INTER ⇄ ACT

WP4: Data Forum

Øystein Godøy, Anna Irrgang

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

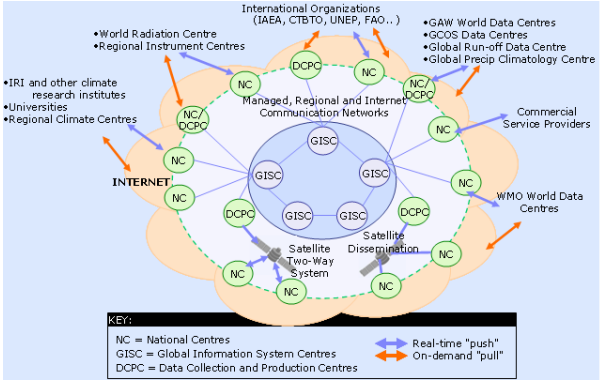
Norwegian Meteorological Institute

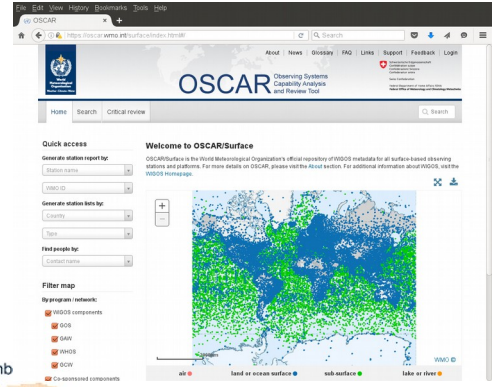
Assumptions

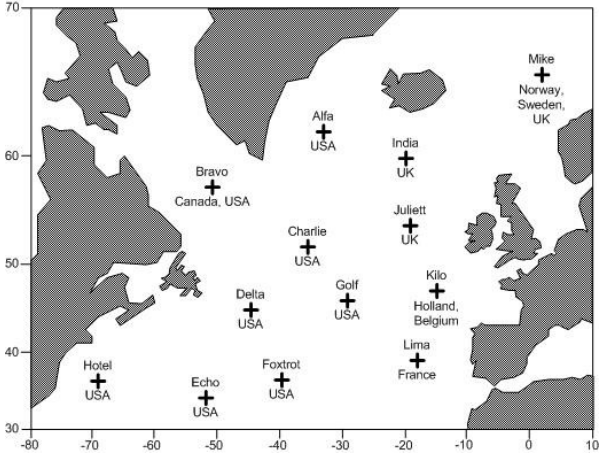
- INTERACT research stations generate data and metadata
  - Long term monitoring
  - Short term process studies
  - External data by individual scientists/ groups
- Research stations archive data and metadata (internal and external)
  - e.g. meteorological data
  - photos, maps, reports etc.
  - list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

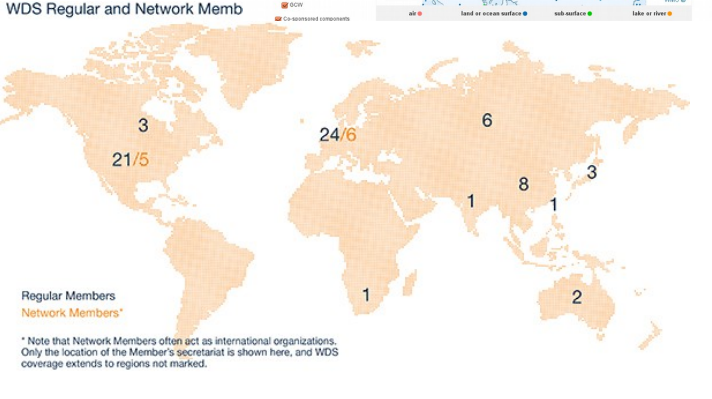
Visibility

- 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
  - ...







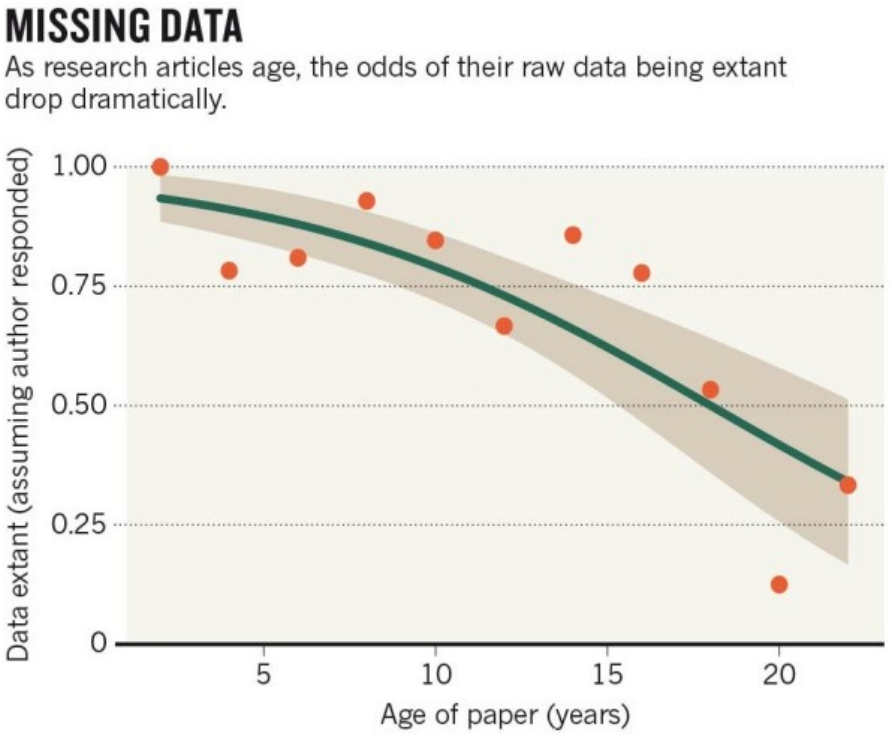


Data Management

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

DATA
not available

PEOPLE
not available



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

Motivation

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

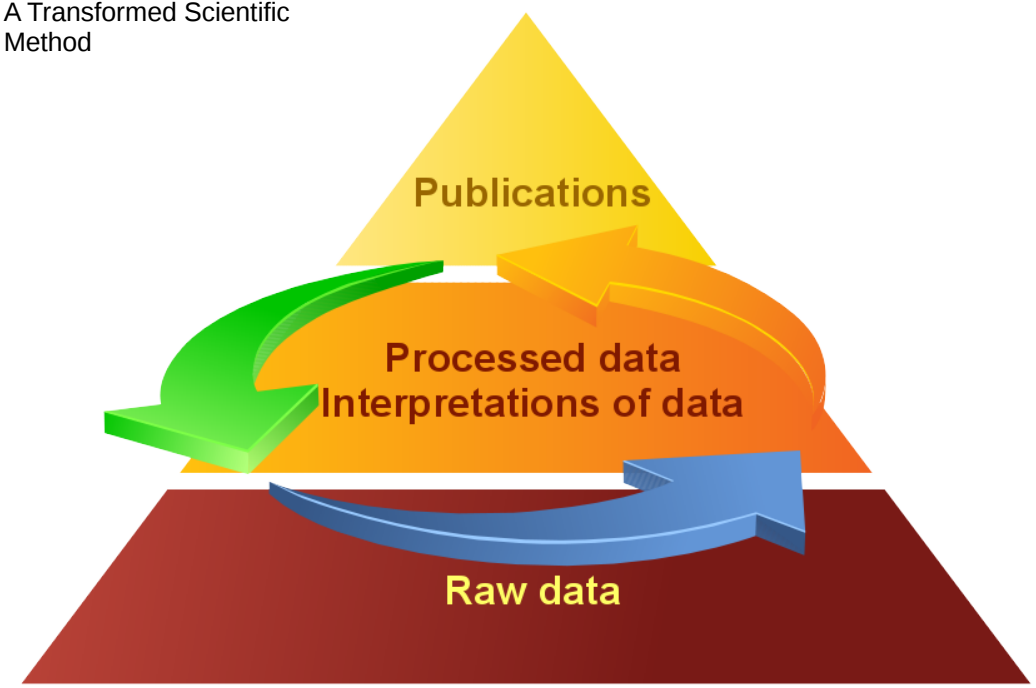
Why bother with structured data management?

- 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

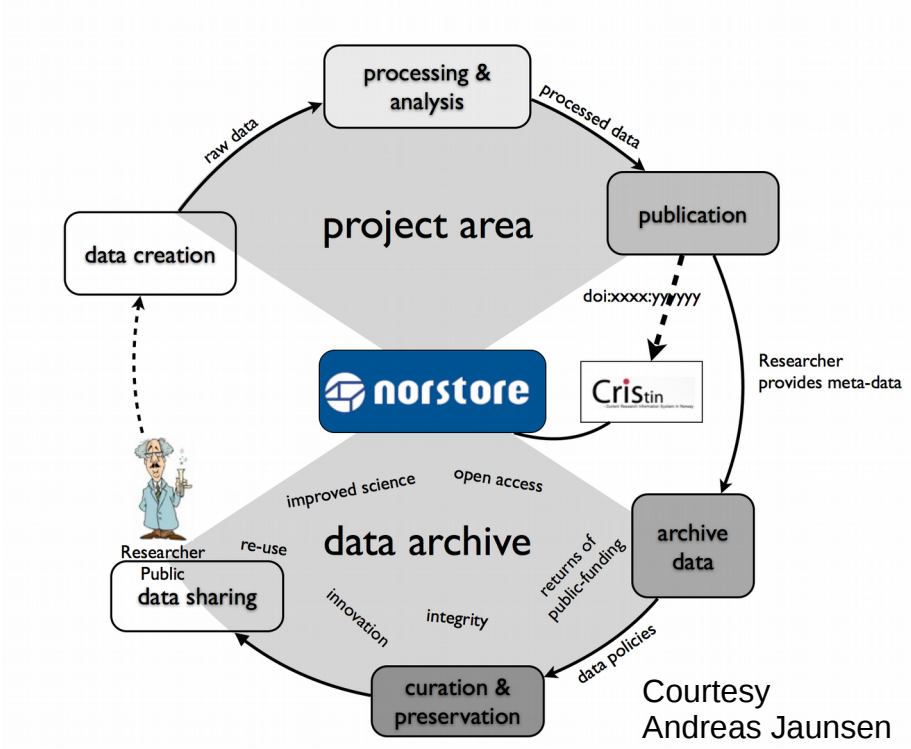
All scientific data online

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

Must put 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

Objectives of the work

- 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

Strategy

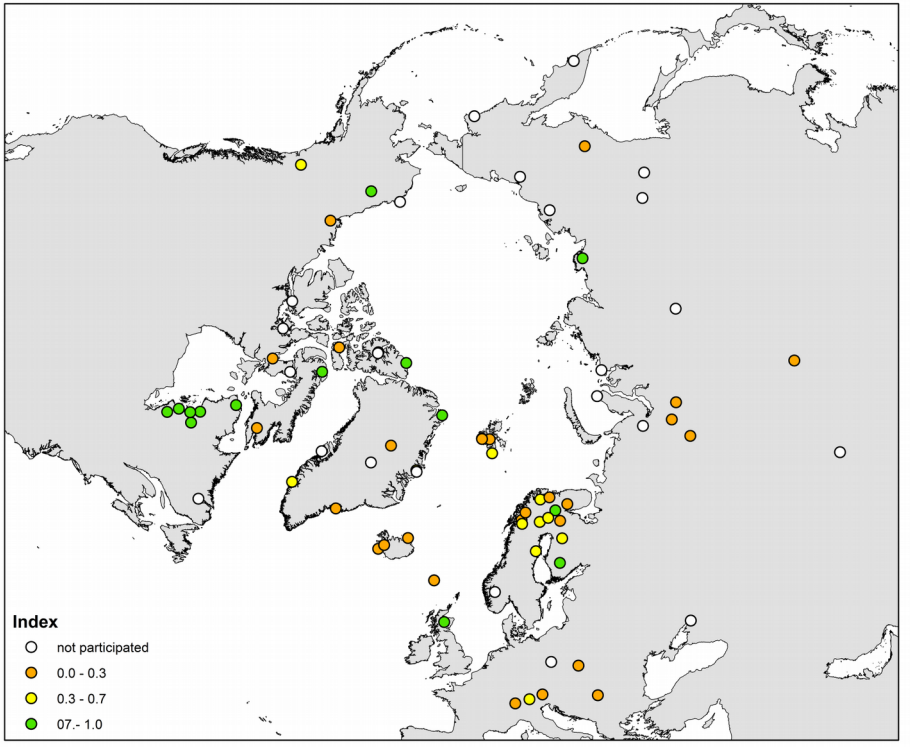
- Initially focus on discovery metadata
  - Ensure interoperability and information flow/streams
  - To establish a unified view of the INTERACT data space
  - Expose this to external frameworks (metadata standards)
- Move to interoperability at the data level when data discovery is working
  - Interoperability 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)
  - Delivered
- D4.2: Report on current data flows (Month 12)
  - Gap analysis and bottlenecks
  - Delivered
- D4.3: Field guide to data repositories (Month 24)
  - Mentoring of potential providers of TA virtual access
  - In progress
- D4.4: Data Policy (Month 24)
  - In progress

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



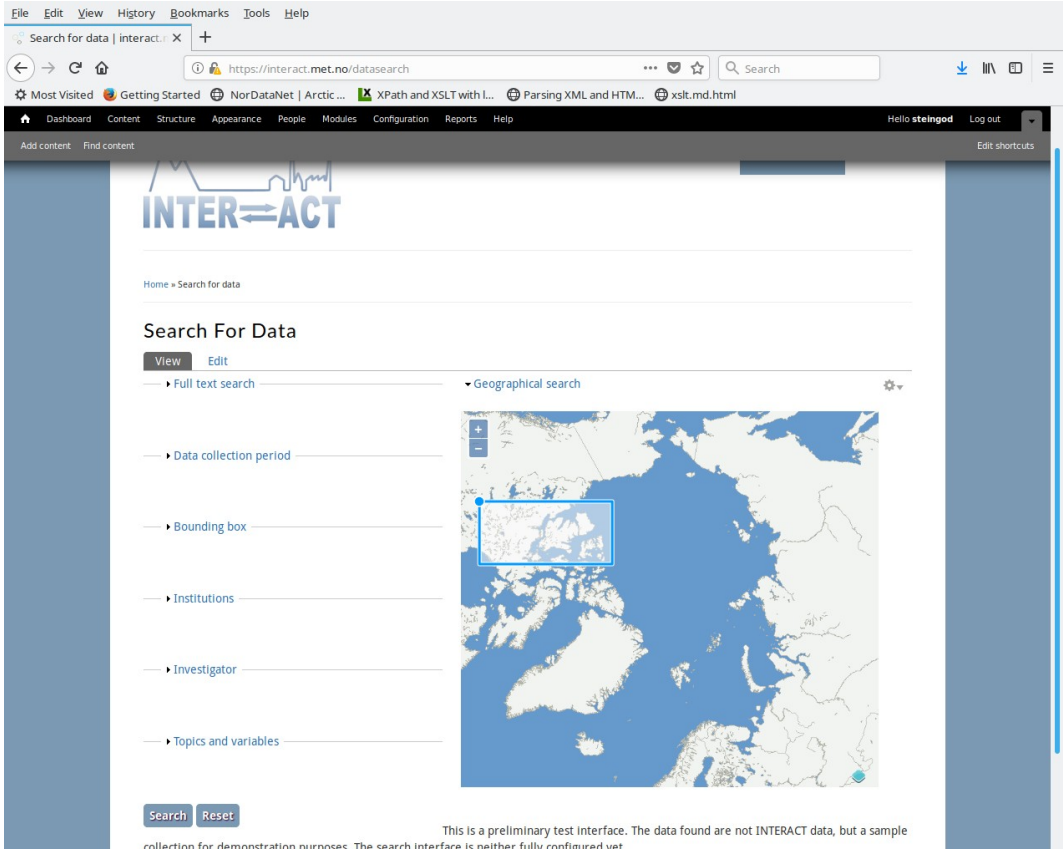
Data management plan

- The purpose of the Data Management Plan is to describe the data that will be created and how it will be shared and preserved.
  - Interact DMP is very high level
- 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.
- This calls for INTERACT datasets described using standardised discovery and use metadata.
  - This shall ensure the data are archived and re-usable for future generations and relevant to technologically driven data analysis developments.
- INTERACT relies on discipline specific efforts to establish interoperability at the data level.
  - But FAIRness decides
- INTERACT promotes free and open access to data in line with the European Open Research Data Pilot (OpenAIRE).

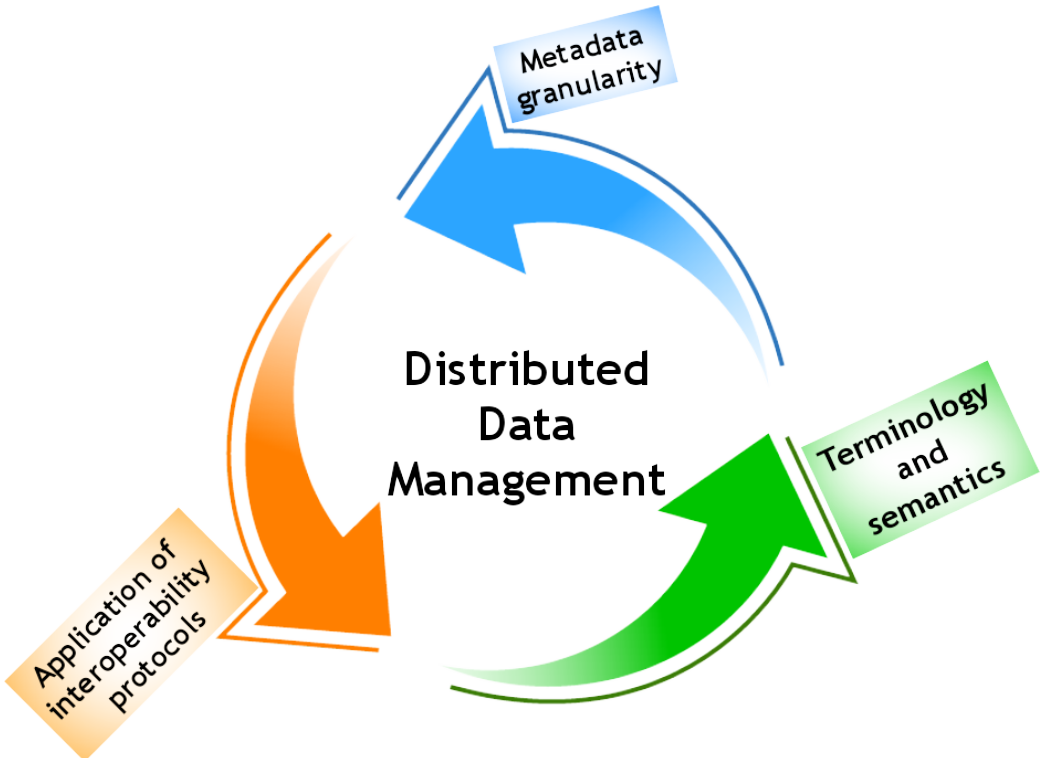
- Selected recommendations
  - metadata and data products shall be free and open (Creative Commons attribution license),
    - although some data may have temporal restrictions
    - to be further detailed in the INTERACT Data Policy
  - shall use self explaining file formats/data encoding
    - Ensuring a lasting legacy
  - shall make data available in a timely fashion
  - data shall be archived in repositories with a long term mandate and in an interoperable form
  - promotes and encourages the implementation of globally resolvable Persistent Identifiers (e.g. Digital Object Identifiers) at each contributing data centre
    - To get credit to data providers

Ongoing work

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

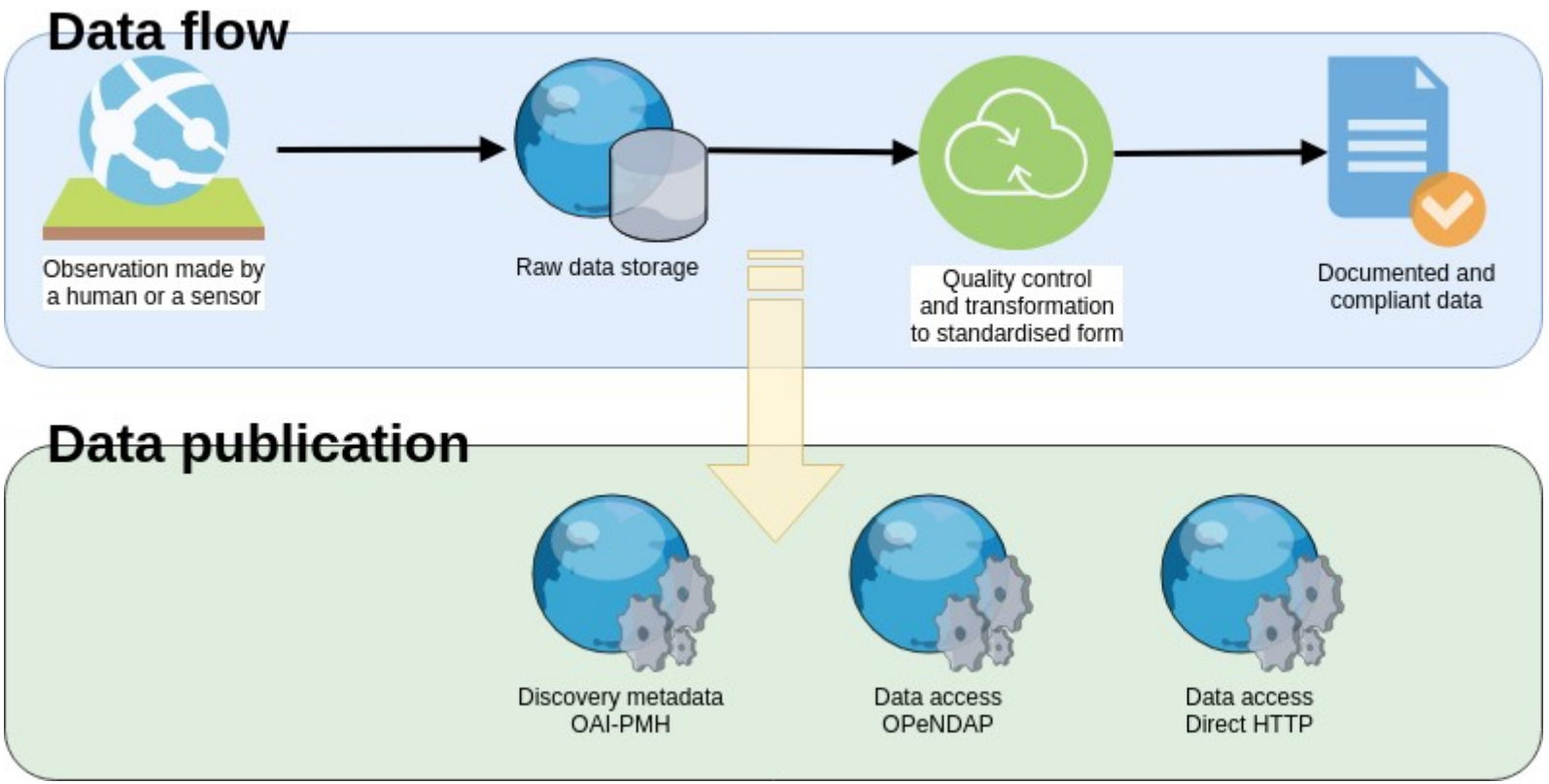


Challenges during integration



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

The promised GCW/SLF software stack



Relevant activities

- SAON/IASC Arctic Data Committee
  - Interoperability Workshop and Assessment Process
    - Frascati November 2016
  - Polar Data Planning Summit
    - Boulder May 2018
  - Polar Data Architecture Workshop
    - Geneva November 2018
- These activities are related to EU-POLARNET and EU-Arctic-Cluster activities
- ENVRI-FAIR starting up, connecting infrastructures to EOSC