INTER ⇄ ACT

WP4: Data Forum

Øystein Godøy, Boris Radosavljević,
Boris Biskaborn, Anna Irrgang

ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR-
UND MEERESFORSCHUNG

Norwegian
Meteorological
Institute

Motivation

- INTERACT research stations generate data and metadata
  - Long term monitoring
  - Short term process studies
  - External data by individual scientists/ groups
- Research stations archive data and metadata (internal and external)
  - e.g. meteorological data
  - photos, maps, reports etc.
  - list of data acquired at the stations
  - information on data collection procedures (field diaries)

Motivation

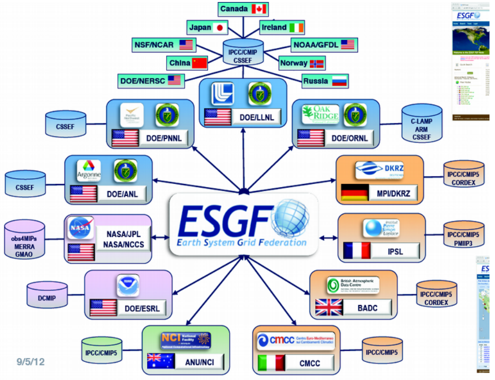
- Gain
  - Interoperability between Arctic station data management and accessibility of metadata and data
  - Increased visibility
- Comply with H2020 requirement for open data access
- Avoid
  - Redundancy of activities
    - unless specifically wanted
  - Loss of data

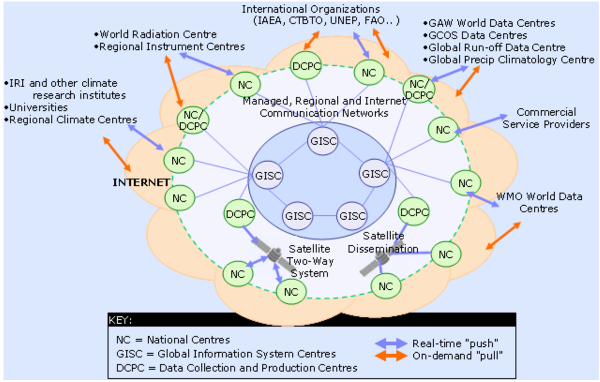
Objectives

- Analyse & identify
  - Current status and potential approaches to unified data management plan and system
  - Identify synergies with external activities
- Mitigate
  - Step by step in a prioritised implementation
  - Basic principles outlined in a data management plan
  - Working with the community through the INTERACT Data Forum
- Establish a demonstrator catalogue
  - Of available datasets
    - Go for the “easy wins” first
  - Through “INTERACT best practises”
- Link research stations
  - with observation networks and data repositories
  - retaining station identity

Visibility

- Regional and global data management frameworks
  - GEO
  - INSPIRE
  - SAON/IASC Arctic Data Committee
  - WMO Information System
  - WMO Integrated Global Observing System
  - ICSU Word Data System
  - ...



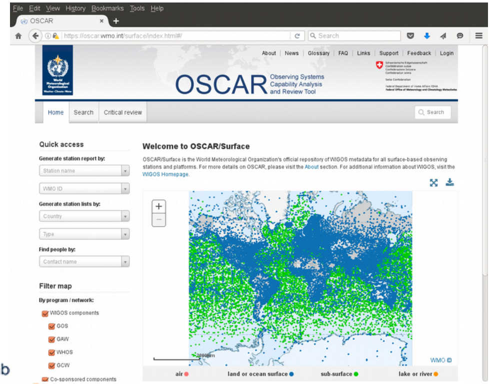


International Organizations
(IAEA, CTBTO, UNEP, FAO..)
•World Radiation Centre
•Regional Instrument Centres
•GAW World Data Centres
•GCOS Data Centres
•Global Run-off Data Centre
•Global Precip Climatology Centre
•IRI and other climate research institutes
•Universities
•Regional Climate Centres
Commercial Service Providers
INTERNET
Managed, Regional and Internet Communication Networks
WMO World Data Centres
Satellite Two-Way System
Satellite Dissemination
NC
DCPC
NC/DCPC
GISC

KEY:

NC = National Centres
GISC = Global Information System Centres
DCPC = Data Collection and Production Centres

Real-time "push"
On-demand "pull"



File Edit View History Bookmarks Tools Help
OSCAR
https://oscar.wmo.int/surface/index.html#/
Search
About | News | Glossary | FAQ | Links | Support | Feedback | Login
OSCAR Observing Systems Capability Analysis and Review Tool
Home
Search
Critical review
Search

Quick access

Generate station report by:
Station name
WMO ID

Generate station lists by:
Country
Type

Find people by:
Contact name

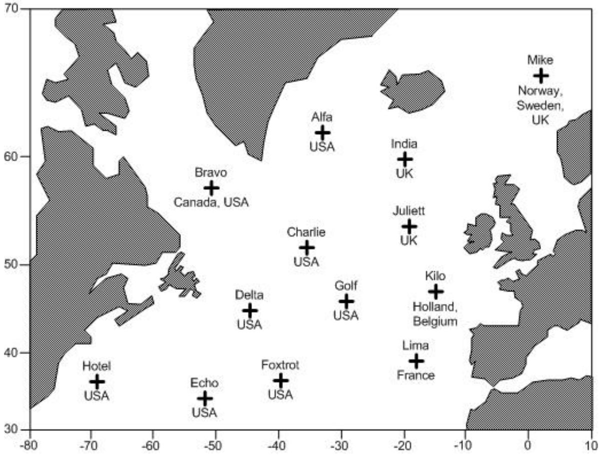
Filter map

By program / network:
☑ WIGOS components
☑ GOS
☑ GAW
☑ WHOS
☑ GCW
☑ Co-sponsored components

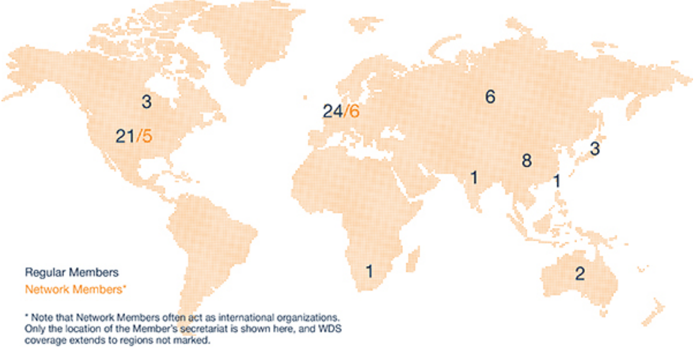
Welcome to OSCAR/Surface

OSCAR/Surface is the World Meteorological Organization's official repository of WIGOS metadata for all surface-based observing stations and platforms. For more details on OSCAR, please visit the About section. For additional information about WIGOS, visit the WIGOS Homepage.

air ● land or ocean surface ● sub-surface ● lake or river ●



WDS Regular and Network Memb



Regular Members
Network Members*

* Note that Network Members often act as international organizations. Only the location of the Member's secretariat is shown here, and WDS coverage extends to regions not marked.

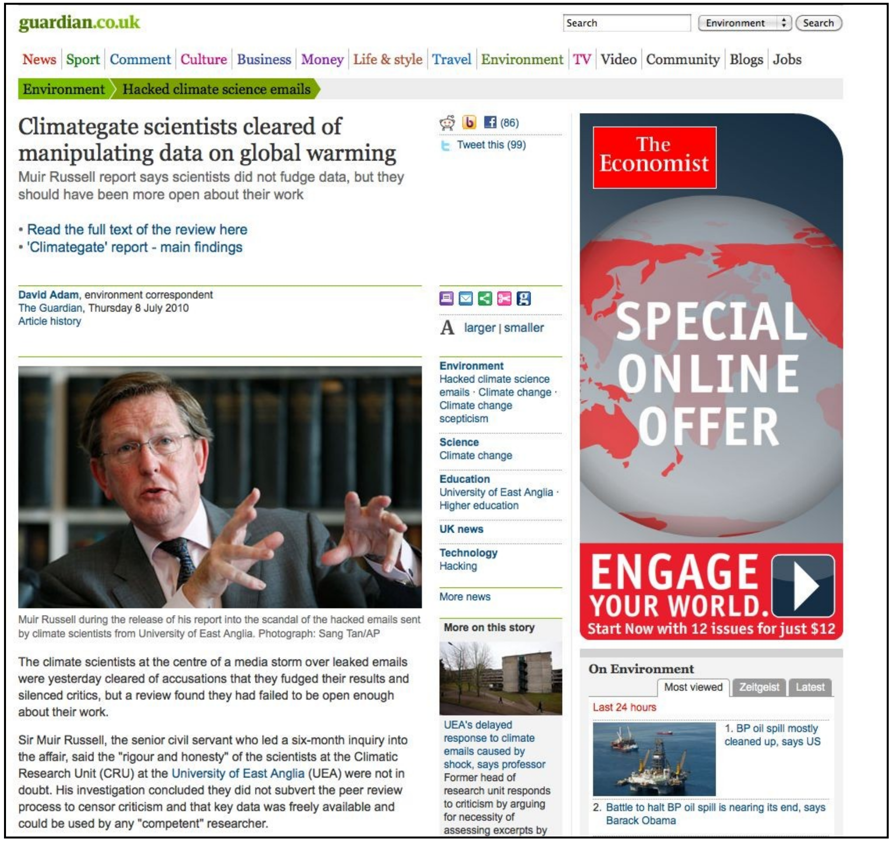
Strategy

- Initially focus on discovery metadata
  - Ensure interoperability and information flow/streams
  - To establish a unified view of the INTERACT data space
  - Expose this to external frameworks (metadata standards)
- Move to interoperability at the data level when data discovery is working
  - Interoperability at the data level is required for interaction with larger frameworks
  - Bundling of similar data is required to be relevant for CalVal activities in larger programmes
  - While retaining the visibility of stations and scientists
- Develop guidance material
  - For stations
  - For scientists
  - Based on existing efforts within disciplines, RDA, ICSU, WMO etc
- Improve visibility and relevance of Interact for e.g. WMO and SAON activities
  - Through interaction with the Arctic Data Committee
  - Being pragmatic, working step by step, towards a long term vision

Deliverables and Milestones

- D4.1: Data Management Plan (Month 6)
- D4.2: Report on current data flows (Month 12)
    - Gap analysis and bottlenecks
- D4.3: Field guide to data repositories (Month 24)
    - Mentoring of potential providers of TA virtual access
- D4.4: Data Policy (Month 24)

Importance of data management


guardian.co.uk
Search
Environment
Search
News | Sport | Comment | Culture | Business | Money | Life & style | Travel | Environment | TV | Video | Community | Blogs | Jobs
Environment
Hacked climate science emails
(86)
Tweet this (99)
A larger | smaller
Environment
Hacked climate science emails · Climate change · Climate change scepticism
Science
Climate change
Education
University of East Anglia · Higher education
UK news
Technology
Hacking
More news
More on this story
On Environment
Most viewed
Zeitgeist
Latest
Last 24 hours

Climategate scientists cleared of manipulating data on global warming

Muir Russell report says scientists did not fudge data, but they should have been more open about their work

- Read the full text of the review here
- 'Climategate' report - main findings

David Adam, environment correspondent
The Guardian, Thursday 8 July 2010
Article history

Muir Russell during the release of his report into the scandal of the hacked emails sent by climate scientists from University of East Anglia. Photograph: Sang Tan/AP

The climate scientists at the centre of a media storm over leaked emails were yesterday cleared of accusations that they fudged their results and silenced critics, but a review found they had failed to be open enough about their work.

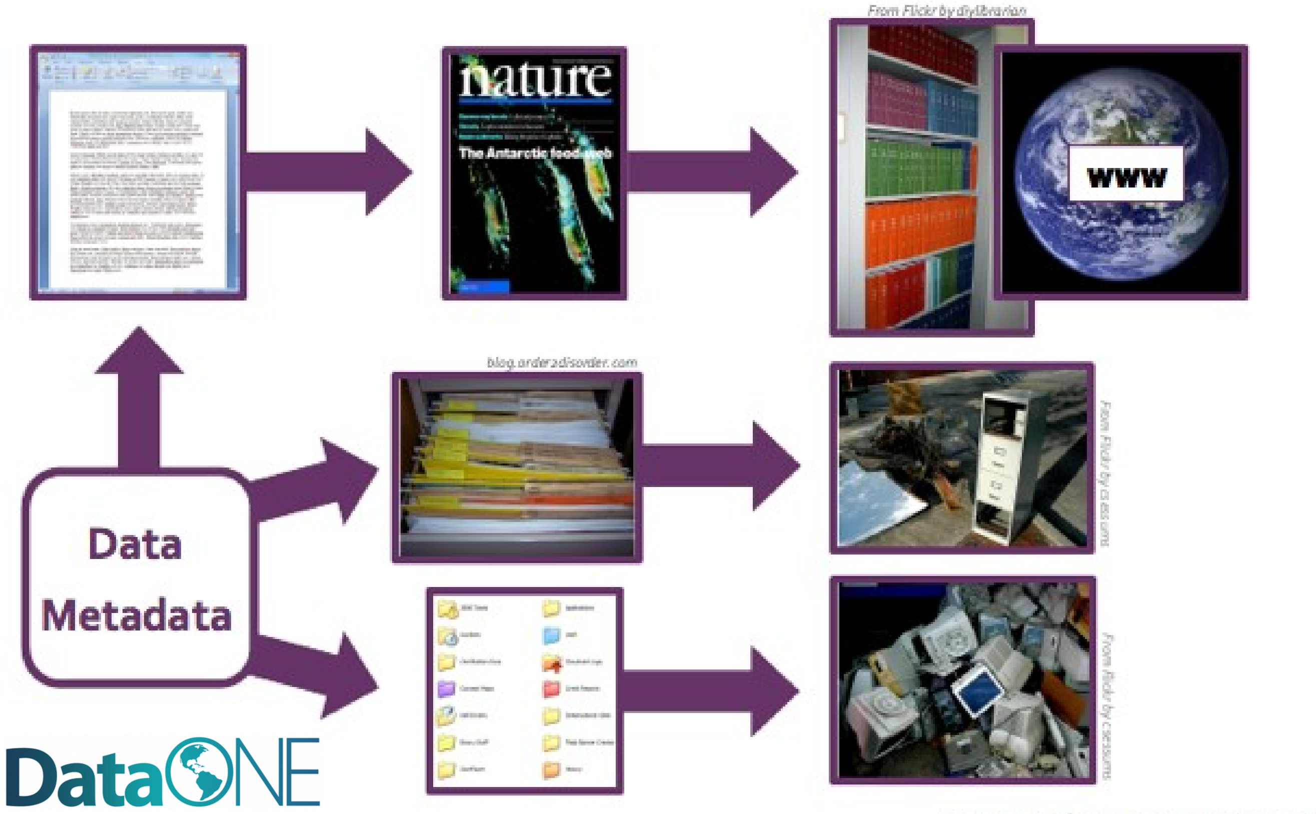
Sir Muir Russell, the senior civil servant who led a six-month inquiry into the affair, said the "rigour and honesty" of the scientists at the Climatic Research Unit (CRU) at the University of East Anglia (UEA) were not in doubt. His investigation concluded they did not subvert the peer review process to censor criticism and that key data was freely available and could be used by any "competent" researcher.

UEA's delayed response to climate emails caused by shock, says professor
Former head of research unit responds to criticism by arguing for necessity of assessing excerpts by

1. BP oil spill mostly cleaned up, says US
2. Battle to halt BP oil spill is nearing its end, says Barack Obama

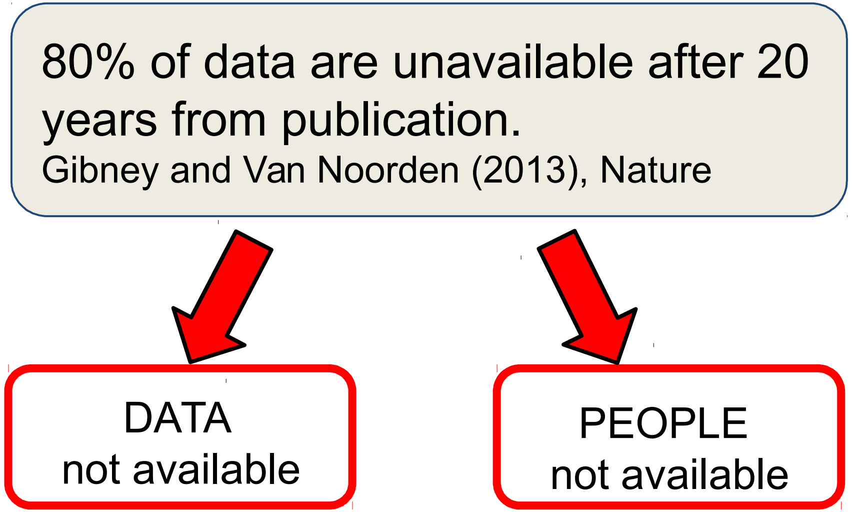
The climate scientists at the centre of a media storm over leaked emails were yesterday cleared of accusations that they fudged their results and silenced critics, but a review found they had failed to be open enough about their work.

The reality today



Recreated from Klump et al. 2006

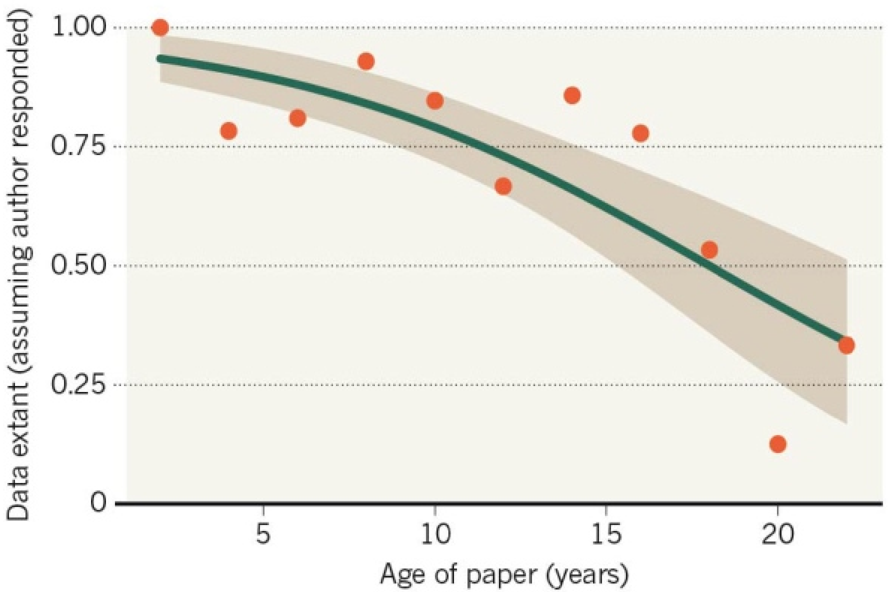
Data Management



80% of data are unavailable after 20 years from publication.
Gibney and Van Noorden (2013), Nature

DATA
not available

PEOPLE
not available

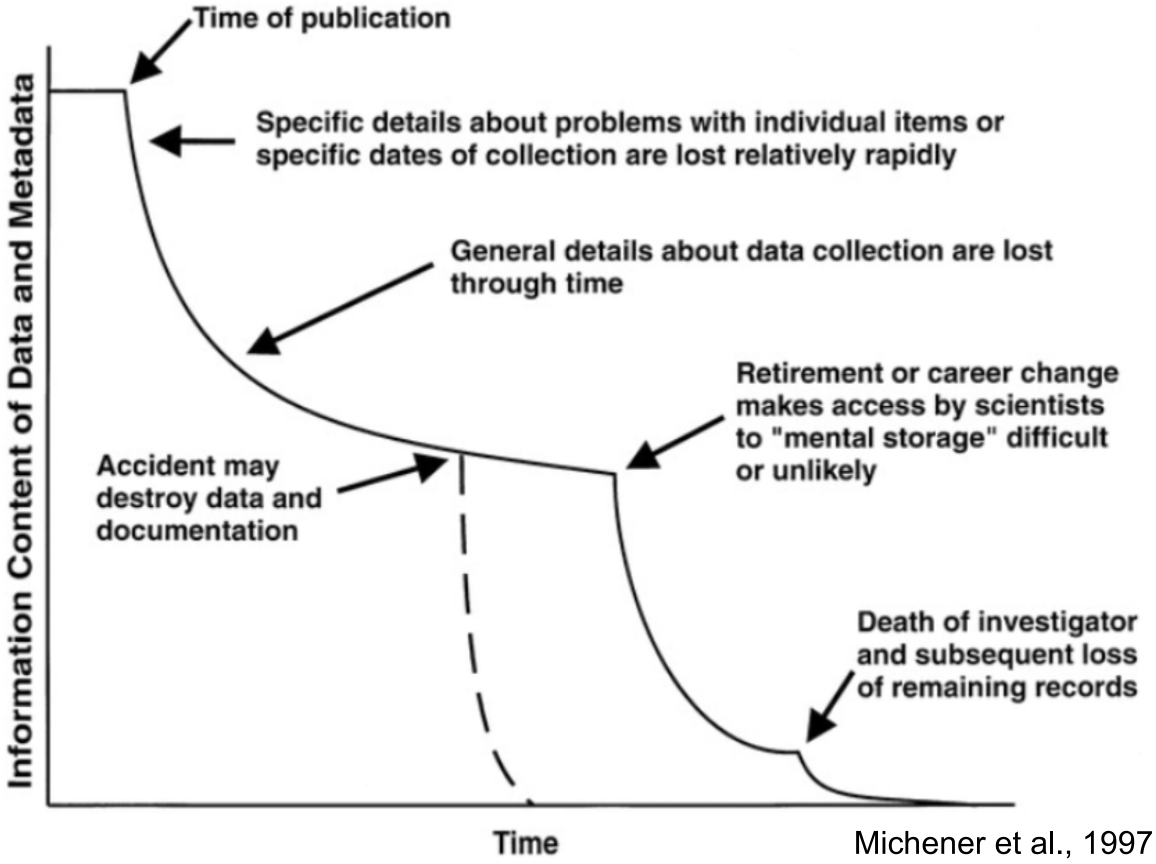


MISSING DATA

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

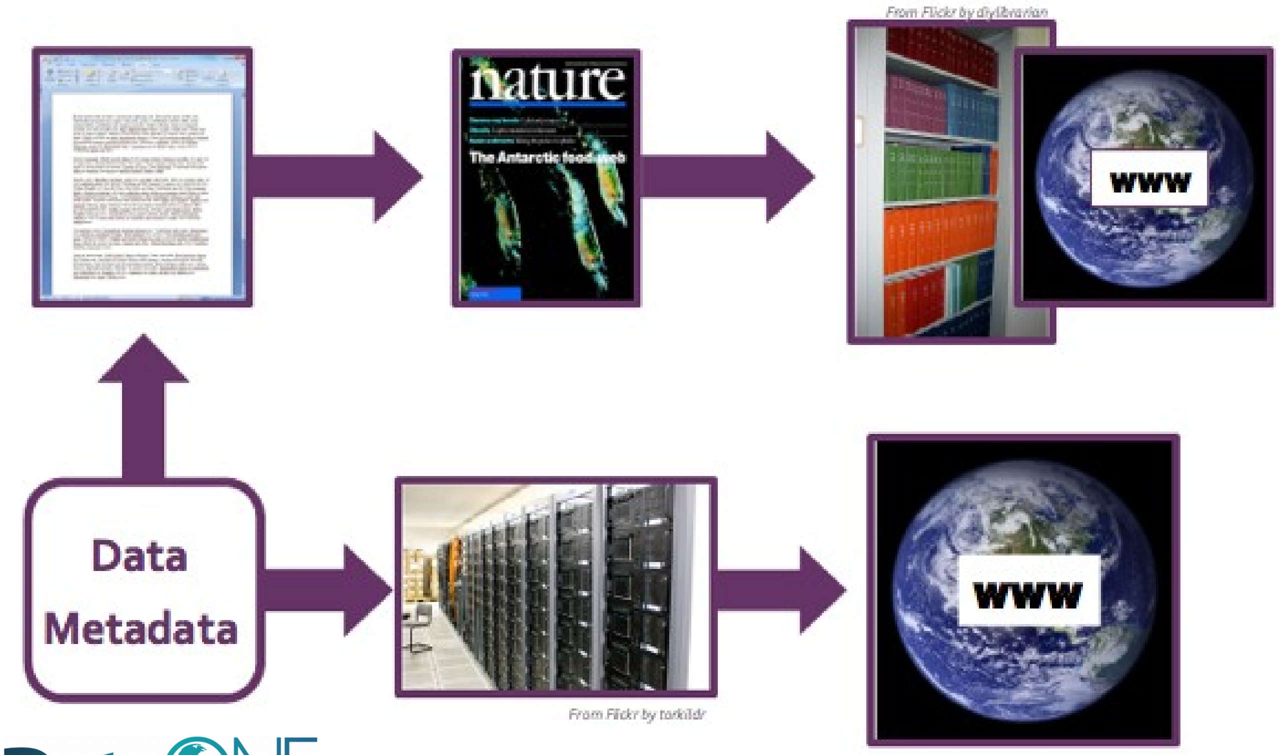
http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416

Poor data practice results in loss of information


Time of publication
Specific details about problems with individual items or specific dates of collection are lost relatively rapidly
General details about data collection are lost through time
Retirement or career change makes access by scientists to "mental storage" difficult or unlikely
Accident may destroy data and documentation
Death of investigator and subsequent loss of remaining records
Information Content of Data and Metadata
Time

Michener et al., 1997, Ecological Applications, 7(1)

The vision for the future



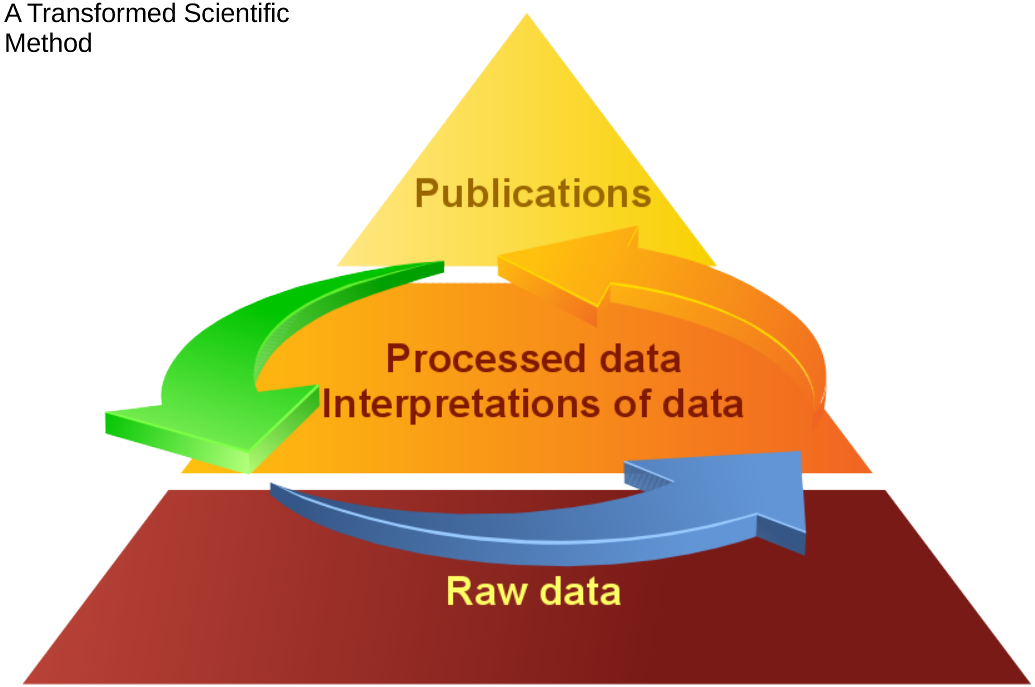
From Flickr by diylibrarian

From Flickr by tarkildr

Recreated from Klump et al. 2006

All scientific data online

Source: Jim Gray on eScience:
A Transformed Scientific Method

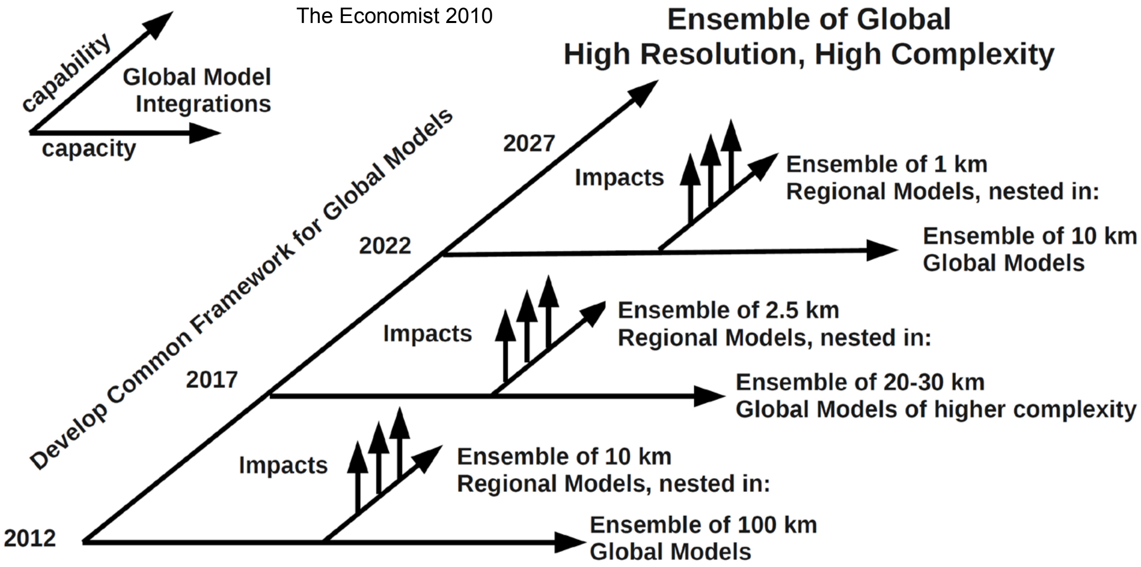


- Many disciplines overlap and use data from other sciences
- Internet can unify data, software and literature
- Go from literature to computation to data back to literature
- Information is at your fingertips for everyone and everywhere
- Potentially Increased Scientific Information Velocity
- Potentially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

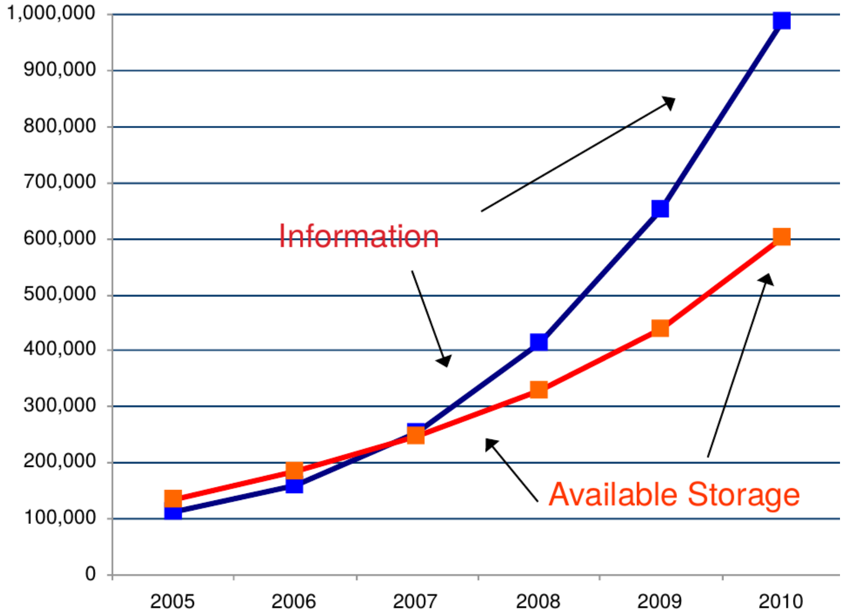
Data deluge

The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

The Economist 2010



Source: John Gantz, IDC Corporation: The Expanding Digital Universe



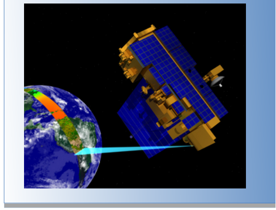
New science

eBird

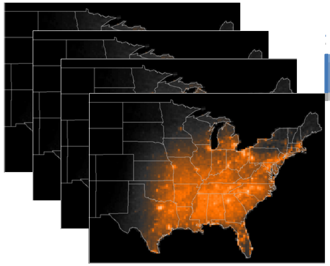
Land Cover

Meteorology

MODIS – Remote sensing data



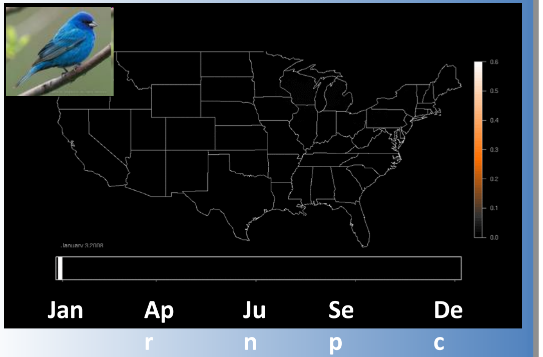
TeraGrid



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

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 supercomputing to determine bird migration routes for particular species.

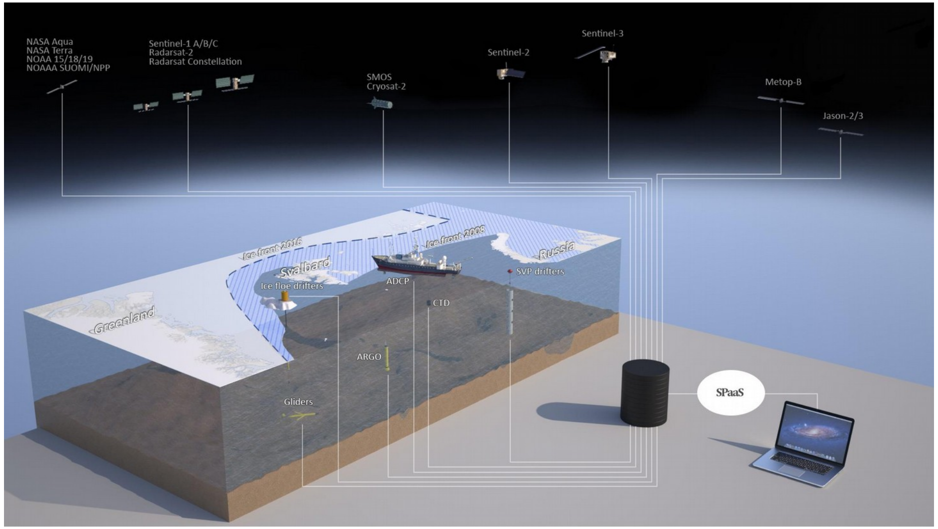
Why bother with structured data management?

- Maximise public investment in data collection and production
- Promote scientific collaboration
- Promote interdisciplinary science
- Promote scientific transparency
- Leave a legacy
- Increase the available material
  - Example from Life Sciences
    - Barends Mons
    - http://confdados.rcaap.pt/wp-content/uploads/2016/09/ConfDados_Barend_Mons.pdf
    - Only 12% of NIH funded datasets are demonstrably deposited in recognized repositories: so over 200,000 ‘invisible’ public datasets can not be re-used effectively.
    - Approximately 50% of funded research not reproducible
    - Prohibitive for scaling effective knowledge discovery

- Science paradigms
  - according to Jim Gray
- – empirical science
  - 1000 years ago
- – theoretical science
  - 200 years ago
- – computational science
  - 20 years ago
- – data exploration science
  - today

Moving towards

- Data management required by funding agencies
- Integration of data centres
- Work flow management
- Scientific Platforms
  - European Open Science Cloud

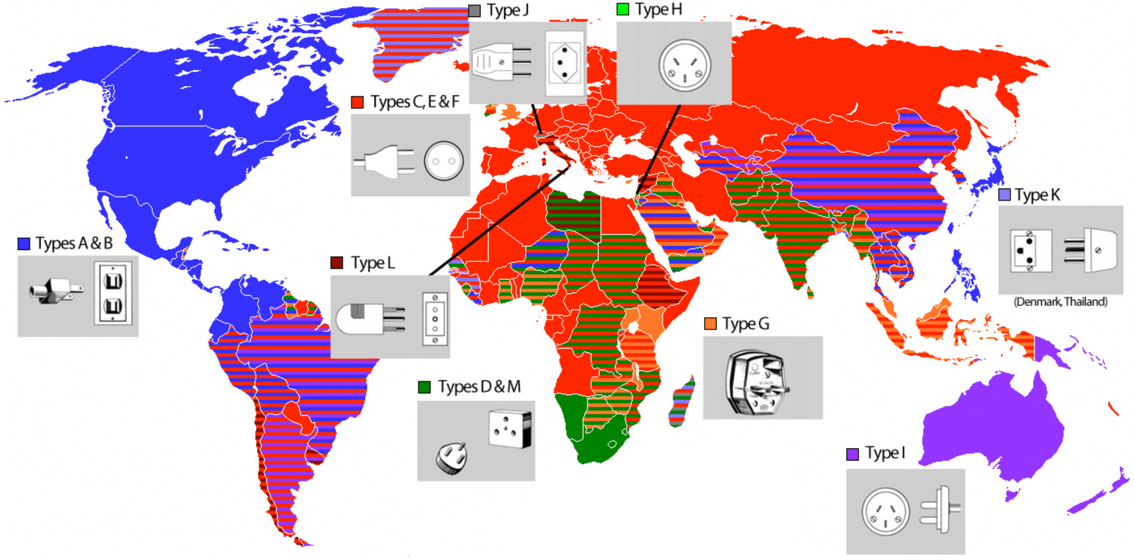


Courtesy of Morten W. Hansen, NERSC

- 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

Benefits of standardised documentation

- Why not use the “Google” approach?



- 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

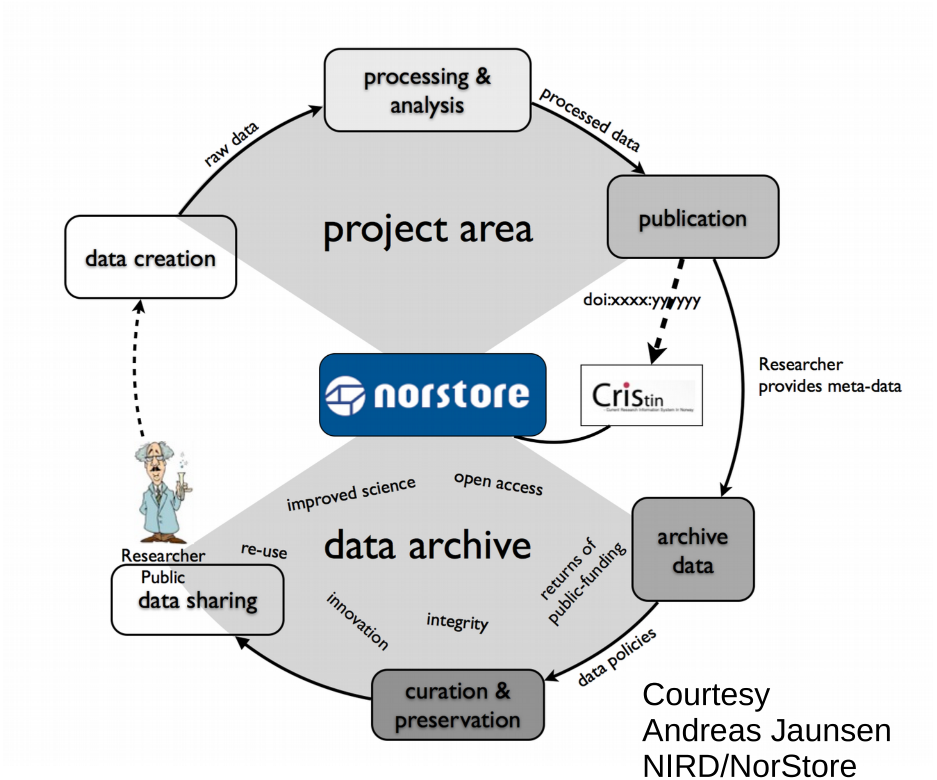
Data and metadata must be connected

- To find data
- To use data

Need to be pragmatic…

- And let computers do the boring part
- But humans need to instruct computers

Data in context



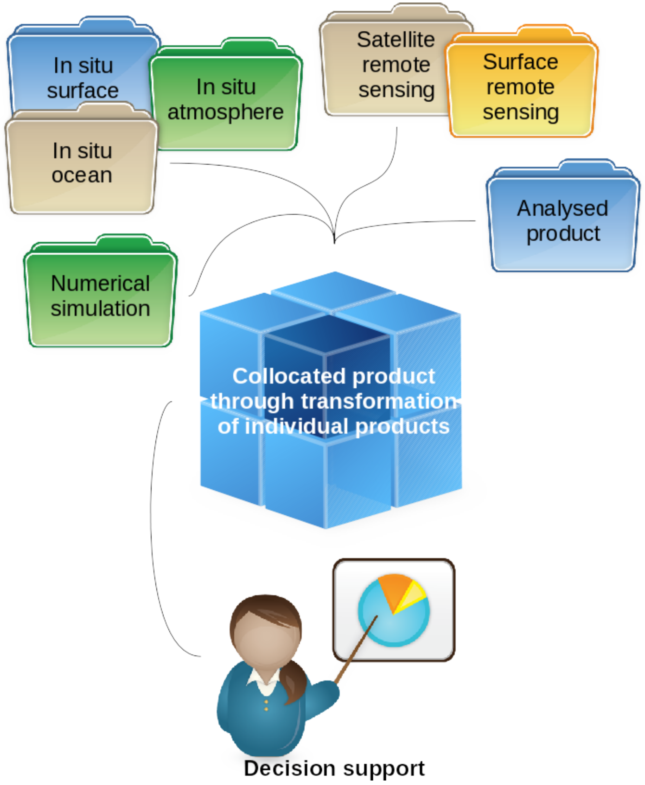
Courtesy
Andreas Jaunsen
NIRD/NorStore

- What's the meaning of a number?
  - Basic metadata are needed for any use of data
- Data can be used in different ways
  - For adequate use of data, adequate information about the data is critical
- The whole is more than the sum of the pieces
  - Smart combination of information has a much larger potential than single observations
- Must
  - Make data talk together
  - Make data traceable
  - Make data count

The FAIR guiding principles

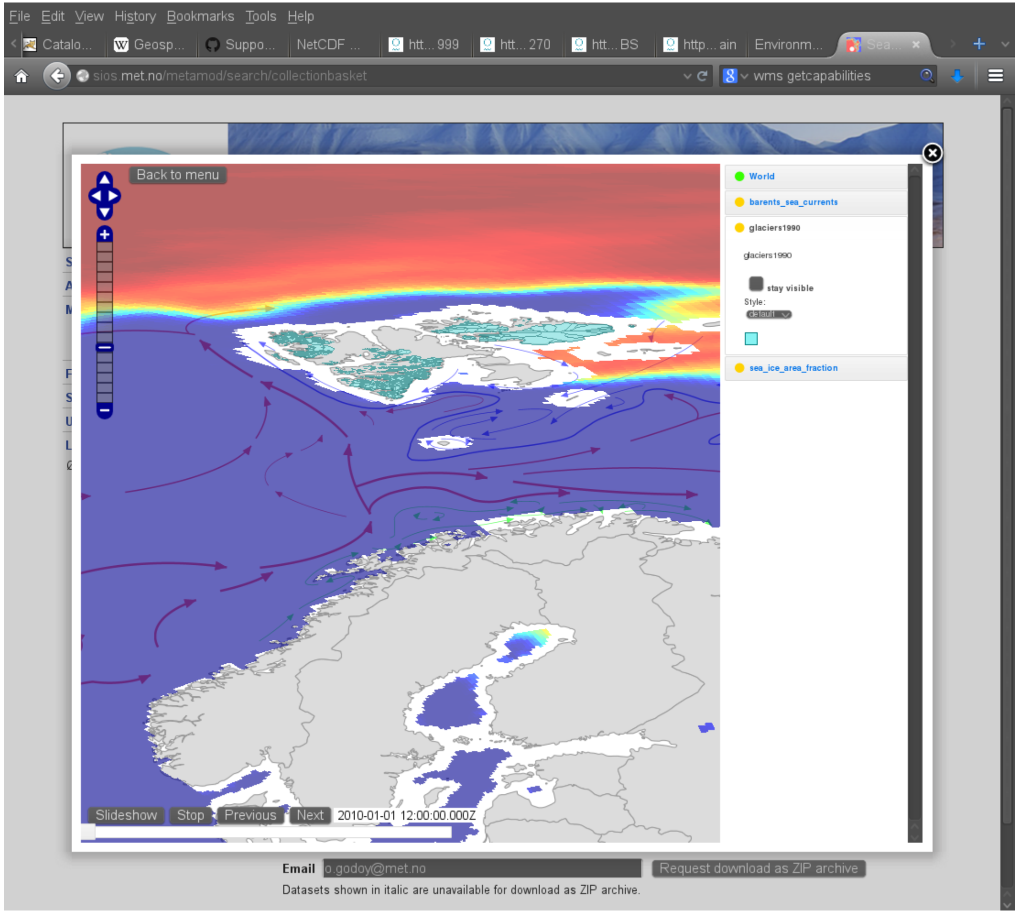
- To be Findable:
  - F1. (meta)data are assigned a globally unique and persistent identifier
  - F2. data are described with rich metadata (defined by R1 below)
  - F3. metadata clearly and explicitly include the identifier of the data it describes
  - F4. (meta)data are registered or indexed in a searchable 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

Approach



- Dataset oriented
  - Metadata driven
- Open data space
  - Higher order services offered when the data space can be constrained
- Net centric
  - Linkages between data centres is vital
  - Implies brokering of metadata and data
- Interdisciplinary
  - Dataset agnostic in the open data space

Demonstrator from SIOS

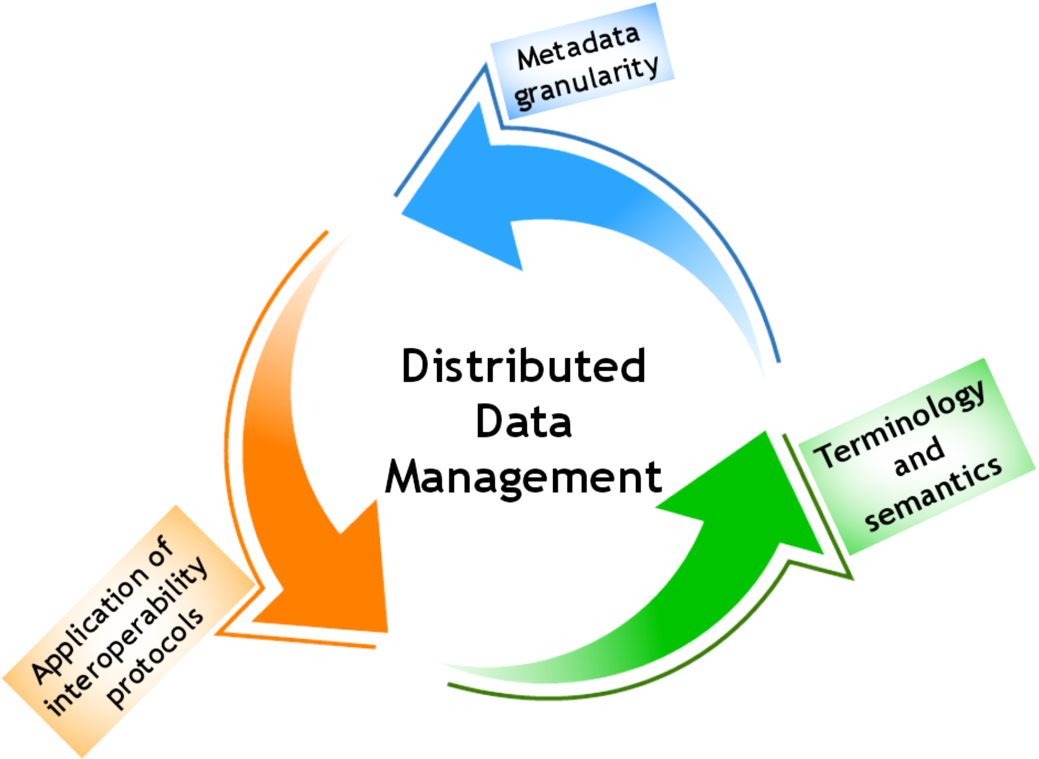

File Edit View History Bookmarks Tools Help
Catalo... Geosp... Suppo... NetCDF ... htt...999 htt...270 htt...BS http...ain Environm... Sea...
sios.met.no/metamod/search/collectionbasket
wms getcapabilities
Back to menu
World
barents_sea_currents
glaciers1990
glaciers1990
stay visible
Style:
default
sea_ice_area_fraction
Slideshow
Stop
Previous
Next
2010-01-01 12:00:00.000Z
Request download as ZIP archive

Email o.godoy@met.no

Datasets shown in italic are unavailable for download as ZIP archive.

- Integrates data using OGC WMS
  - Norwegian Polar Institute (NO)
  - Institute of Marine Research (NO)
  - Norwegian Meteorological Institute (NO)
- OAI-PMH
  - GCMD DIF
  - ISO19115

Challenges during integration


Metadata granularity
Distributed Data Management
Terminology and semantics
Application of interoperability protocols

- Interoperability
  - Discovery Metadata
    - Exchange Protocols (✓)
    - Structures (✓)
    - Semantics/terminology (-)
  - Data
    - Exchange Protocols (✓)
    - Formats (-)
    - Use metadata (✓)
    - Semantics/terminology (-)
    - Common data model (-)
- Cultural
  - Sharing data...

Data management plan

- The purpose of the Data Management Plan is to describe the data that will be created and how it will be shared and preserved.
- The goal is a unified view of INTERACT data that will improve the impact of INTERACT and individual stations.
- The basic principles of INTERACT data management is that INTERACT is following a metadata driven approach.
- INTERACT datasets are described using standardised discovery metadata.
  - This shall ensure the data are archived and re-usable for future generations and relevant to technologically driven data analysis developments.
- INTERACT shall rely on discipline specific efforts to establish interoperability at the data level.
- INTERACT promotes free and open access to data in line with the European Open Research Data Pilot (OpenAIRE).

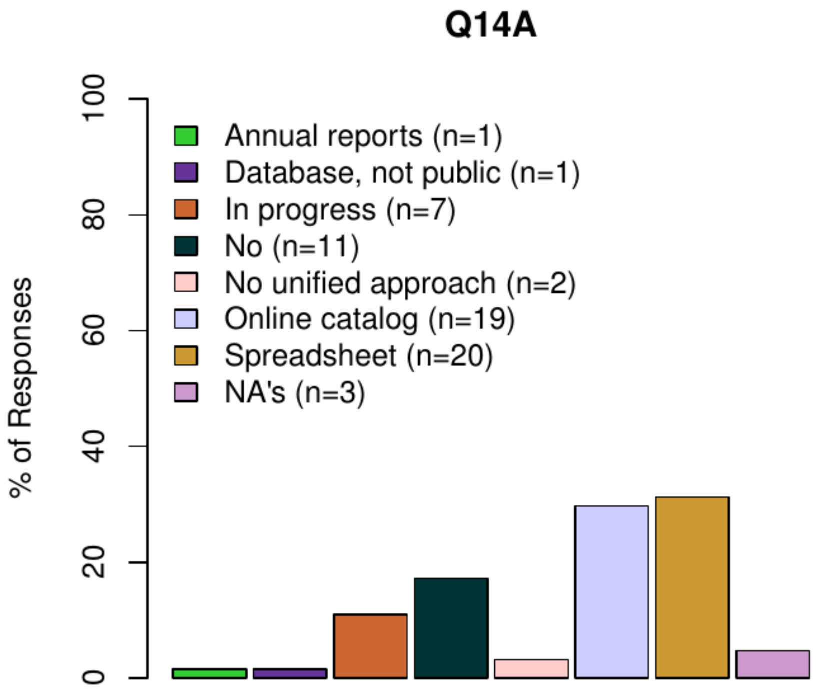
- Selected recommendations
  - metadata and data products shall be free and open (Creative Commons attribution license),
    - although some data may have temporal restrictions
  - shall use self explaining file formats/data encoding
  - shall use the NetCDF format following the Climate and Forecast Convention where possible
  - shall make data available in a timely fashion
  - data shall be archived in repositories with a long term mandate
  - promotes and encourages the implementation of globally resolvable Persistent Identifiers (e.g. Digital Object Identifiers) at each contributing data centre

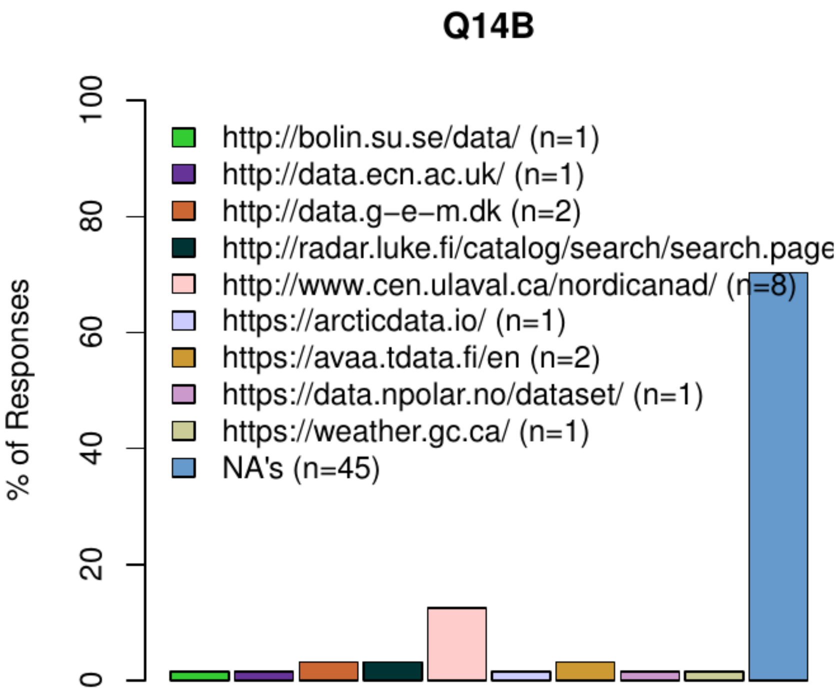
Current state of data management

- Survey circulated among station managers
  - used to guide development of Data Management Plan
- 16 multiple choice and free text questions addressing FAIR guiding principles
  - Wilkinson et. al (2016)
- 78% of INTERACT stations responded

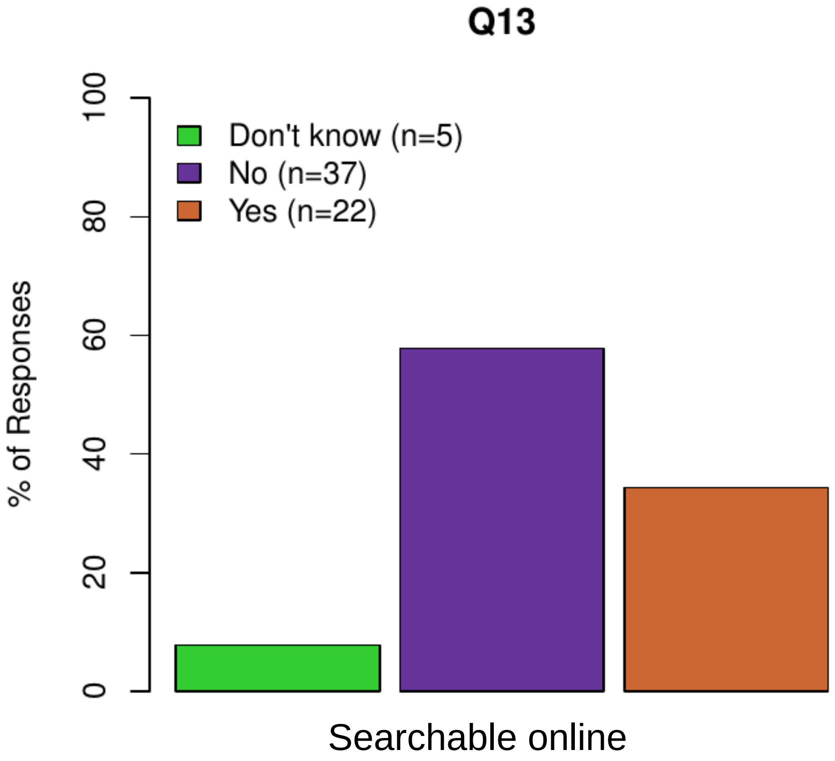
- Considerable lack of knowledge on data management requirements
- For many stations responsibilities are unclear
  - resulting in low or lacking data integrity/security

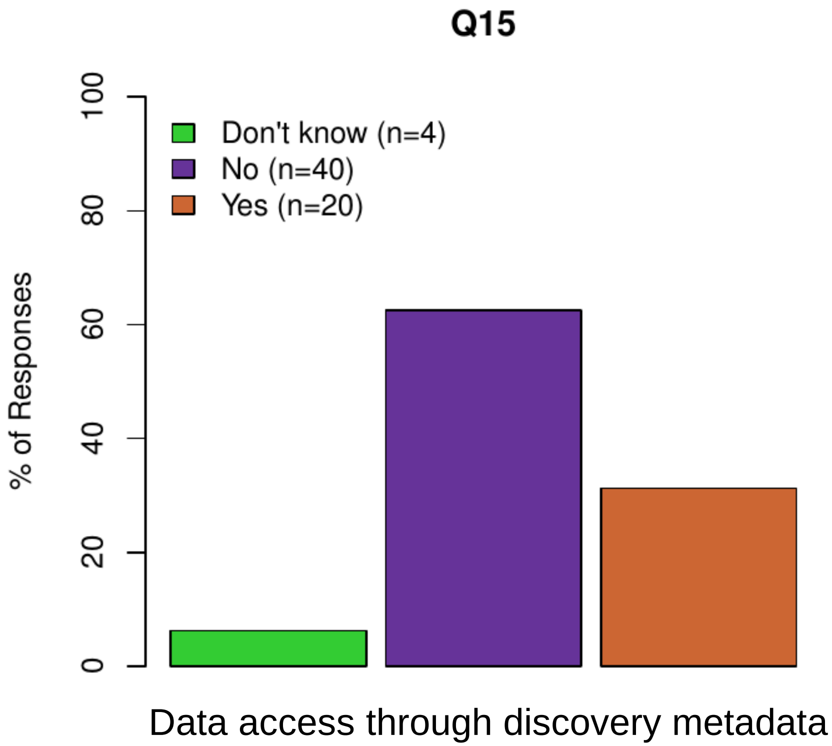
Catalog service



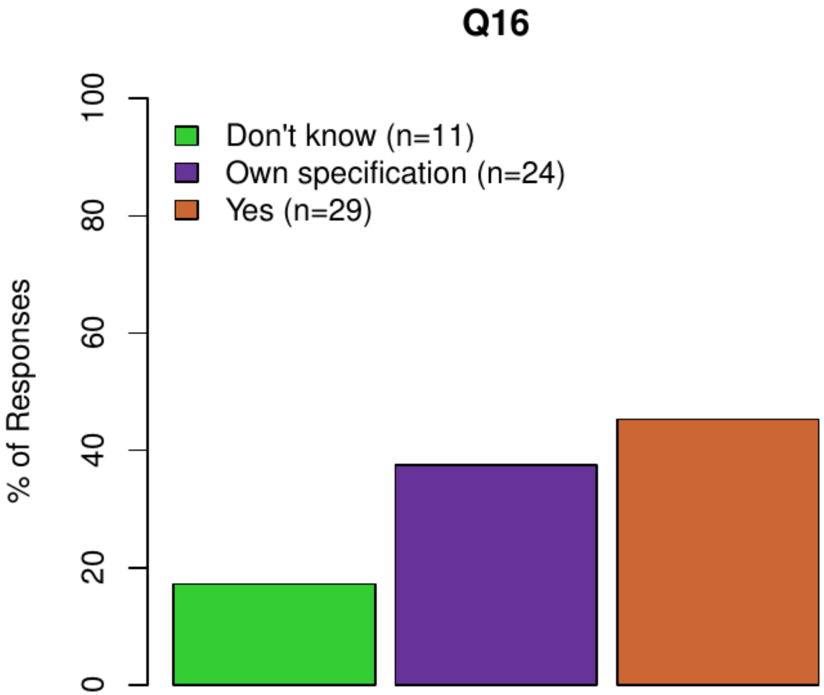


Discovery metadata

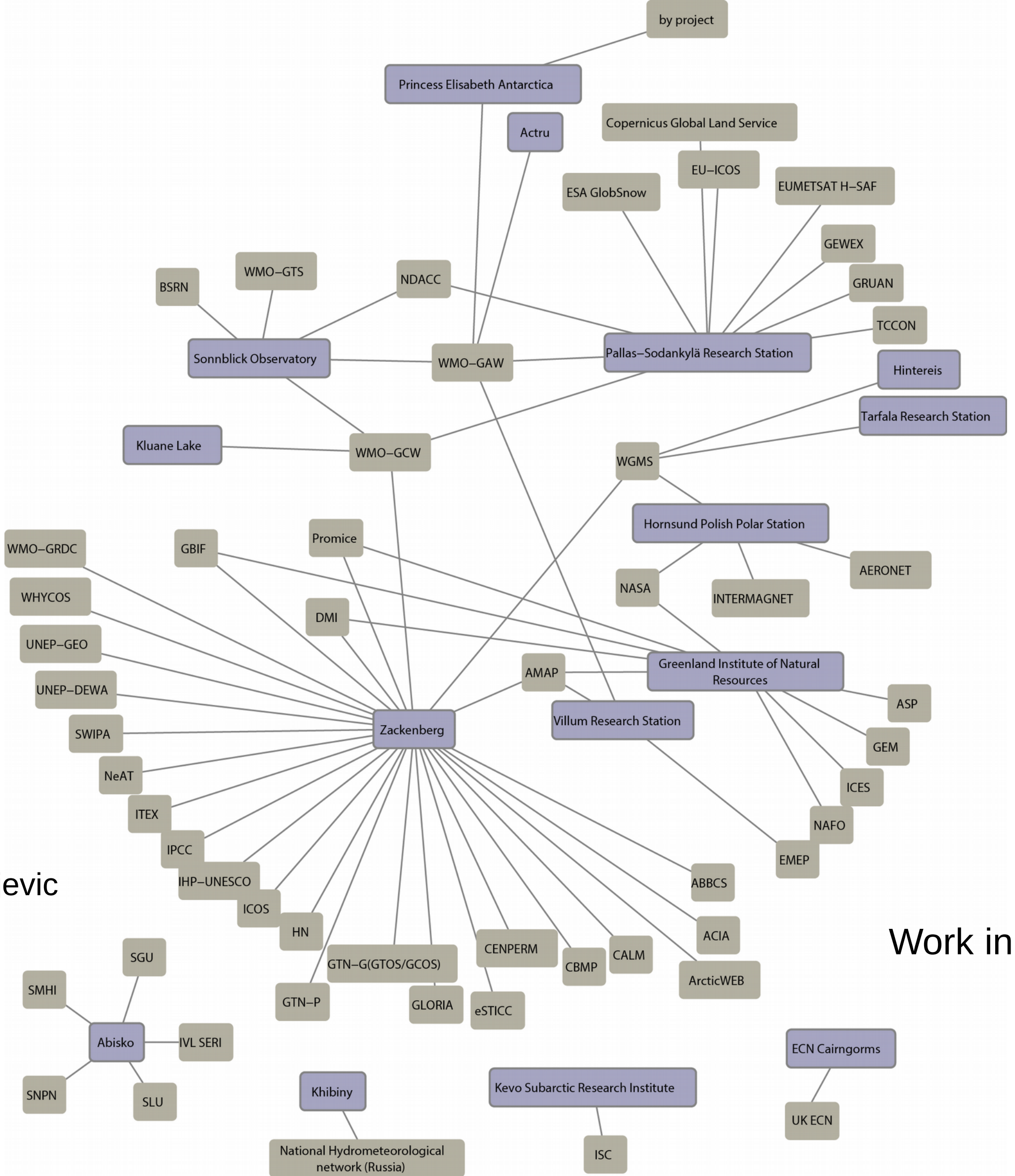




Standardised format



- Yes is probably falling into the “Own specification”



by project

Princess Elisabeth Antarctica

Actru

Copernicus Global Land Service

EU–ICOS

ESA GlobSnow

EUMETSAT H–SAF

GEWEX

GRUAN

TCCON

WMO–GTS

BSRN

NDACC

Sonnblick Observatory

WMO–GAW

Pallas–Sodankylä Research Station

Hintereis

Tarfala Research Station

Kluane Lake

WMO–GCW

WGMS

Hornsund Polish Polar Station

AERONET

NASA

INTERMAGNET

WMO–GRDC

GBIF

Promice

WHYCOS

DMI

UNEP–GEO

Greenland Institute of Natural Resources

AMAP

UNEP–DEWA

Villum Research Station

Zackenberg

SWIPA

ASP

GEM

NeAT

ICES

ITEX

NAFO

IPCC

EMEP

IHP–UNESCO

ABBCS

ICOS

HN

ACIA

CENPERM

CALM

CBMP

GTN–G(GTOS/GCOS)

ArcticWEB

SGU

SMHI

GLORIA

eSTICC

GTN–P

Abisko

IVL SERI

ECN Cairngorms

SNPN

SLU

Khibiny

Kevo Subarctic Research Institute

UK ECN

National Hydrometeorological network (Russia)

ISC

Courtesy of
Boris Radosavljevic

Work in progress

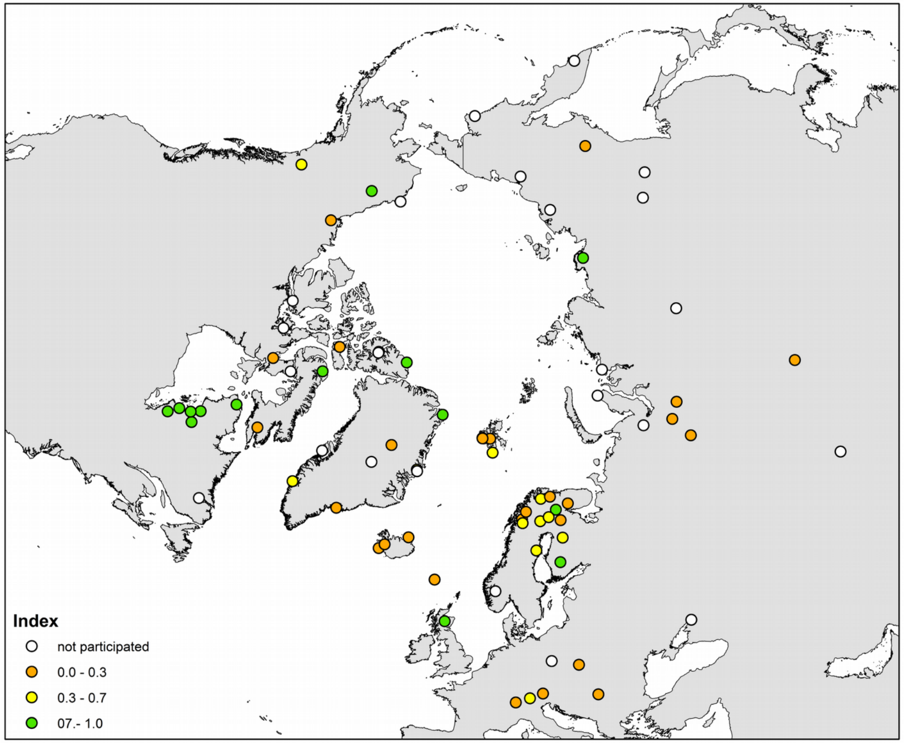
Summary

- If data are
  - Well-organized
  - Documented
  - Preserved
  - Accessible
  - Verified as to accuracy and validity
- Result is
  - High quality data
  - Easy to share and re-use in science
  - Citation and credibility to the researcher
  - Cost-savings to science

- The data deluge has created a surge of information that needs to be well-managed and made accessible.
- The cost of not doing data management can be very high.
- Be cognizant of best practices and tools associated with the data lifecycle to manage your data well.
- Many benefits are associated with the act of managing data, including the ability to find, access, understand, integrate, and re-use data.

Next steps

- Evaluate the interoperability status of data centres identified so far
  - Establishing demonstrator of the unified data portal
- Drafting data management and interoperability guidelines for field stations
  - Need to engage the community
  - Adapting to external forcing mechanisms
- Drafting INTERACT data policy
  - Need to engage the community
  - Ethical principles and behaviour
  - Adapting to external forcing mechanisms


Index
○ not participated
● 0.0 - 0.3
● 0.3 - 0.7
● 07.- 1.0