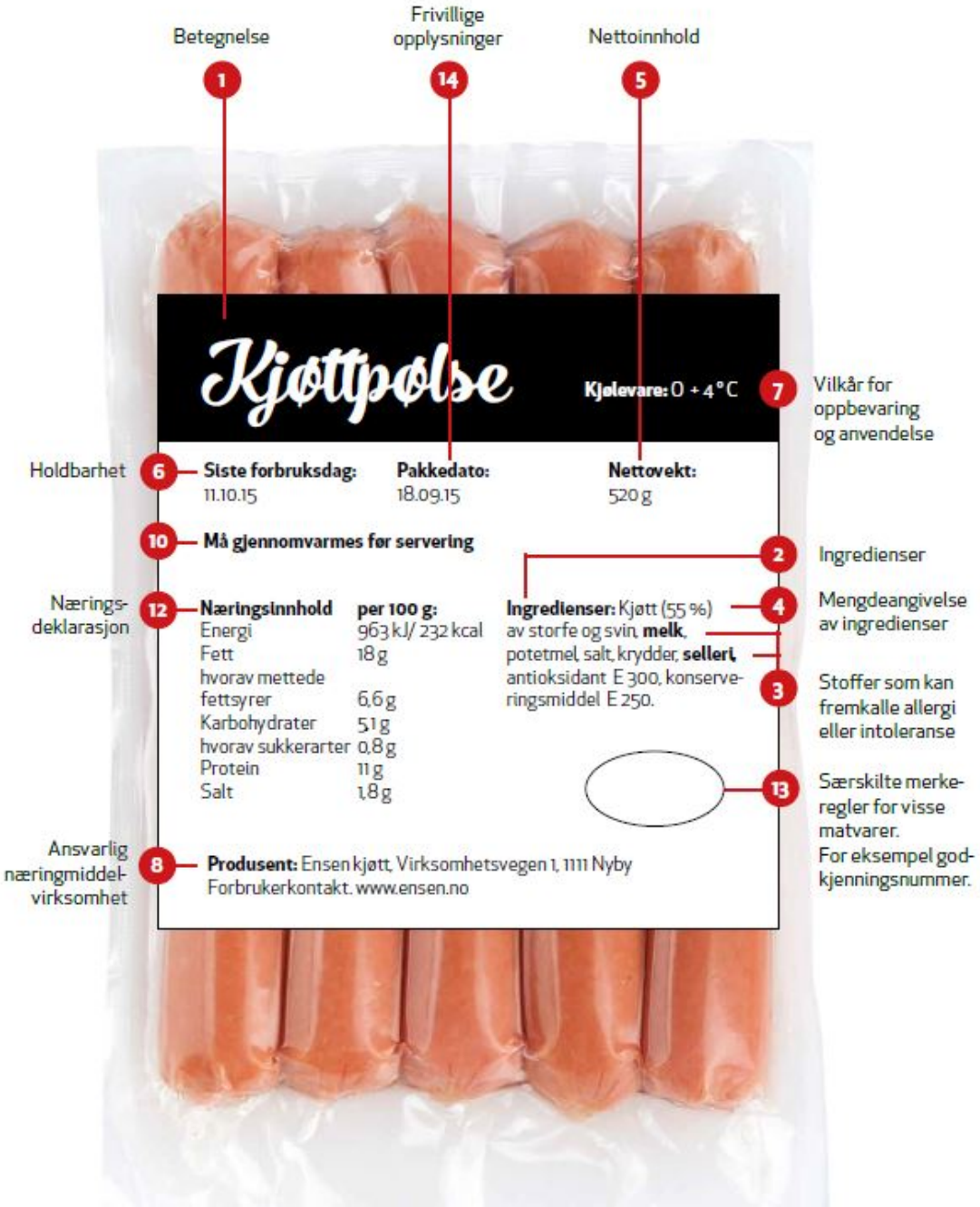
NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Types of metadata

- Discovery metadata
  - ***who*** measured, simulated or analysed ***what***, ***where***, and ***when*** as well as ***conditions for reuse*** and ***access mechanisms*** for the data
  - to enable users to find appropriate data for the task
- Use metadata
  - ***identification of the variables/parameters*** generated, ***units of variables/parameters***, how ***missing values*** are encoded, definition of ***grid and map projections*** for gridded data, ***methodology applied in space or time*** to achieve the values in a dataset etc
  - to enable users to properly understand the data found
- Site metadata
- Configuration metadata

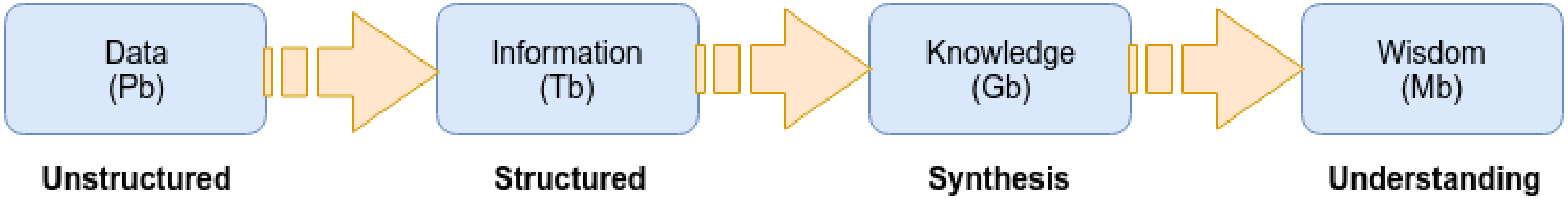
NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Eksempel på merking

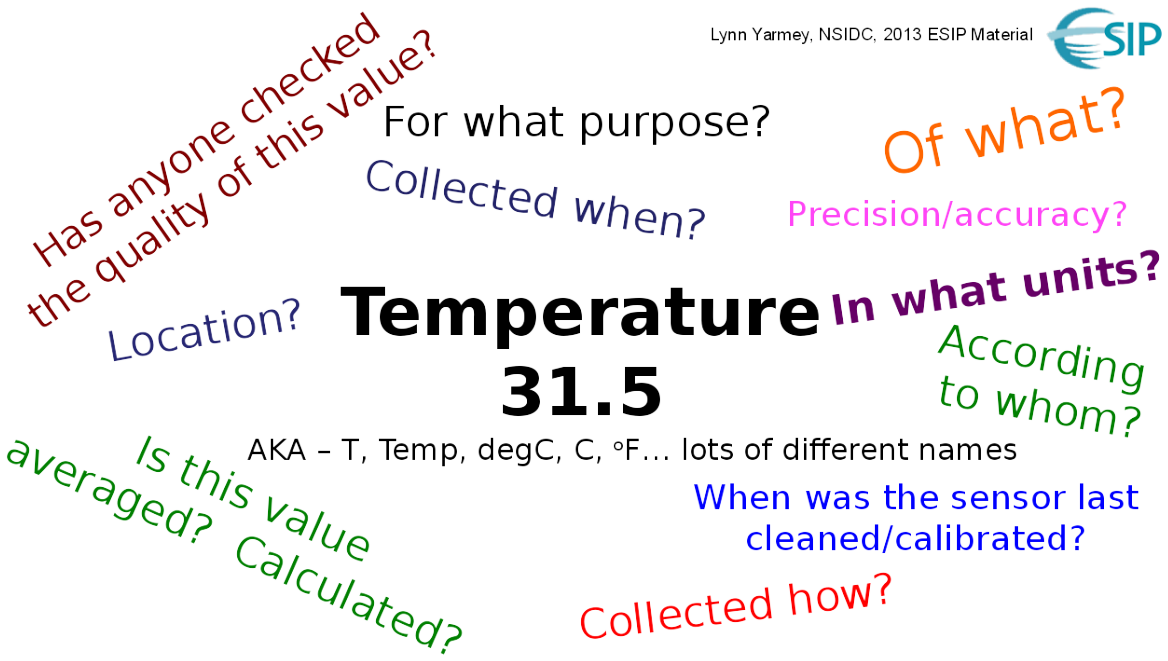


NorDataNet Norwegian Scientific Data Network GeoAcce

DIKW chain


Data
(Pb)
Information
(Tb)
Knowledge
(Gb)
Wisdom
(Mb)
Unstructured
Structured
Synthesis
Understanding

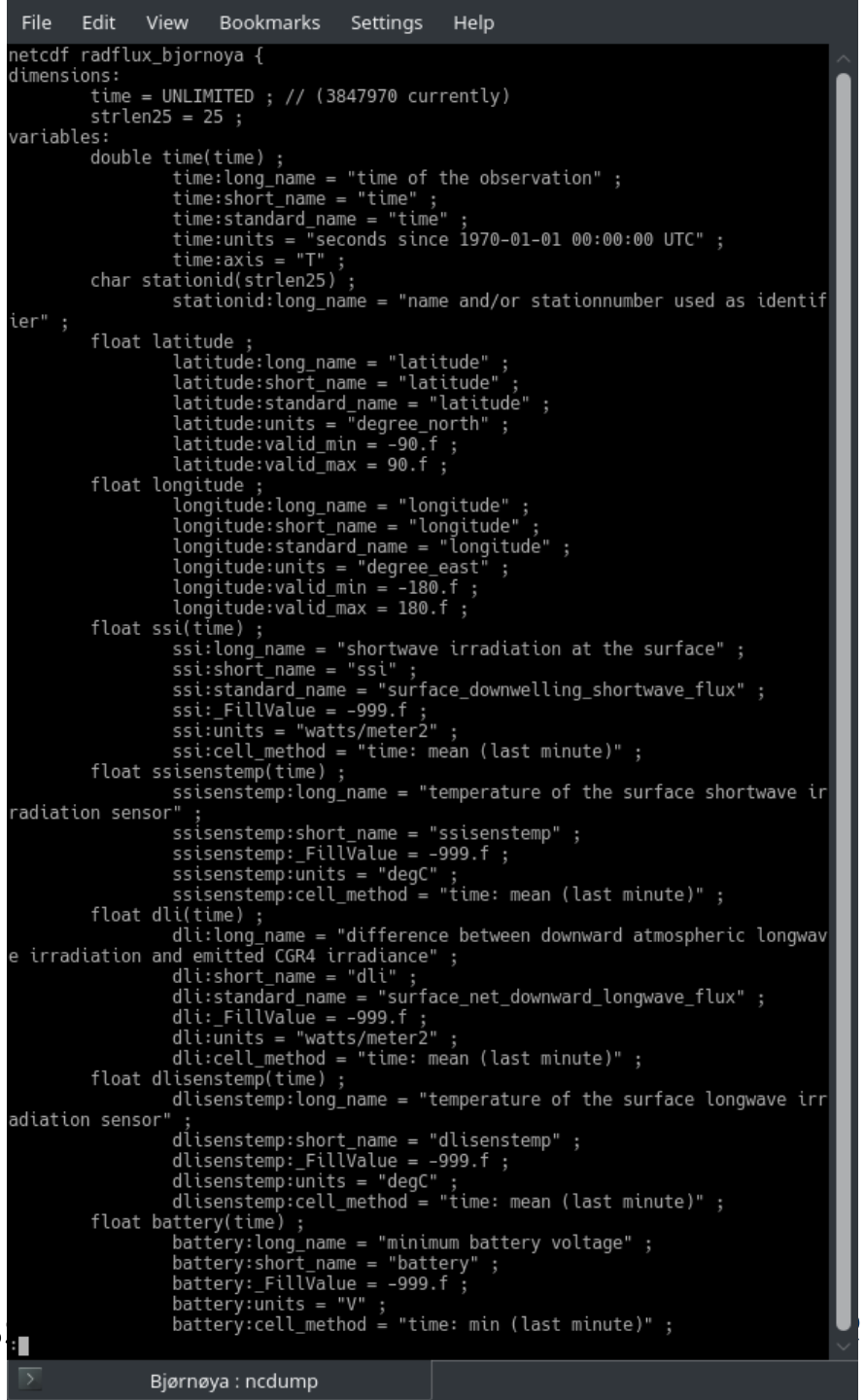
Data in context

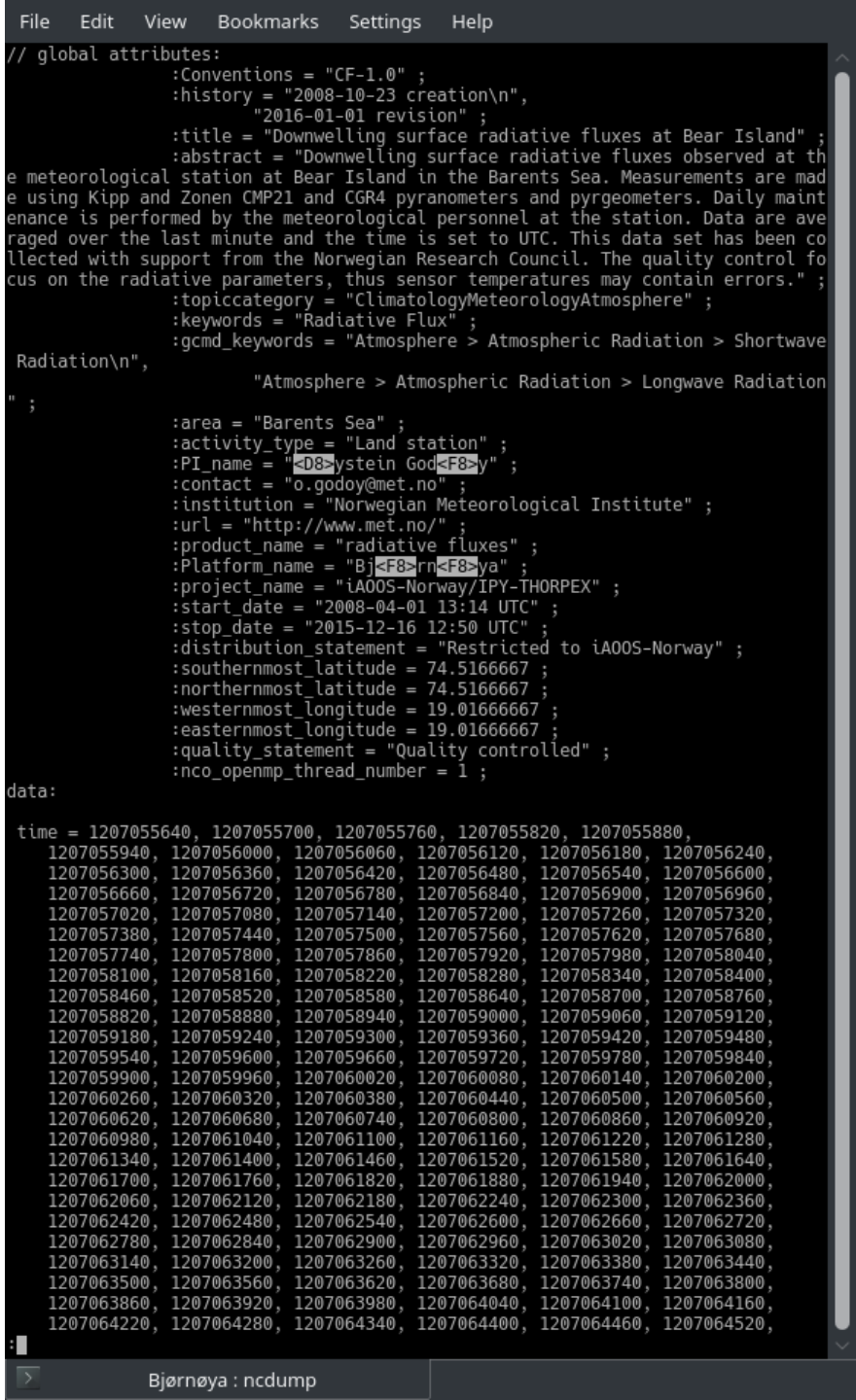


- Make data “talk”
  - Standardised documentation
    - e.g. Climate and Forecast convention
- What is a number?
  - Use metadata is required to efficiently reuse data
  - Describe the context of observations using e.g. WIGOS metadata

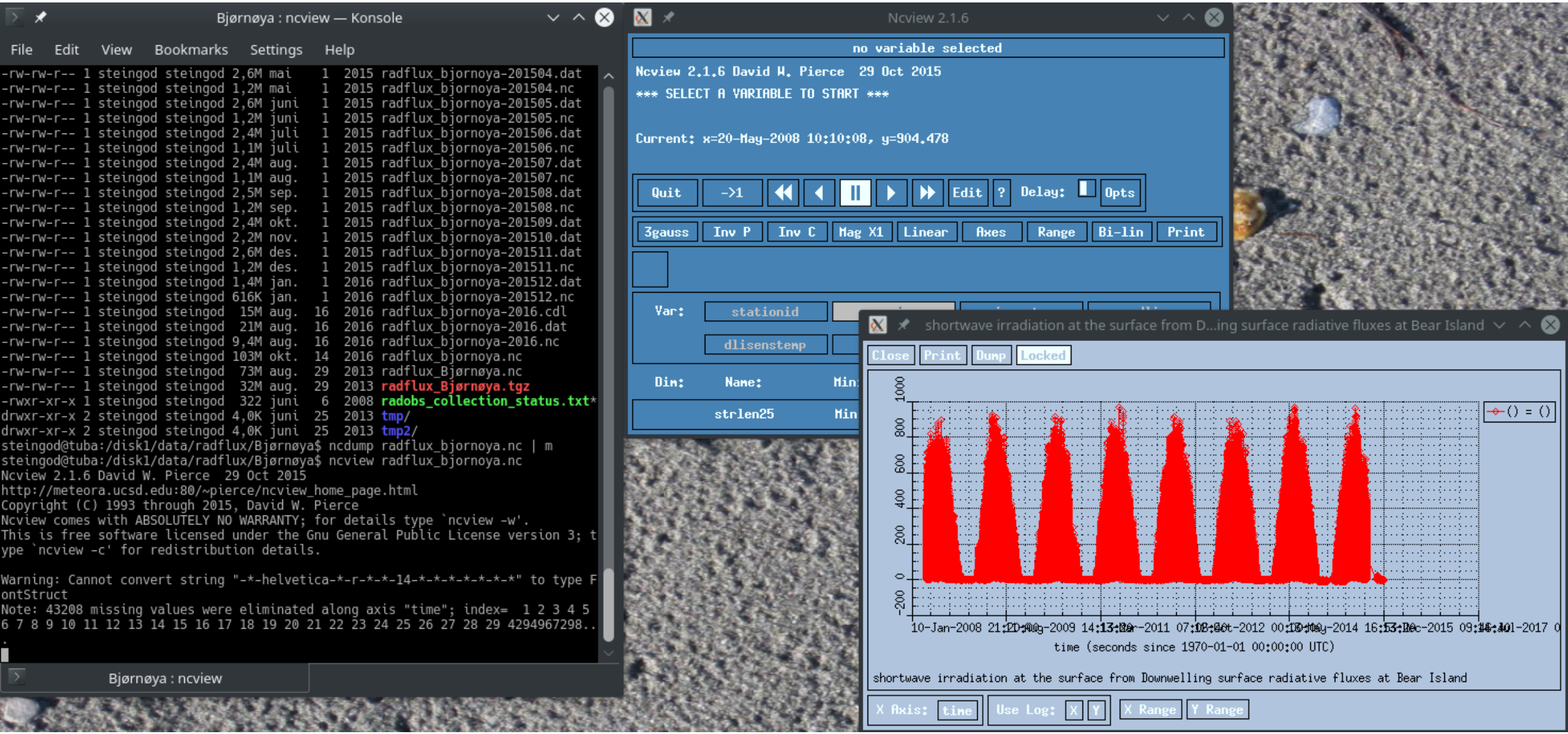
NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Courtesy
Andreas Jaunsen
NorStore





NorDataNet Norwegian Scientific Data Network GeoAcces



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Norwegian Scientific Data Network

You are testing you dataset "radflux_bjornoya-2016.nc" against CF-1.6 convention

Your dataset is not compliant with the required test(s).

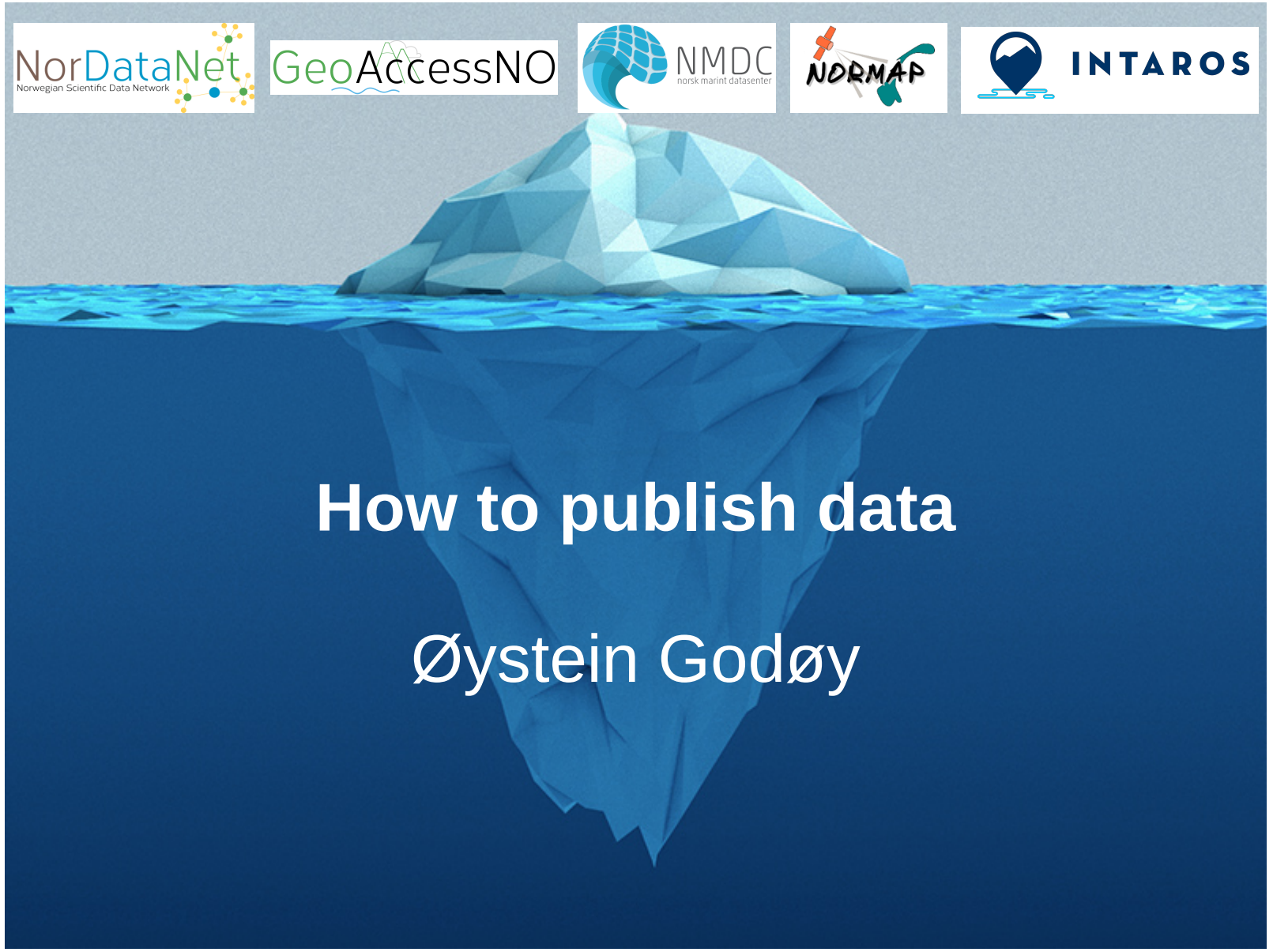
Your Dataset Scored 204 Out Of 206 Points

During the cf:1.6 check

Scoring Breakdown:

High Priority

Name	Score
§2.2 Valid netCDF data types	9/9
§2.4 Unique dimensions	9/9
§3.1 Variable battery contains valid CF units	3/3
§3.1 Variable dli contains valid CF units	3/3
§3.1 Variable dli's units are appropriate for the standard_name surface_net_downward_longwave_flux	1/1
§3.1 Variable dlisenstemp contains valid CF units	3/3
§3.1 Variable latitude contains valid CF units	3/3
§3.1 Variable latitude's units are appropriate for the standard_name latitude	1/1
§3.1 Variable longitude contains valid CF units	3/3
§3.1 Variable longitude's units are appropriate for the standard_name longitude	1/1
§3.1 Variable ssi contains valid CF units	3/3
§3.1 Variable ssi's units are appropriate for the standard_name surface_downwelling_shortwave_flux	1/1
§3.1 Variable ssisenstemp contains valid CF units	3/3
§3.1 Variable time contains valid CF units	3/3


NorDataNet
Norwegian Scientific Data Network
GeoAccessNO
NMDC
norsk marint datasenter
NORMAP
INTAROS
How to publish data
Øystein Godøy

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

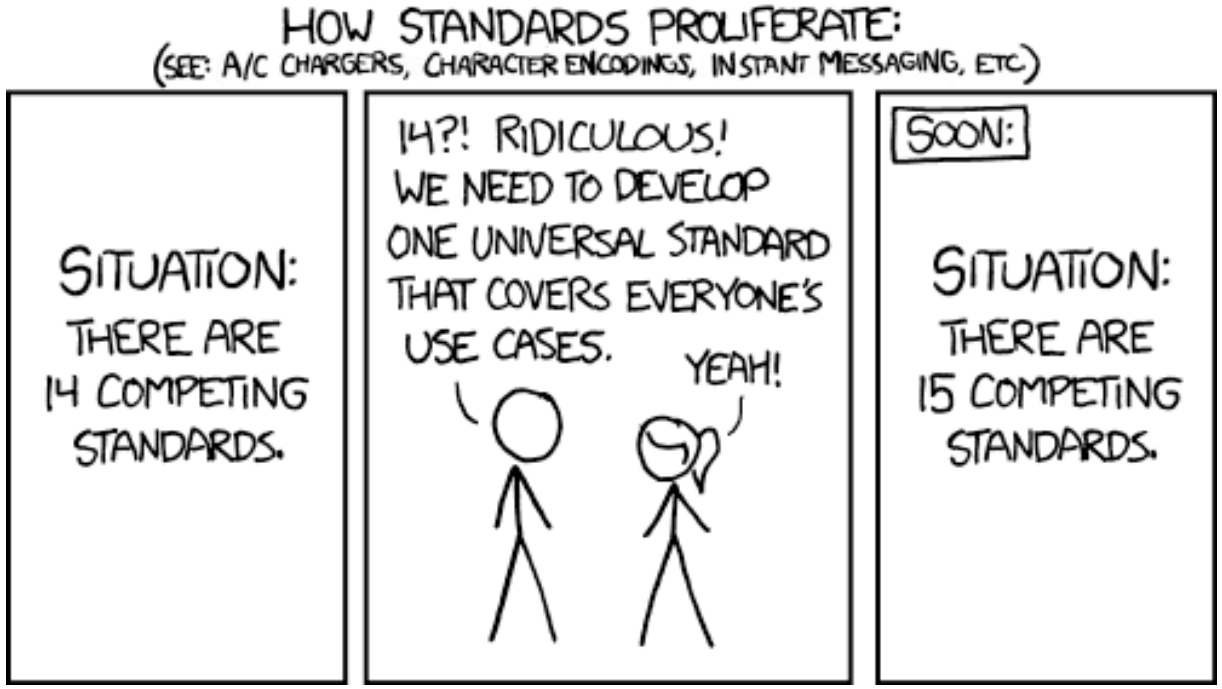
NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Data Formats: Choosing and Adopting Community Accepted Standards

- Most projects (rightly so) focus on the content of their data files, you need to consider the format as well.
- Since you captured or created the data, and stored them in your own files, you know
  - how the data are organized,
  - how to read them,
  - how to use them,
  - characteristics of the data that could constrain their use.
- The goal of a good data format is to make it easier for others to read the data too.
- Many hours have gone into developing standards for formats – try to learn from them.

Why use community standards

- If you try to develop your data format from scratch, you will forget something.
- Build on the experience and improvements built into the community standards over years of use.
- Tools and analysis software natively support reading community standard data.
- Reduce development effort and support reuse.
- Positive feedback – they are more likely to be adopted by others.



http://xkcd.com/927/

Use self describing data formats

- Self-describing data formats have become a well accepted way of archiving and disseminating scientific data.
- Before self-describing data formats became widely used, each project often invented their own data formats, often raw binary or even ASCII.
- These approaches had a number of problems:
  - Machine dependent byte ordering or floating point organizations
  - Required a 'key' to be able to open the file and read the right data.
  - A new custom reader is needed for each different data organization. Working in a new language could be very difficult since you have to redevelop the reader anew.

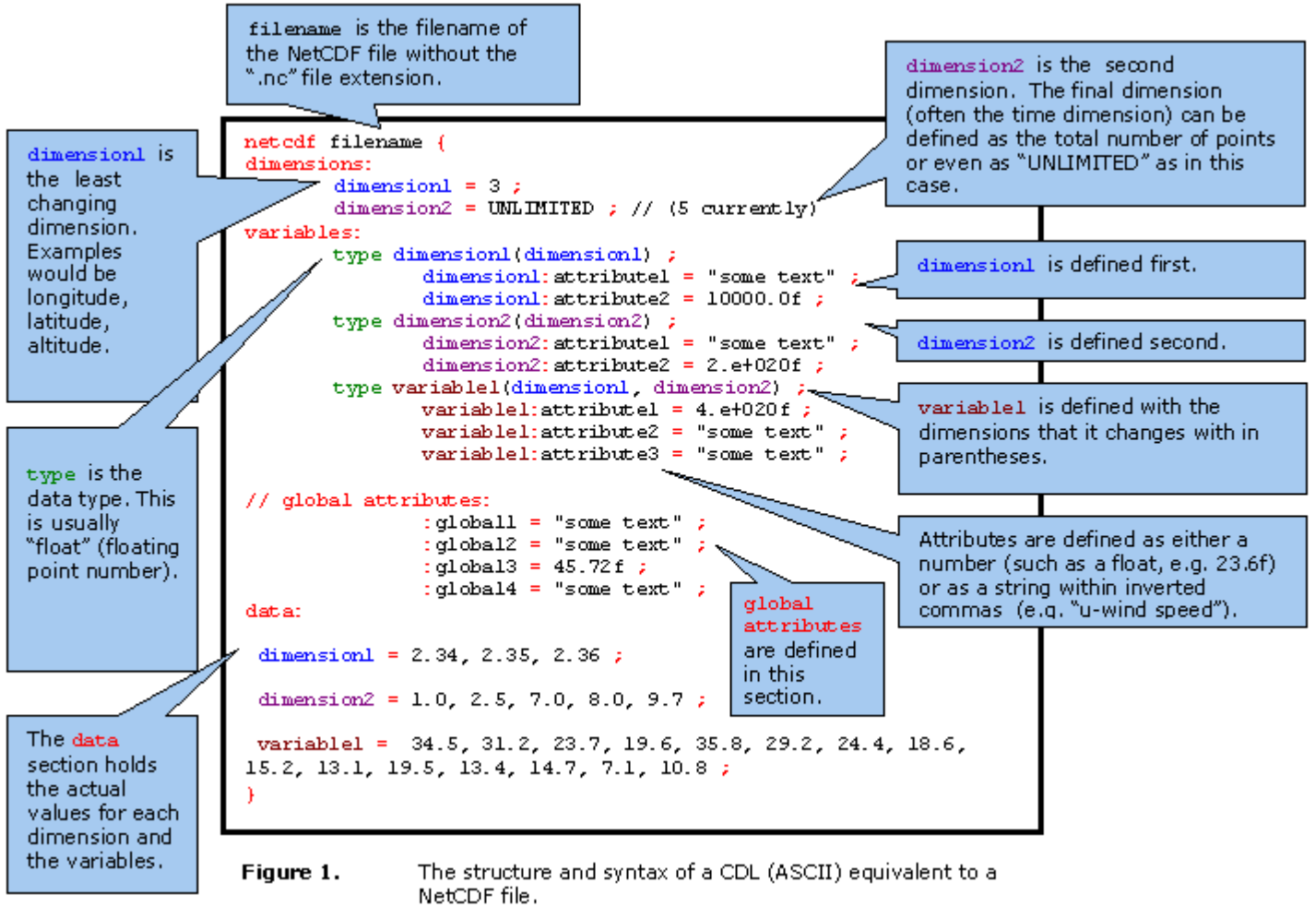
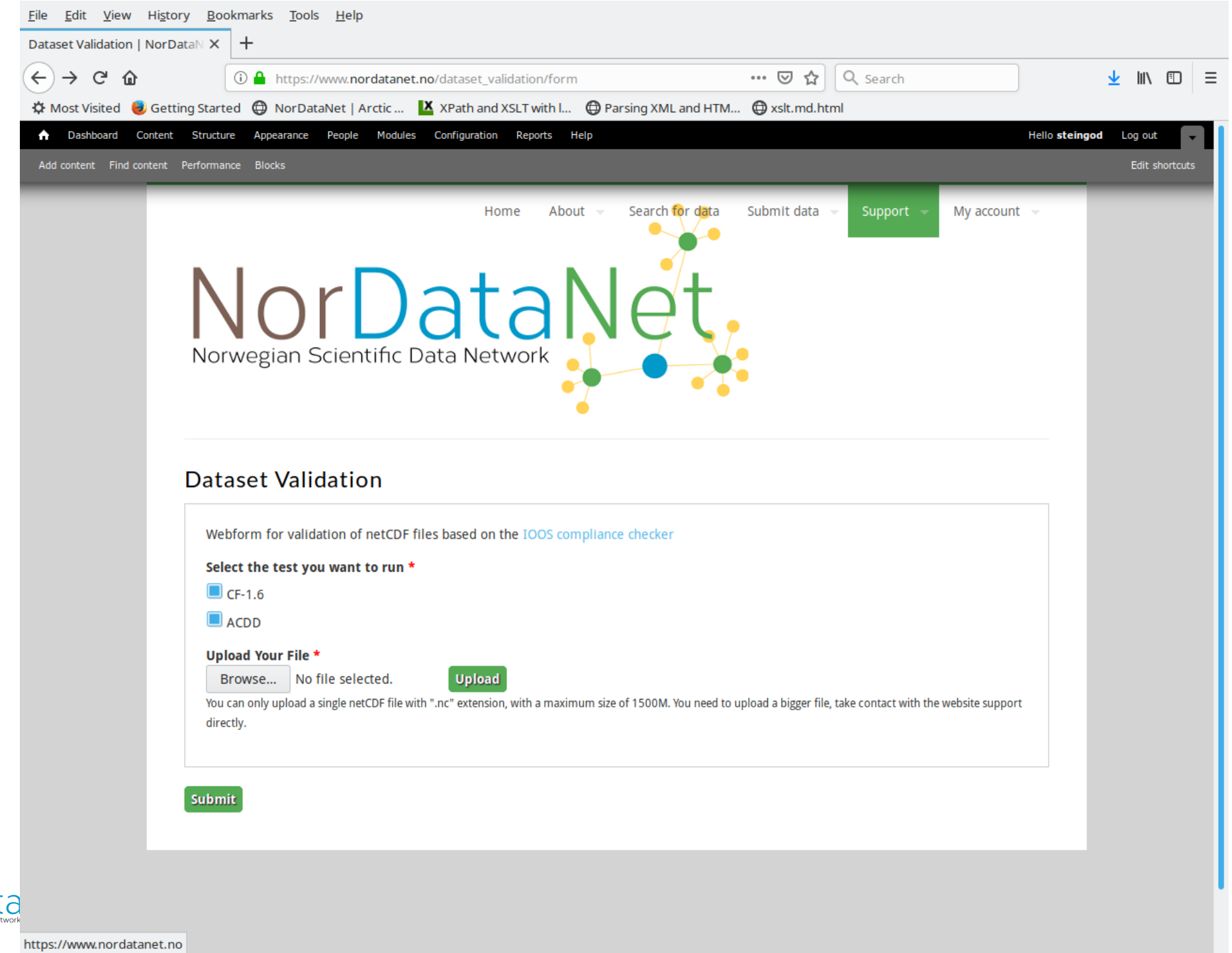


Figure 1. The structure and syntax of a CDL (ASCII) equivalent to a NetCDF file.

http://artefacts.ceda.ac.uk/formats/netcdf/index_cf.html

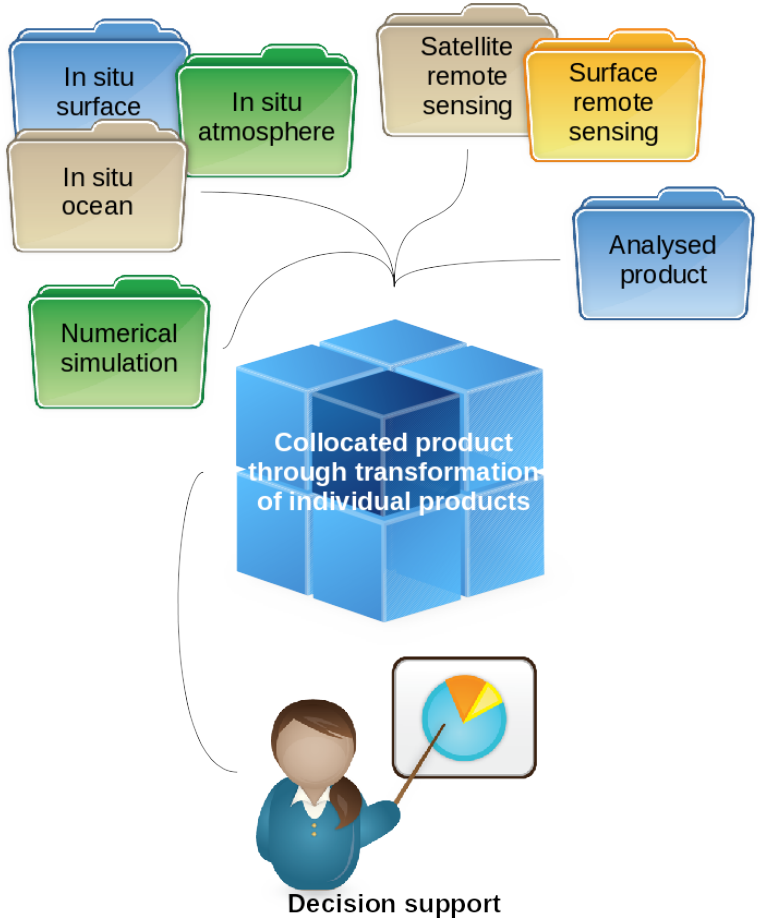

File Edit View History Bookmarks Tools Help
Dataset Validation | NorDataN
https://www.nordatanet.no/dataset_validation/form
Search
Most Visited
Getting Started
NorDataNet | Arctic ...
XPath and XSLT with l...
Parsing XML and HTM...
xslt.md.html
Dashboard Content Structure Appearance People Modules Configuration Reports Help
Hello steingod
Log out
Add content Find content Performance Blocks
Edit shortcuts
Home
About
Search for data
Submit data
Support
My account
NorDataNet
Norwegian Scientific Data Network
Dataset Validation
Webform for validation of netCDF files based on the IOOS compliance checker
Select the test you want to run *
CF-1.6
ACDD
Upload Your File *
Browse...
No file selected.
Upload
You can only upload a single netCDF file with ".nc" extension, with a maximum size of 1500M. You need to upload a bigger file, take contact with the website support directly.
Submit
https://www.nordatanet.no

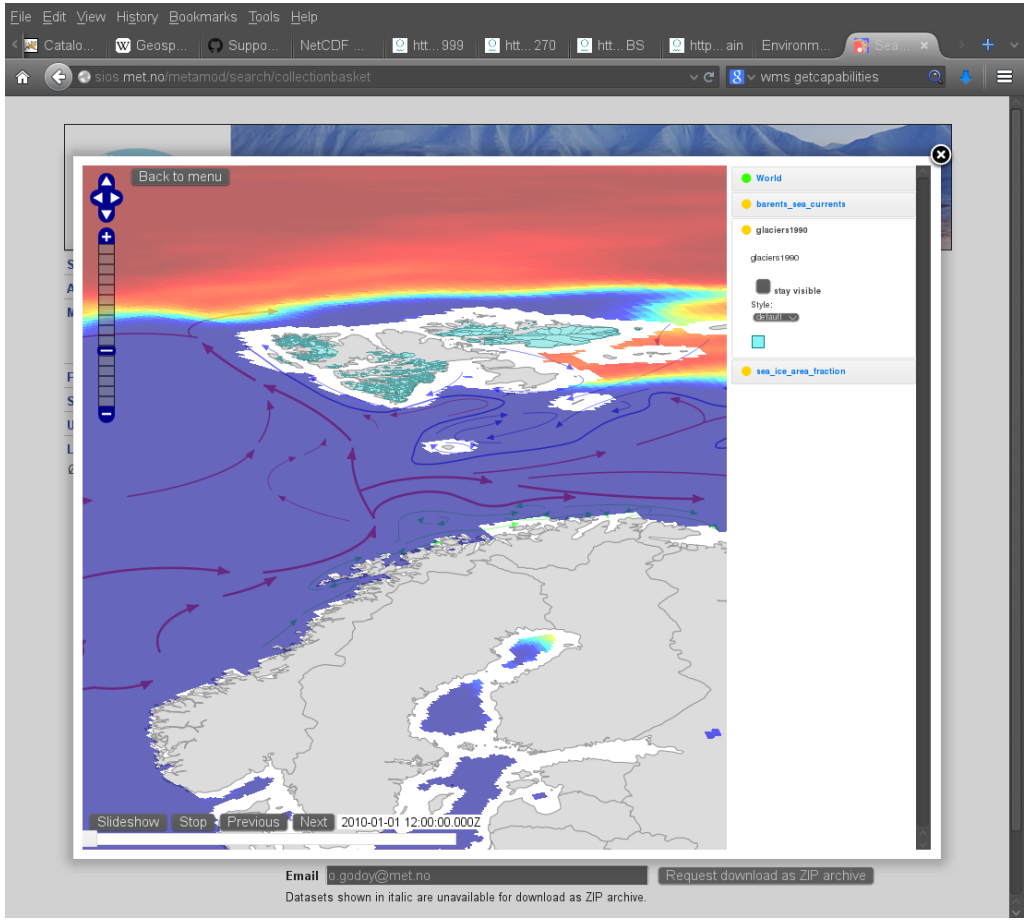
Long term data archives

DataSpan

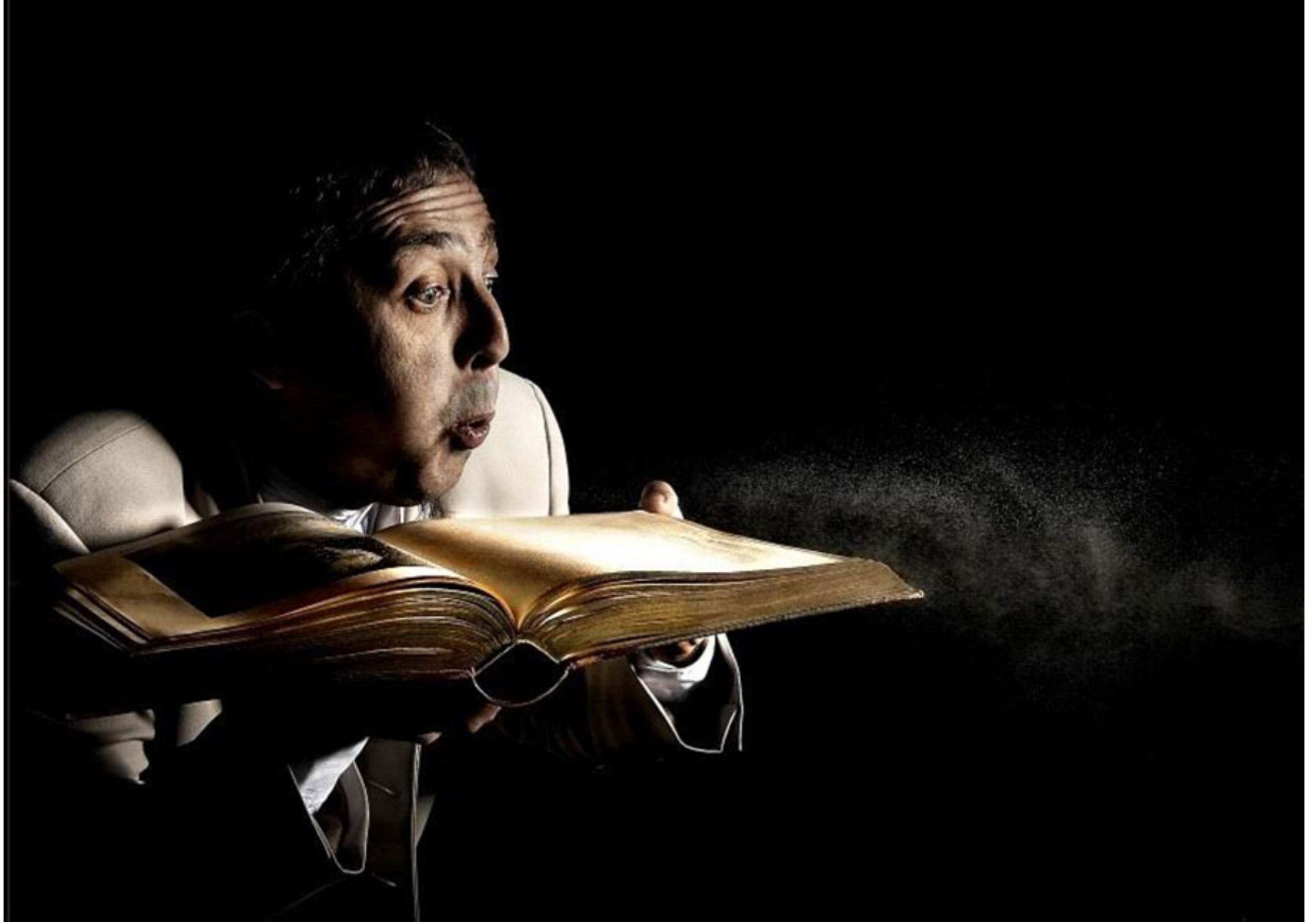
NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Combining data





NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC 2 norsk marint datasenter NORMAP INTAROS

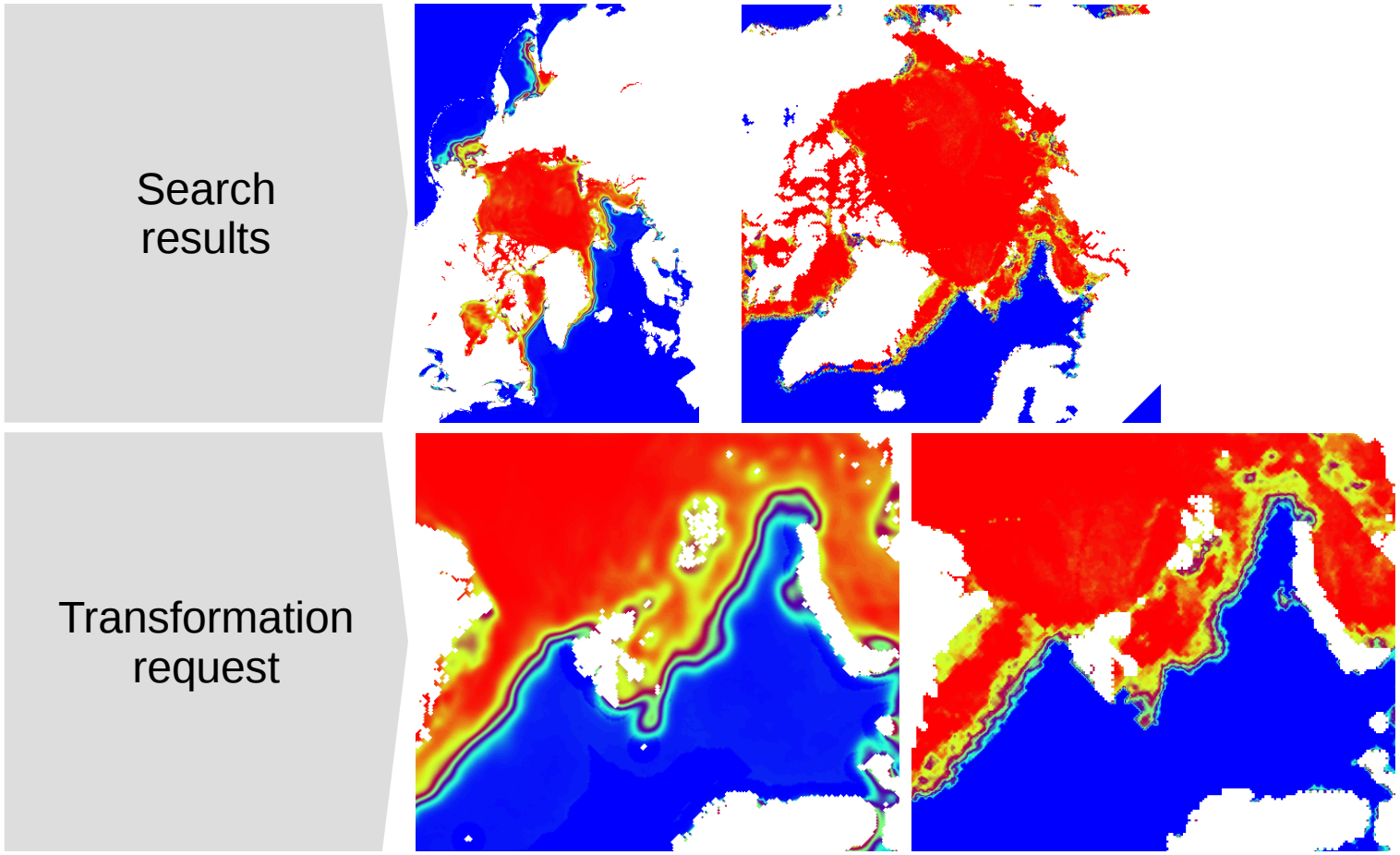


NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC 2 norsk marint datasenter NORMAP INTAROS


NorDataNet
Norwegian Scientific Data Network
GeoAccessNO
NMDC
norsk marint datasenter
NORMAP
INTAROS
Various slides
Øystein Godøy

Transformations

Transformations allow users to do
comparisons of products
and to extract
tailored products for their specific need

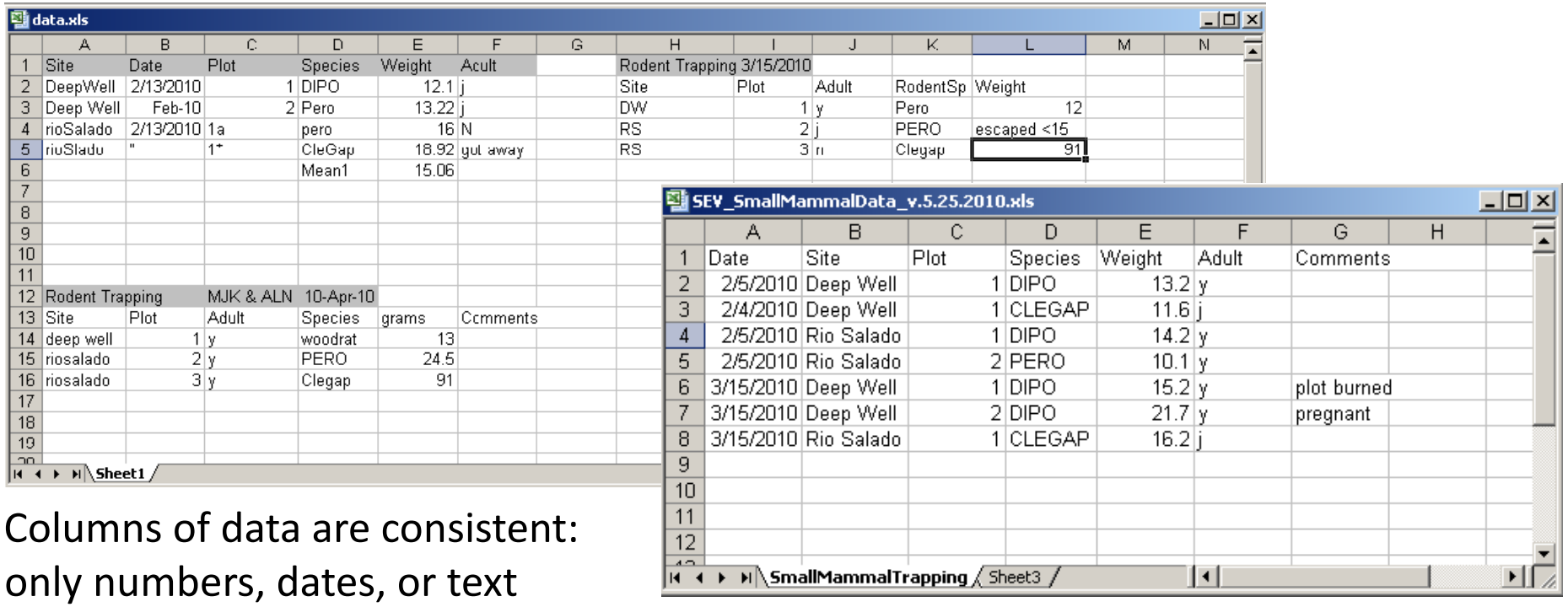


NetCDF

- NetCDF is a container you can put almost anything into
- Standardised formulations exist for
  - Gridded data
  - Timeseries
  - Trajectories
    - Including time series of trajectories
  - Profiles
    - Including time series of profiles
  - Geometries (next release)
- Can easily be served as data streams using OPeNDAP
- Can be integrated directly in tools like
  - R
  - Matlab
  - Ferret
  - Python
  - Excel
    - Check e.g. NETCDF4Excel on GitHub
- If you access data through OPeNDAP you do not have to download data

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Best Practices



data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping	3/15/2010					
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1*	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

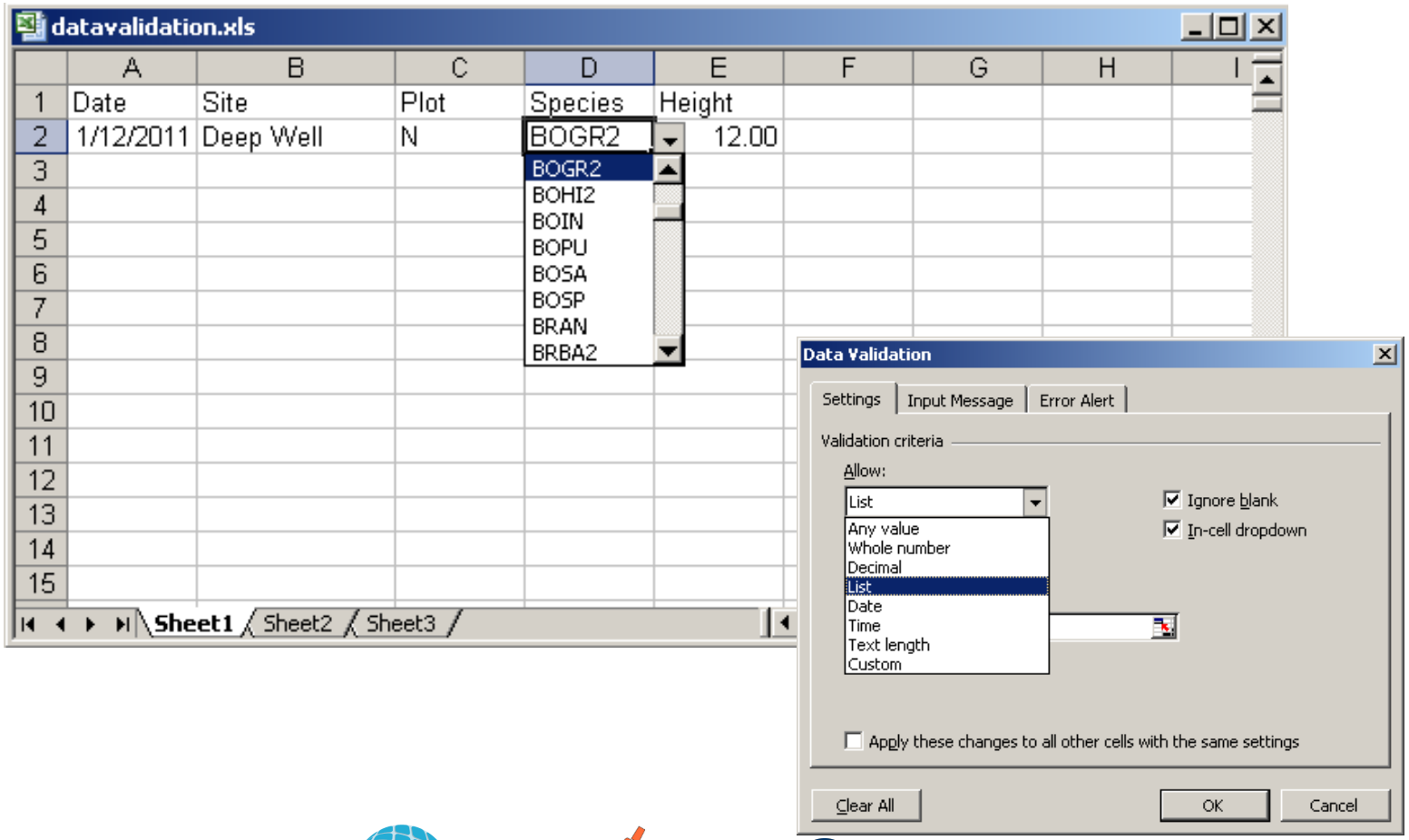
SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

SmallMammalTrapping Sheet3

- Columns of data are consistent: only numbers, dates, or text
- Consistent Names, Codes, Formats (date) used in each column
- Data are all in one table, which is much easier for a statistical program to work with than multiple small tables which each require human intervention

If you use Excel or similar



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS DataONE

C2S3 Ny-Alesund merged_170704.ods - LibreOffice Calc

File Edit View Insert Format Sheet Data Tools Window Help

Glacier	Holtedahlfonna		**Weather Condit**	5 m.s-1 wind, clear sky, sampling for C2S3 project											
Area	Ablation, Stake 2		**Coordinates/GP**	**North**		78.931367									
Date	17.04.16			**East**		13.304033									
Time	afternoon			**Altitude (m a.s.l.)**											
Observer(s)	TM, JCG, AS, EB														
General comments	110 cm pit														
	Sampling for Iso, Ions, Bio and BC														
Temperatures:															
Air temp 1m [°C]		-14.5													
Air/snow temp [°C]															
Snow/ice temp [°C]		-8.2													
Comments		stratigraphy by JCG and AS													
Layer, from top to bottom	**Top [cm]**	**Bottom [cm]**	**Depth center**	**Type 1**	**Type 2**	**iameter min [mm**	**iameter max [mm**	**Hardness**	**Hardness code**	**Depth Center [cm]**	**Density [g/cm³]**	**Depth [cm]**	**Temp [deg C]**		
1	0	1	0.5	wind packed	and sun crust	0.2	0.2	1f	3	4	136	5	-11.6		
2	1	8	4.5	wind broken		0.1	0.3	Fist	1	14	309	8	-12.1		
3	8	20	14	faceted	clusters	0.1	0.1	1f	3	23	334	10	-12.1		
4	20	26	23	rounded and faceted		0.1	0.3	pencil	4	32	381	20	-11.3		
5	26	27	26.5	melt refreeze	clusters			knife	5	39	482	26	-11.7		
6	27	35	31	faceted		0.3	0.3	pencil	4	47	475	35	-11.1		
7	35	37	36	depth hoar	clusters	1	1	4 f	2	58	306	37	-11.1		
8	37	42	39.5	faceted		0.4	0.4	pencil	4	62	394	42	-10.8		
9	42	57	49.5	faceted		0.5	0.5	1f	3	66	361	50	-10.3		
10	57	60	58.5	depth hoar		1	1	4 f	2	75	325	57	-10		
11	60	63	61.5	faceted		0.4	0.4	1f	3	82	265	60	-9.9		
12	63	71	67	melt refreeze				pencil	4	100	207	71	-9.5		
13	71	73	72	faceted		1	1	4 f	2			79	-9.3		
14	73	79	76	melt refreeze				knife	5			85	-9.1		
15	79	80	79.5	horizontal ice layer				ice	6			91	-8.7		
16	80	85	82.5	faceted		0.5	0.5	1f	3			100	-8.6		
17	85	88	86.5	depth hoar		1.5	1.5	4 f	2			108	-8.5		
18	88	91	89.5	melt refreeze				pencil	4			110	-8.2		
19	91	108	99.5	depth hoar		2	2	Fist	1						
20	108	110	109	horizontal ice layer				ice	6						

KVG1 Stake 2 | KVG 1.5 Stake 3 | KVG 2 Stake 6 | KVG 3 Stake 8 | **HTD 1 Stake 2** | HTD 2 Stake 6 | HTD 3 Stake 10 | ALB_1 | ALB_2 | ALB_3 | BC | ISOTOPE | IONS | MICROBIO | SSA

Find All | Search Formatted Display String | Match Case

Sheet 5 of 15 | PageStyle_HTD 1 Stake 2 | Sum=0 | 110%

C2S3 Ny-Alesund merged_170704.ods - LibreOffice Calc

File Edit View Insert Format Sheet Data Tools Window Help

C13 | stratigraphy by AS

Glacier	Kongsvegen		Weather Conditi	Little wind, 13 minus, almost fully clear, nice day in general, sampling for C2S3 project												
Area	Ablation, Stake 2		Coordinates/GI	North	8751951.45	78.83018										
Date	13.04.16			East	451534.753	12.75838										
Time	afternoon			Altitude (m a.s.l.)	230.728											
Observer(s)	TM, JCG, AS, EB, HH															
General comments	Little wind, partly cloudy, 20 cm snow pit, not much melt refrozen stuff, most likely all is superimposed ice															
	Sampling for Iso, Ions, Bio and BC															
Temperatures:																
Air temp 1m [°C]		-13.9														
Air/snow temp [°C]																
Snow/ice temp [°C]		-11.6														
Comments		stratigraphy by AS														
Layer, from top to bottom	**Top [cm]**	**Bottom [cm]**	**Depth center**	**Type 1**	**Type 2**	**Diameter min [mm]**	**Diameter max [mm]**	**Hardness**	**Hardness code**	**Depth Center [cm]**	**Density [g/cm³]**	**Depth [cm]**	**Temp [deg C]**			
1	0	1	0.5	wind packed		0.1	0.1	pencil	4	2	147	1	-12.6			
2	1	5	3	wind broken		0.2	0.3	1 finger	3	10	318	5	-12.4			
3	5	16	10.5	clustered rounded		0.3	0.4	pencil	4	18	224	16	-11.6			
4	16	20	18	faceted				1 finger	4			20	-11.6			

Preferably do not split content over columns and cells

Does this belong under general comments?

Use full names

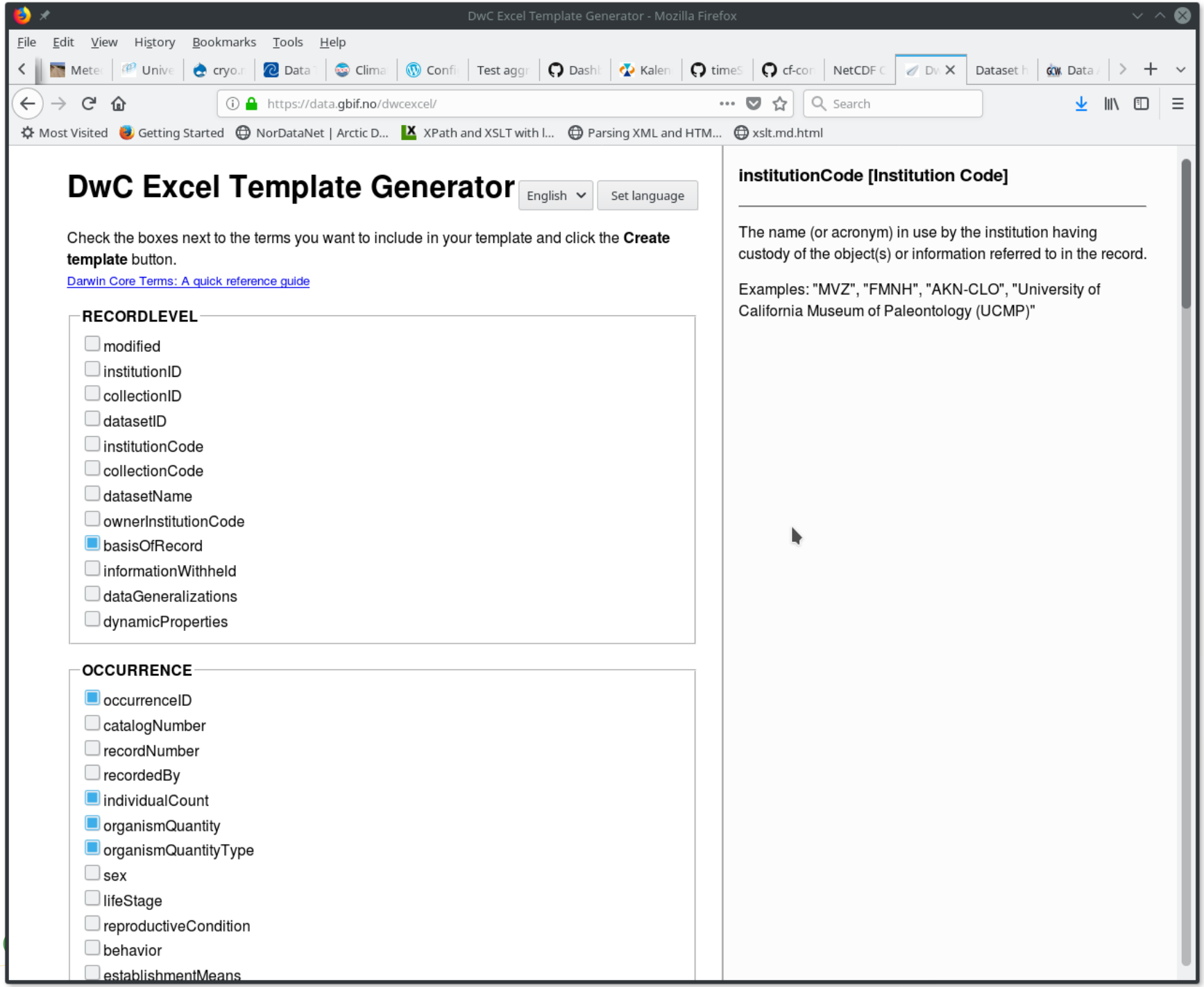
What is type? Explain vocabularies.

How to interprete, which standard?

KVG1 Stake 2 | KVG 1.5 Stake 3 | KVG 2 Stake 6 | KVG 3 Stake 8 | HTD 1 Stake 2 | HTD 2 Stake 6 | HTD 3 Stake 10 | ALB_1 | ALB_2 | ALB_3 | BC | ISOTOPE | IONS | MICROBIO | SSA

Find All | Search Formatted Display String | Match Case

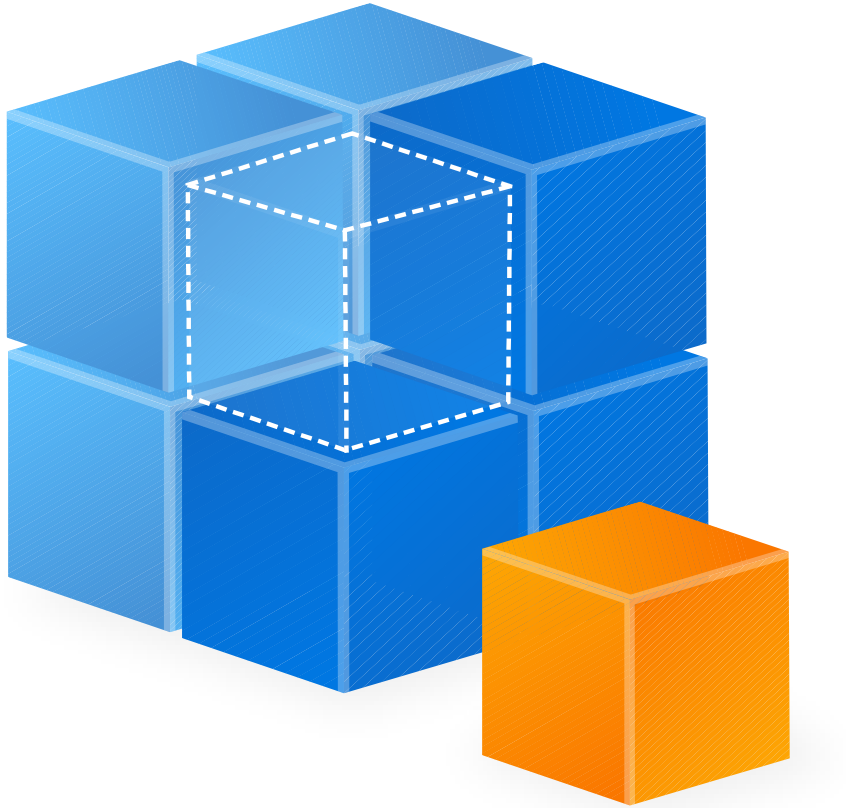
Sheet 1 of 15 | PageStyle_KVG1 Stake 2 | Sum=0 | 110%



NorDataNet
Norwegian Scientific Data Network

Specific Arctic challenges

- Monitoring by both scientific and operational communities
  - Collaboration needed
- Scientific communities good a archiving
- Operational communities good at real time exchange and combining efforts
- Bridging needed



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

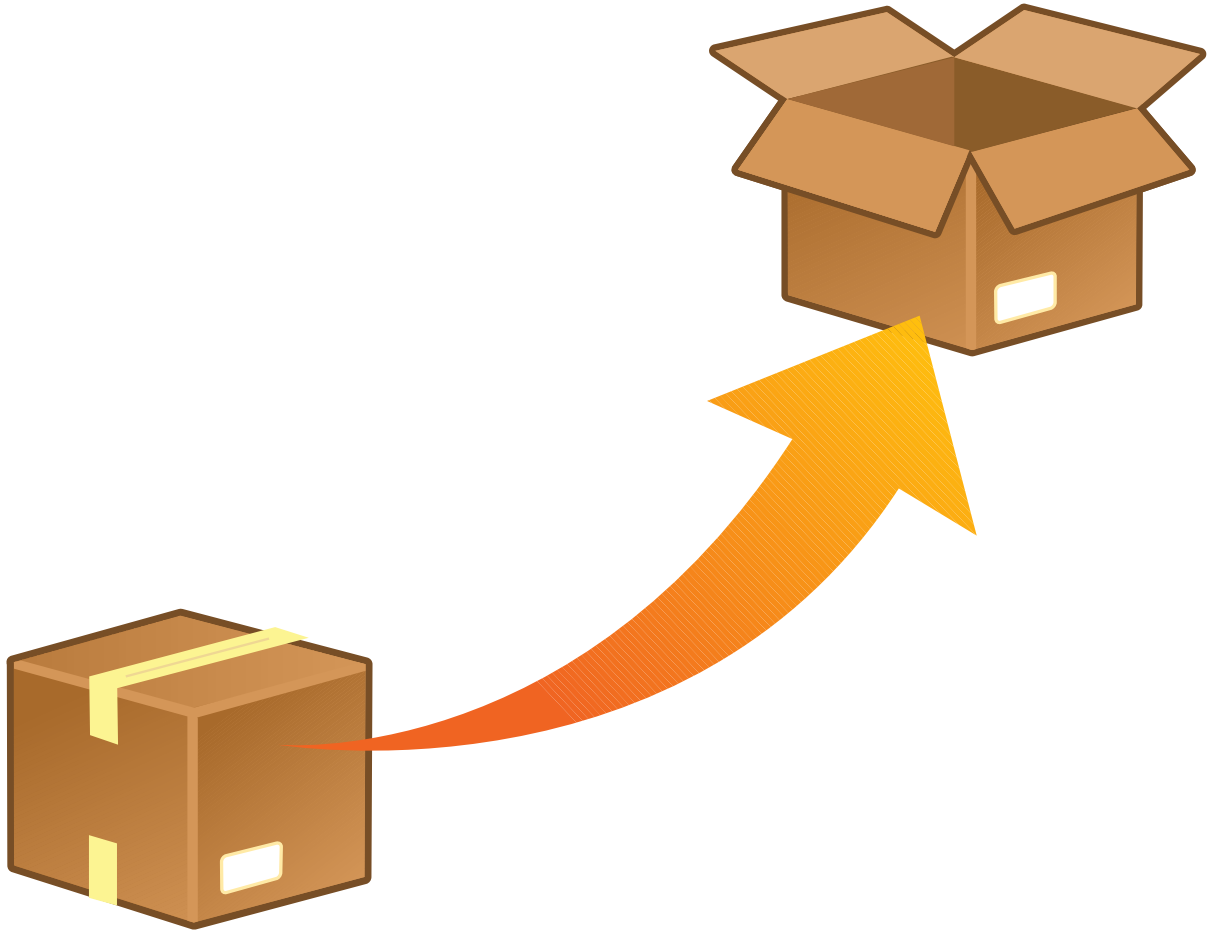
Summary

- Document and share data
- Use standards
  - And contribute
- Track usage
  - Across communities

NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

Issues for discussion

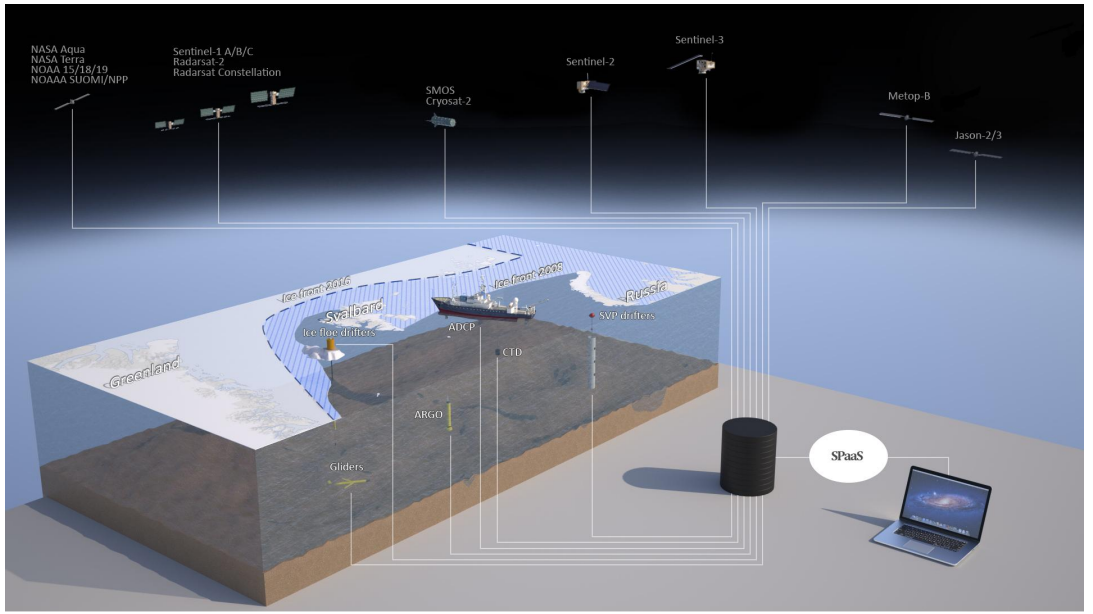
- Sharing data
  - Bottlenecks sharing data in a proper form?
  - Gaps in education?
  - Mitigation?
- Experience using data generated by others?
  - Bottlenecks and mitigation
- Standardisation
  - Necessary?
  - How to navigate?



NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS

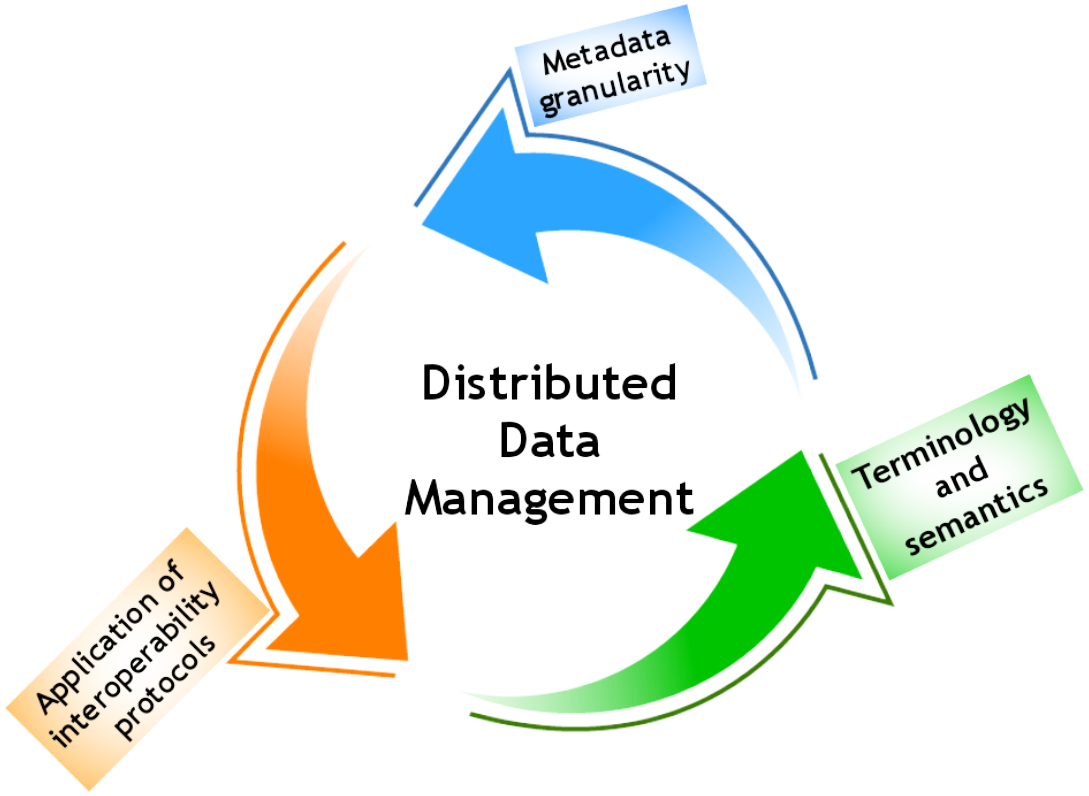
Moving towards

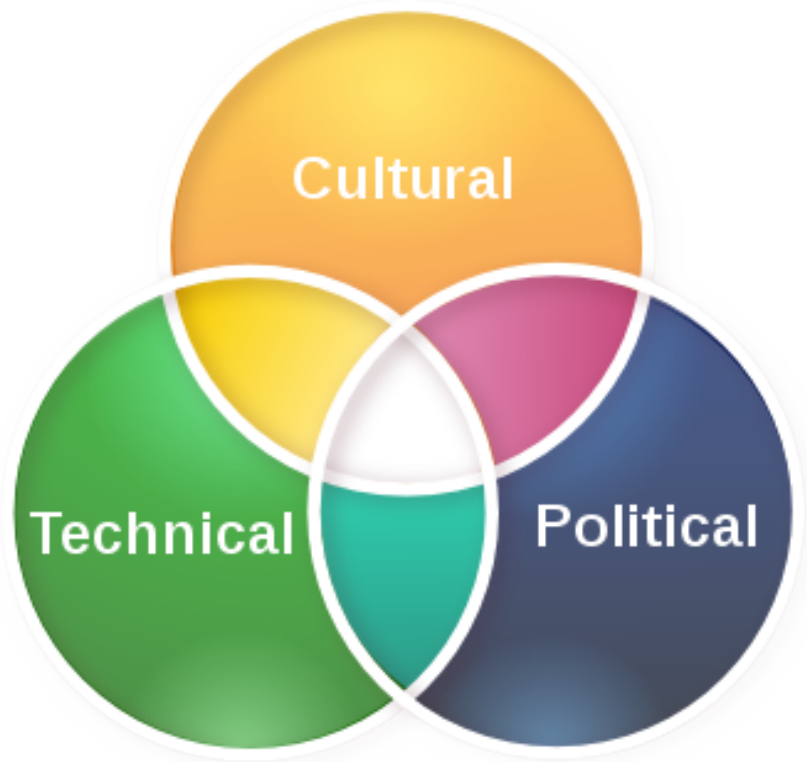
- Data management required by funding agencies
- Integration of data centres
- Work flow management
- Funding agency requirements
  - Projects must have a data plan
  - Data underlying scientific publications have to be open
  - Data plan (DCC)
    - Data summary
    - FAIR data
      - Making data findable, including provisions for metadata
      - Making data openly accessible
      - Making data interoperable
      - Increase data re-use
    - Allocation of resources
    - Data security
    - Ethical aspects
- Scientific Platforms
  - European Open Science Cloud



Courtesy of Morten W. Hansen, NERSC

Challenges





NorDataNet Norwegian Scientific Data Network GeoAccessNO NMDC norsk marint datasenter NORMAP INTAROS