SIOS Core Data as Analysis Ready Data supporting Science for Service and Society

Øystein Godøy

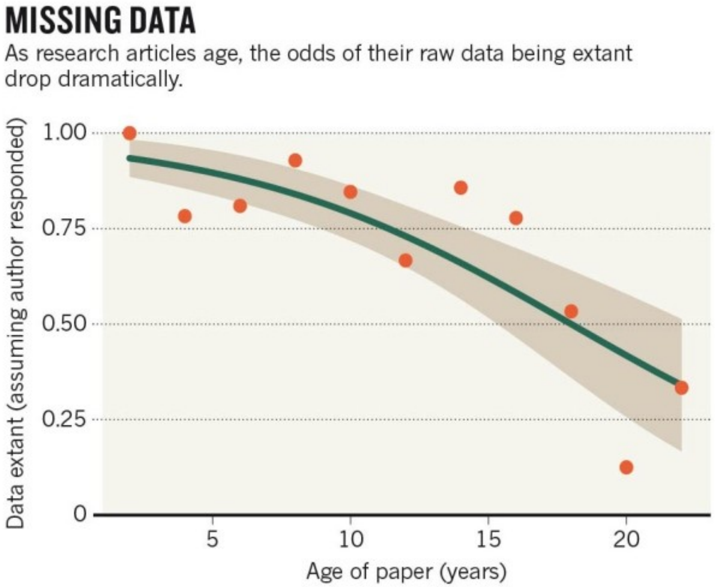
Photo: Malin Daase

SIOS

SVALBARD INTEGRATED ARCTIC EARTH OBSERVING SYSTEM

Why bother with structured data management?

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

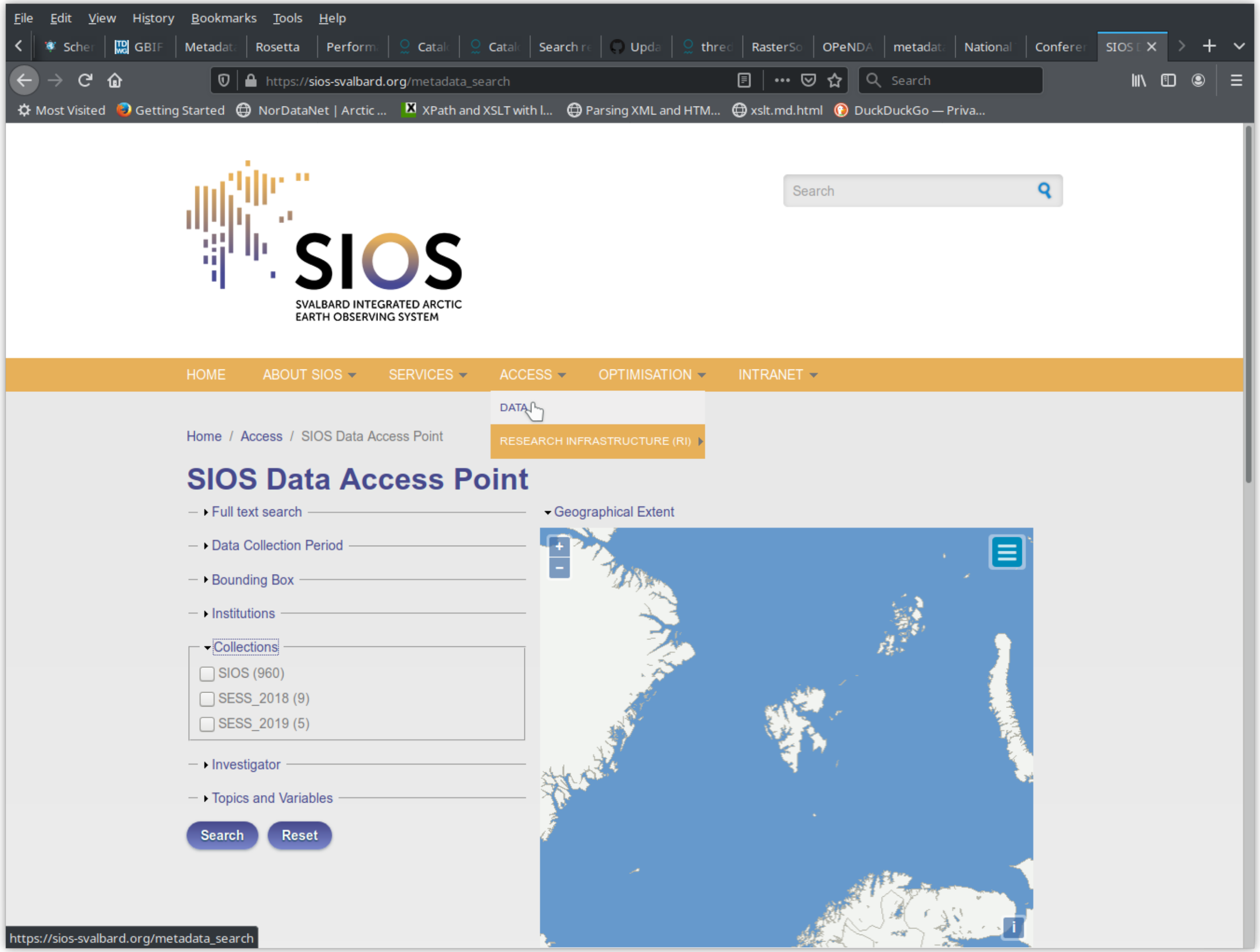

MISSING DATA

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

- 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

SIOS Data Management System

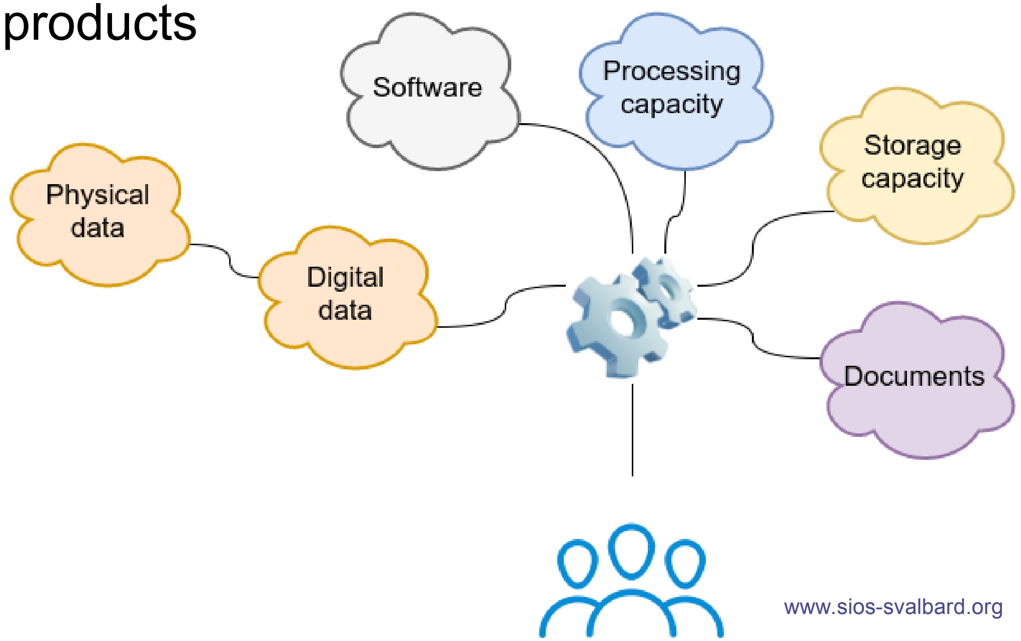
Data Management Service


File Edit View History Bookmarks Tools Help
Sche GBIF Metadat Rosetta Perform Catal Catal Search r Upda threc RasterSo OPeNDA metadat National Confere SIOS
https://sios-svalbard.org/metadata_search
Search
Most Visited Getting Started NorDataNet | Arctic ... XPath and XSLT with l... Parsing XML and HTM... xslt.md.html DuckDuckGo — Priva...
SIOS
SVALBARD INTEGRATED ARCTIC EARTH OBSERVING SYSTEM
Search
HOME ABOUT SIOS SERVICES ACCESS OPTIMISATION INTRANET
DATA
RESEARCH INFRASTRUCTURE (RI)
Home / Access / SIOS Data Access Point

SIOS Data Access Point

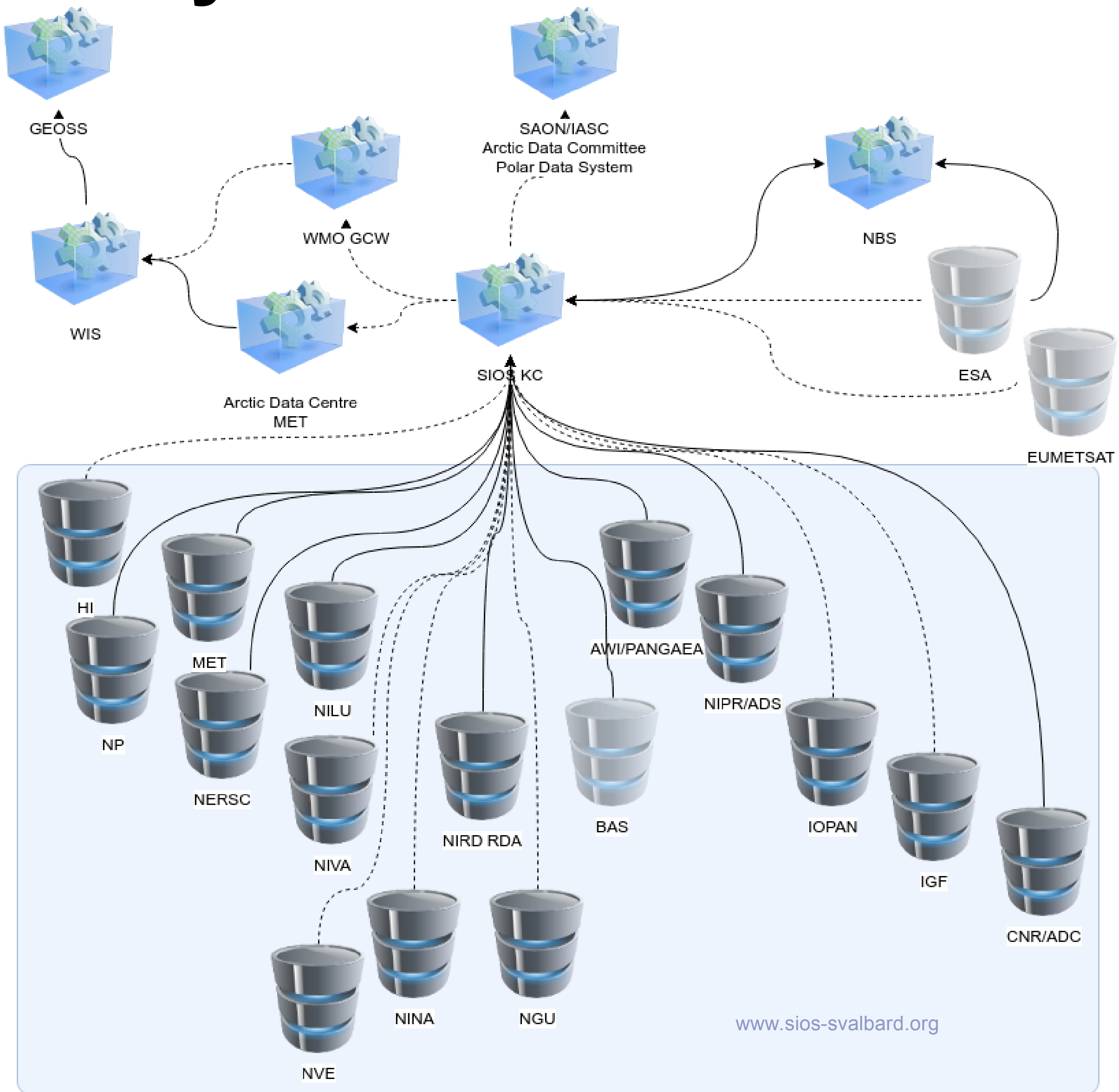
Full text search
Data Collection Period
Bounding Box
Institutions
Collections
SIOS (960)
SESS_2018 (9)
SESS_2019 (5)
Investigator
Topics and Variables
Search
Reset
Geographical Extent
https://sios-svalbard.org/metadata_search

- Main principles
  - Open and free data sharing
  - Distributed data centres
  - International standards
  - Harmonisation of SIOS Core data
- Interdisciplinary combinations of data and products



The SIOS Data Management System

- Integration of existing data centres into a unified system.
  - Linked to larger frameworks
- Each data centre has its own procedures and technical solutions tailored to the needs of that data centre.
- SIOS will not change this, but bridge
  - using internationally accepted interoperability standards and technologies
  - Which can be added as a layer between the data and the SIOS Data Management System



SIOS

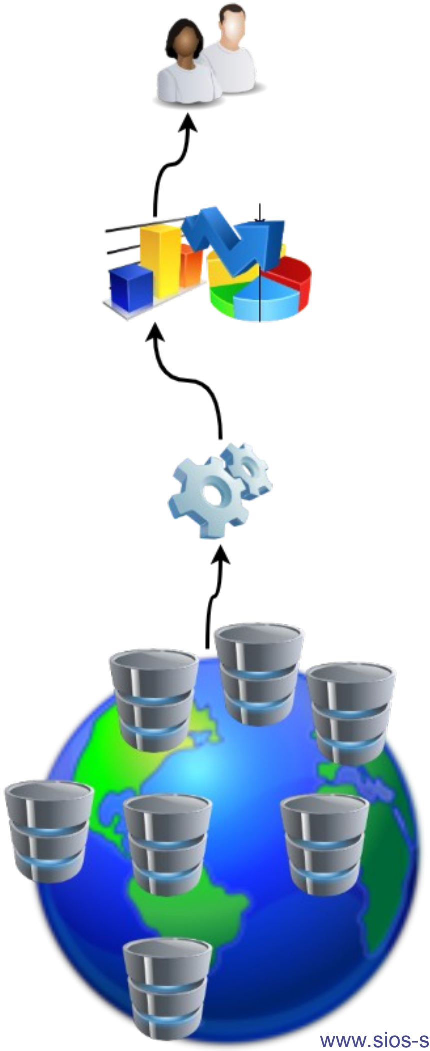
SIOS Core Data

- SIOS Core Data are Earth System Science data for Svalbard that fulfil the defined criteria of scientific requirements, data availability and collecting commitment.
- Data that fulfil these criteria, but are not yet available online are defined as SIOS Core Data Candidates.

- Criteria for SIOS Core Data
  - Scientific requirements
    - SIOS core data are essential in answering the Earth System Science questions defined in the SIOS infrastructure optimisation report, and further update in SESS reports.
  - Data availability
    - SIOS core data should be available through the SIOS Data Management System (SDMS). SIOS core data candidates should be available as soon as possible and at latest one year after data collection.
  - Collecting commitment
    - For SIOS core data there should be a more than 5-year commitment from the providing institute to maintain the measurement and to make the data available through SDMS.

Harmonisation of SIOS core data

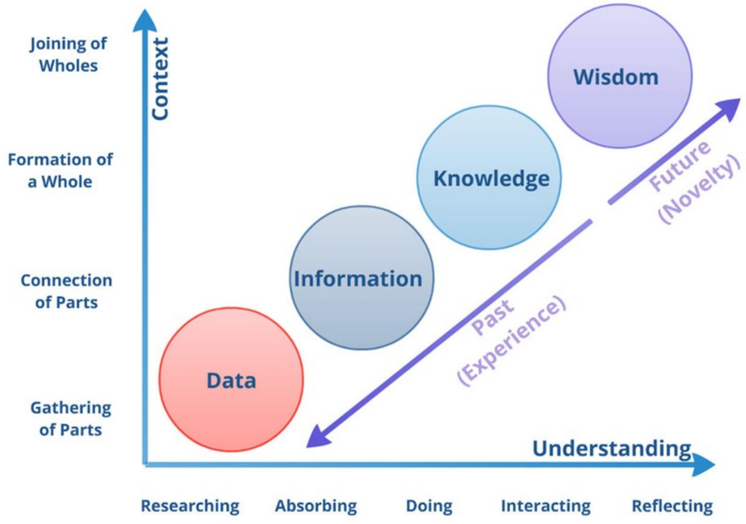
- E.g. GTOS statements
  - Datasets should be harmonised to the extent possible to allow integration of institutional, national and regional datasets into a usable global information resource.
  - Harmonisation seeks to bring together various types, levels and sources of data in such a way that they can be made compatible and comparable, and thus useful for *decision making*.
- Implies harmonisation at the levels of
  - Observation protocols
  - Encoding and publishing of data

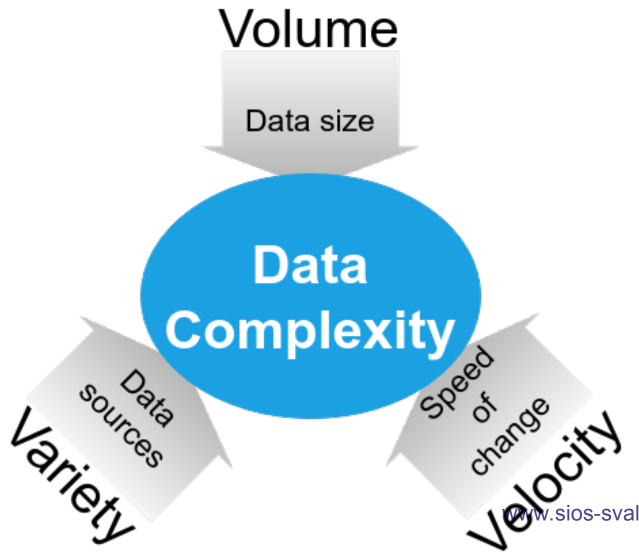


The challenge of data

- Need to integrate data across data providers (silos), communities and languages to simplify data consumption.
  - With the support of emerging technologies.
  - Challenged by heterogeneous data policies.
- Need to combine different types of data.
  - E.g. in situ, remote sensing, numerical simulations and LTK.
  - While minimising the human effort required.
- Need to switch from 80% of human effort on massaging data and 20% on use to the opposite.
- In this context we need to transition from data to knowledge and understanding through connection of the “dots”
  - From a fragmented to consolidated view.

Russ Ackoff “From Data to Wisdom” -
Journal of Applied Systems Analysis, Volume 16, 1989 p 3-9





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*

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

The ata
man

- To be F
  - F1. (pers
  - F2. (by F
  - F3. i iden
  - F4. (sear
- To be A
  - A1. usin
  - A1.1 impl
  - A1.2 auth
  - A2. are

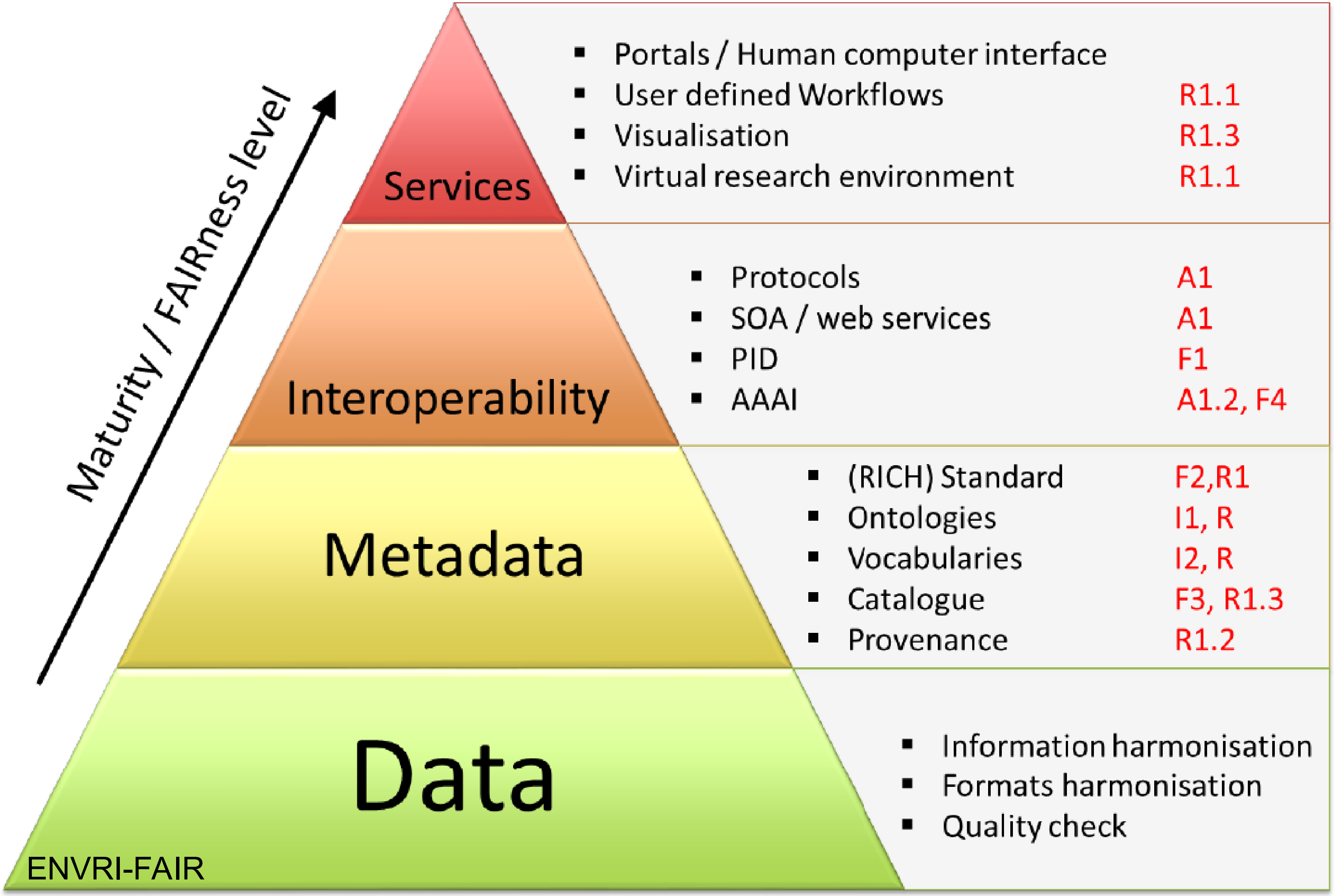
shared, wledge

ow FAIR

es to

a es ar and

tailed



Maturity / FAIRness level

Services

Portals / Human computer interface	
User defined Workflows	R1.1
Visualisation	R1.3
Virtual research environment	R1.1

Interoperability

Protocols	A1
SOA / web services	A1
PID	F1
AAAI	A1.2, F4

Metadata

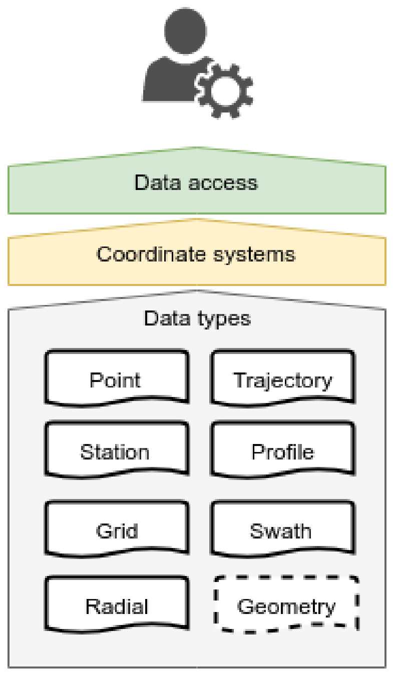
(RICH) Standard	F2,R1
Ontologies	I1, R
Vocabularies	I2, R
Catalogue	F3, R1.3
Provenance	R1.2

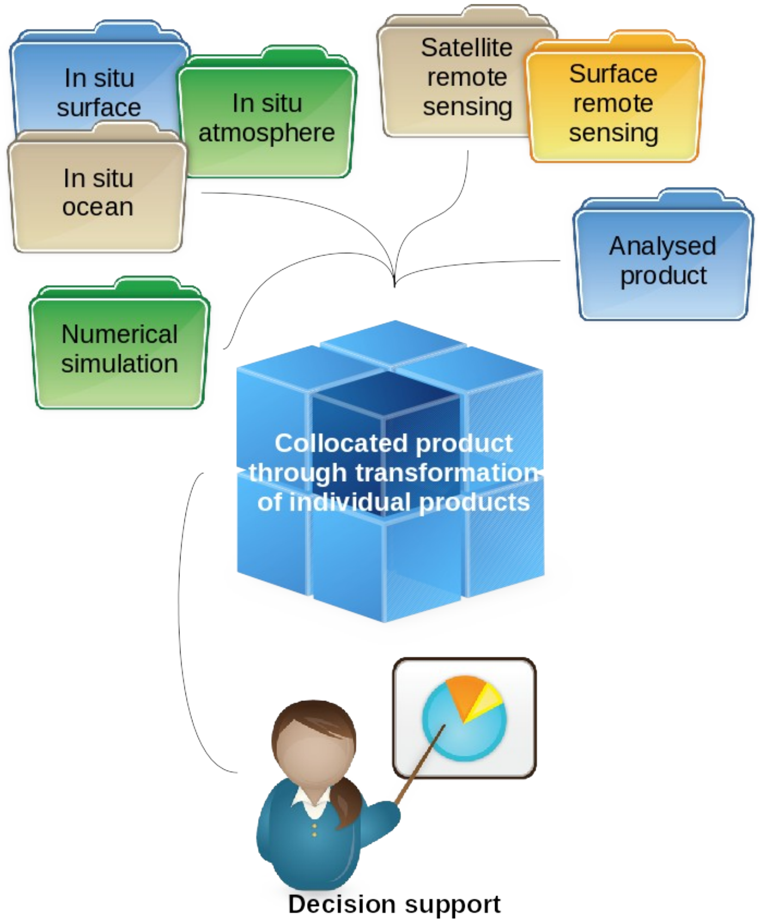
Data

- Information harmonisation
- Formats harmonisation
- Quality check

ENVRI-FAIR

The SIOS approach to data management

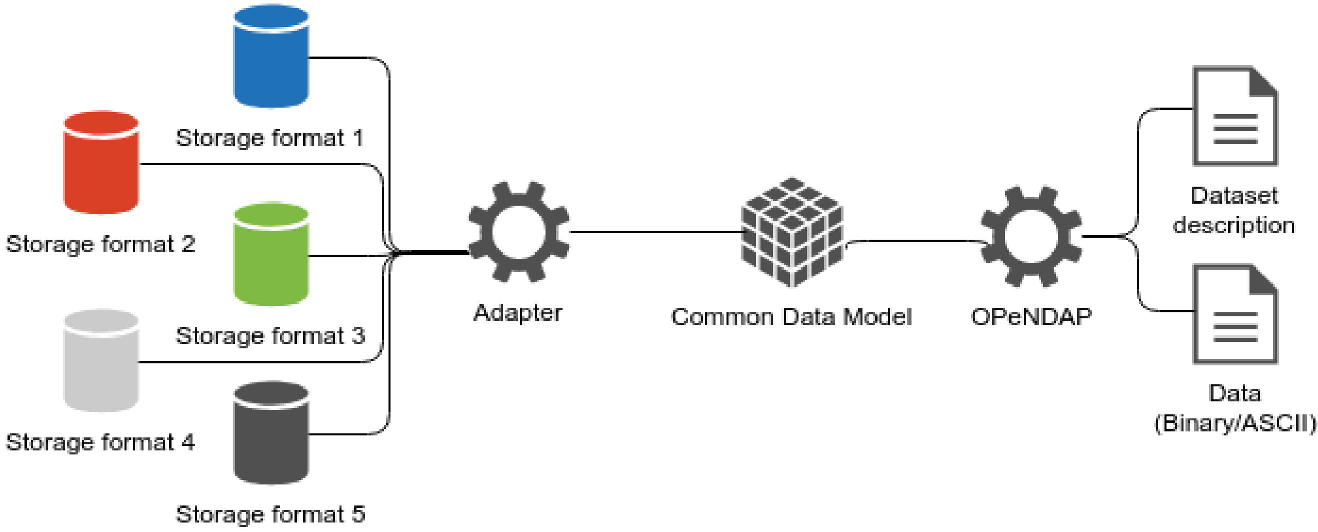




- Open data space
  - Interdisciplinary
  - Higher order services offered when the data space can be constrained
- Net centric
  - Connecting data centres
    - Using standardised interfaces and behaviour
- Dataset oriented
  - Metadata driven
    - Discovery and use metadata
      - Observation facility metadata
  - Identifies web services for datasets through standardised discovery metadata

SDMS Functionality

- Prioritised
  - Data discovery,
    - understood as the process of finding relevant datasets across the distributed data repositories contributing to SDMS.
  - Retrieval of data,
    - understood as the process of downloading data identified in the previous step.
  - Visualisation of data,
    - understood as the process of generating a graphical interpretation of a dataset (either as a map, a time series or appropriate) for data identified previous step.
  - Transformation of data,
    - understood as the process of reformatting, reprojecting, subsetting and combining different datasets into a new dataset.
  - Data submission through well developed documentation, best practices, interfaces and tools.
  - Long term preservation of data sets through mandated data archives

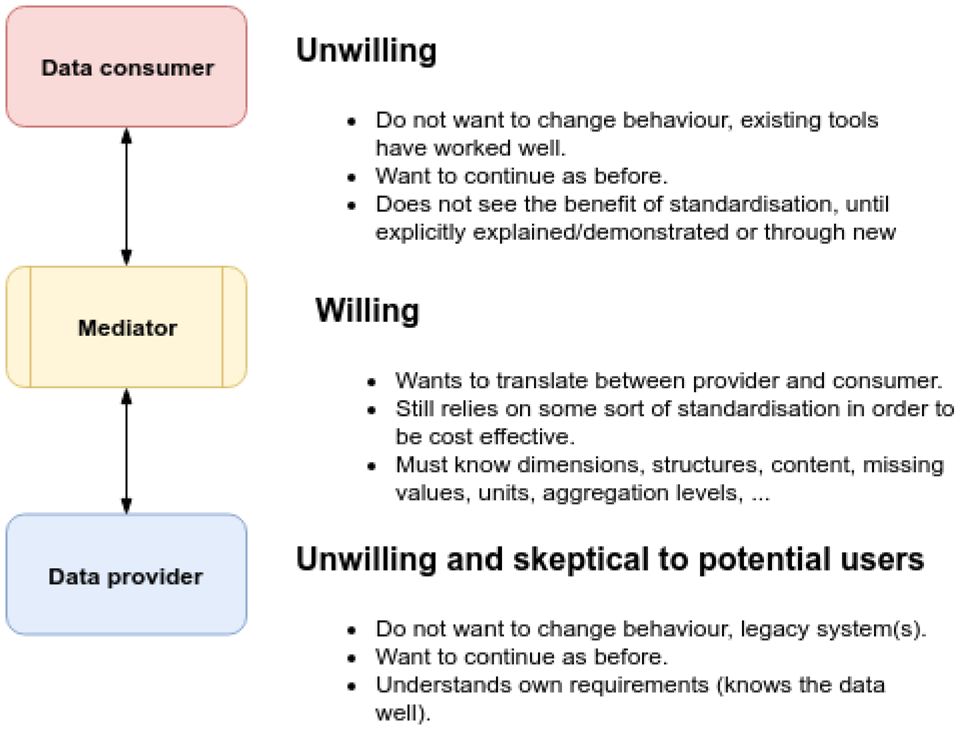


Types of metadata for datasets

Type	Purpose	Description	Examples
Discovery metadata	Used to find relevant data	Discovery metadata are also called index metadata and are a digital version of the library index card. It describes who did what, where and when, how to access data and potential constraints on the data. It shall also link to further information on the data like site metadata.	ISO19115 GCMD DIF ACDD MMD
Use metadata	Used to understand data found	Use metadata are describing the actual content of a dataset and how it is encoded. The purpose is to enable the user to understand the data without any further communication. It describes content of variables using standardised vocabularies, units of variable, encoding of missing values, map projections etc.	Climate and Forecast Convention BUFR GRIB DwCA
Configuration metadata	Used to tune portal services for datasets for users.	Configuration metadata are used to improve the services offered through a portal to the user community. This can be e.g. how to best visualise a product.	MMD
Site metadata	Used to understand data found	Site metadata are used to describe the context of observational data. It describes the location of an observation, the instrumentation, procedures etc. To a certain extent it overlaps with discovery metadata, but more so it really extends discovery metadata. Site metadata can be used for observation network design.	WIGOS OGC O&M

Specific challenges

- Currently most data are non standardised
- Lacking (a common) understanding for standardisation requirements
  - Especially for the core data
- Non standardised interfaces to data are frequently used
  - Invent your own “standard”
- Lacking understanding for the importance of use metadata
  - Enabling reuse across communities and generations
- Lacking understanding for the importance of semantic standardisation
  - How to described the content of datasets
    - In a manner allowing translation between communities and languages
- Crediting all involved parties
  - Scientists, institutions, data centres, ….



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).

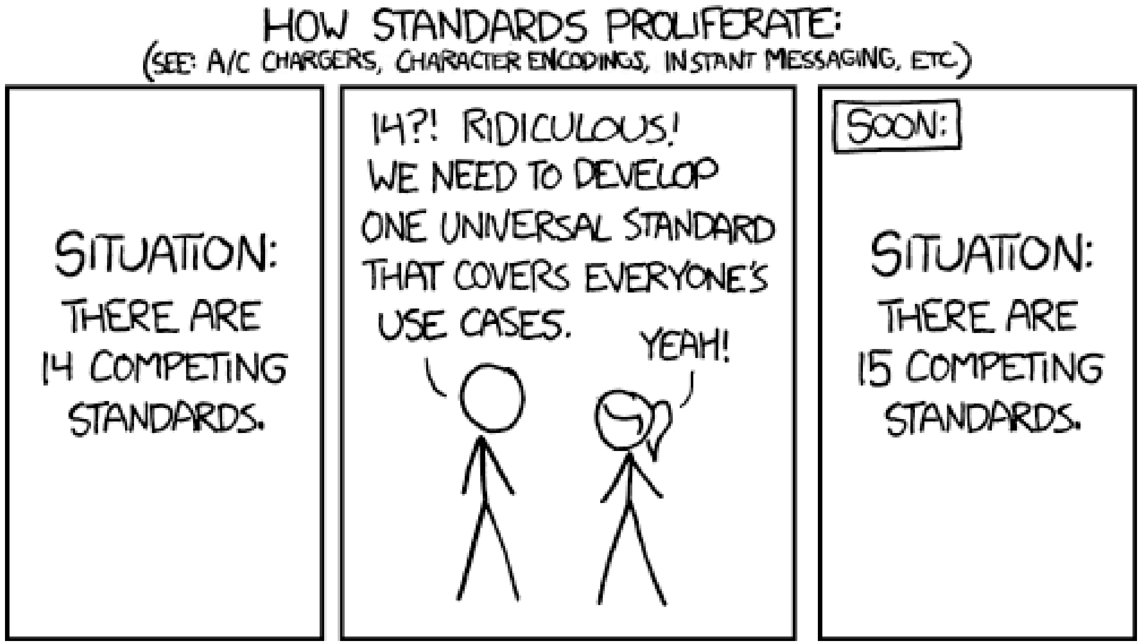
Data Formats: Choosing and Adopting Community Accepted Standards

- Most projects (rightly so) focus on the content of their data files
  - But 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.

Adopted from Curt Tilmes, NASA, 2013
ESIP Material

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.

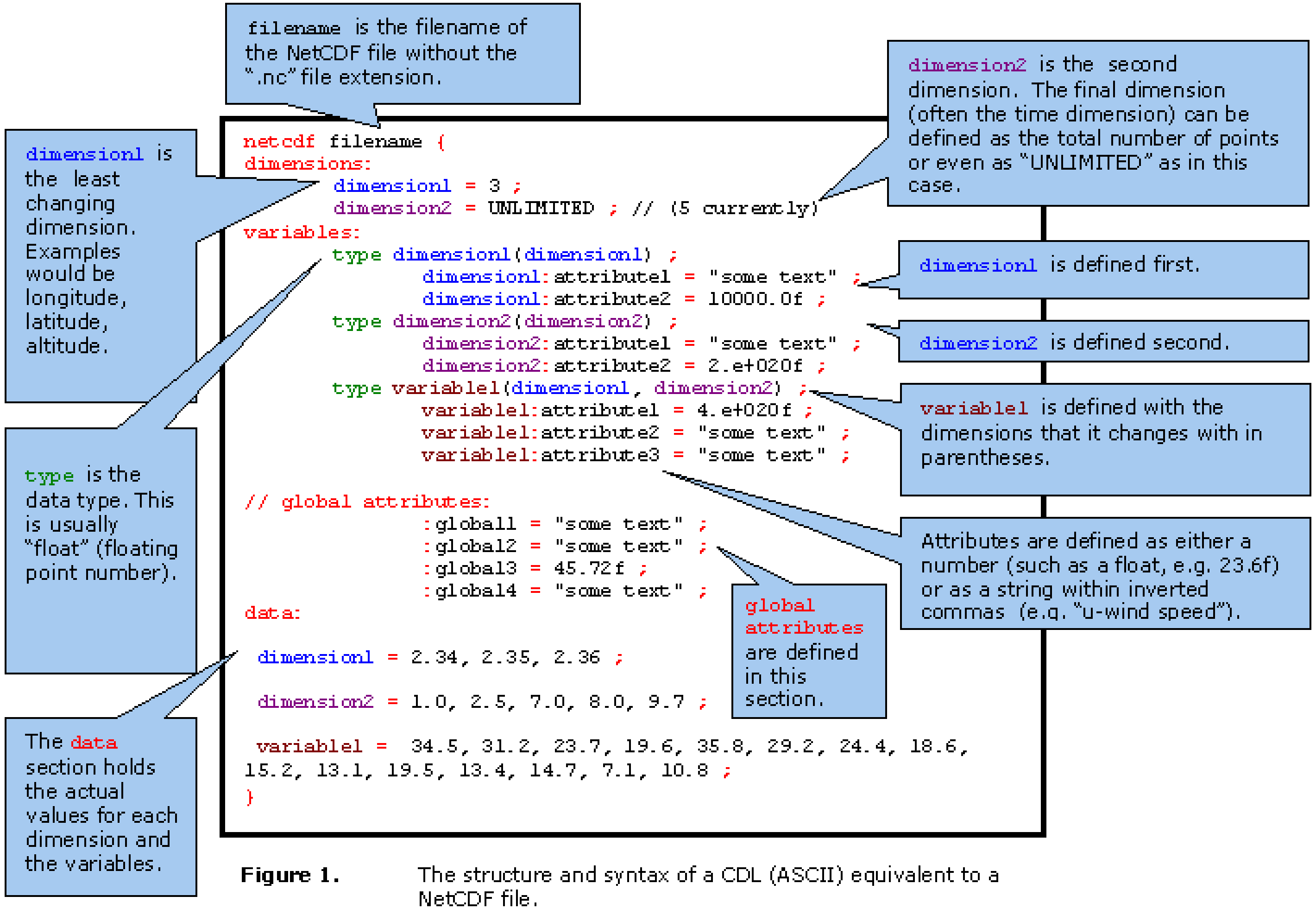

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

http://xkcd.com/927/

Curt Tilmes, NASA, 2013 ESIP Material

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.



```
netcdf filename {
dimensions:
        dimension1 = 3 ;
        dimension2 = UNLIMITED ; // (5 currently)
variables:
        type dimension1(dimension1) ;
                dimension1:attribute1 = "some text" ;
                dimension1:attribute2 = 10000.0f ;
        type dimension2(dimension2) ;
                dimension2:attribute1 = "some text" ;
                dimension2:attribute2 = 2.e+020f ;
        type variable1(dimension1, dimension2) ;
                variable1:attribute1 = 4.e+020f ;
                variable1:attribute2 = "some text" ;
                variable1:attribute3 = "some text" ;

// global attributes:
                :global1 = "some text" ;
                :global2 = "some text" ;
                :global3 = 45.72f ;
                :global4 = "some text" ;
data:

 dimension1 = 2.34, 2.35, 2.36 ;

 dimension2 = 1.0, 2.5, 7.0, 8.0, 9.7 ;

 variable1 =  34.5, 31.2, 23.7, 19.6, 35.8, 29.2, 24.4, 18.6,
15.2, 13.1, 19.5, 13.4, 14.7, 7.1, 10.8 ;
}
```

filename is the filename of the NetCDF file without the ".nc" file extension.

dimension1 is the least changing dimension. Examples would be longitude, latitude, altitude.

dimension2 is the second dimension. The final dimension (often the time dimension) can be defined as the total number of points or even as "UNLIMITED" as in this case.

dimension1 is defined first.

dimension2 is defined second.

variable1 is defined with the dimensions that it changes with in parentheses.

type is the data type. This is usually "float" (floating point number).

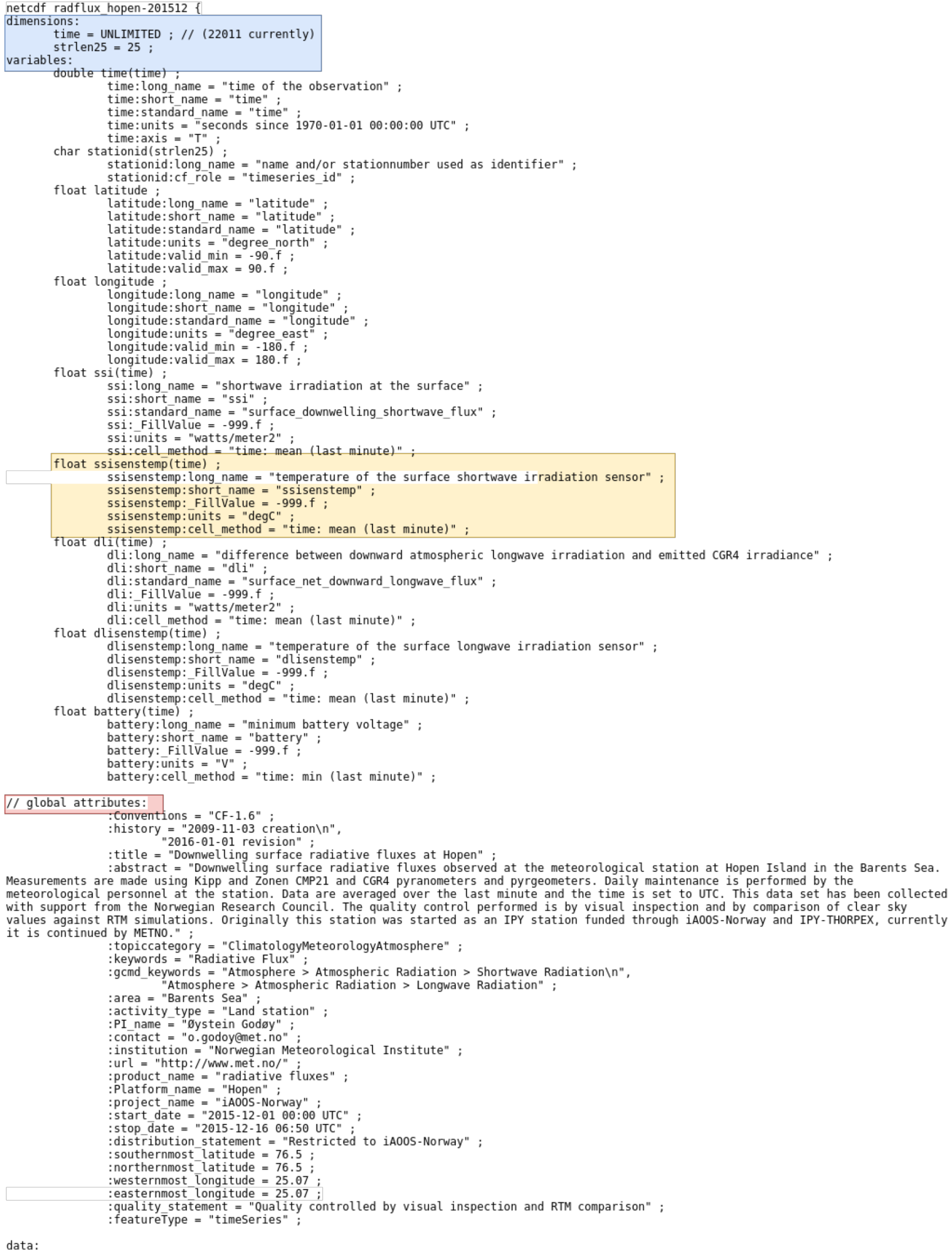
Attributes are defined as either a number (such as a float, e.g. 23.6f) or as a string within inverted commas (e.g. "u-wind speed").

global attributes are defined in this section.

The data section holds the actual values for each dimension and the variables.

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



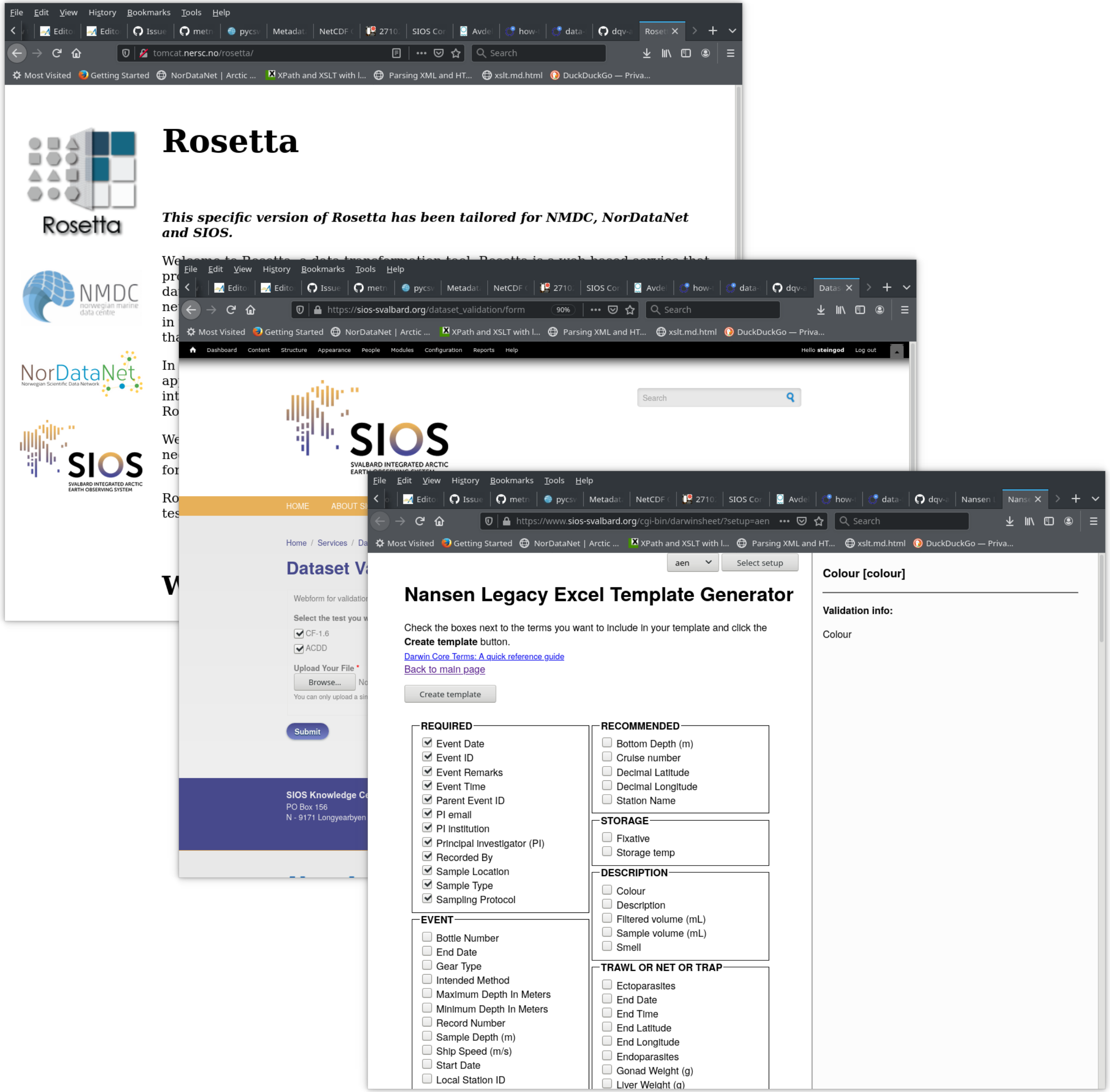
```
netcdf radflux_hopen-201512 {
dimensions:
        time = UNLIMITED ; // (22011 currently)
        strlen25 = 25 ;
variables:
        double time(time) ;
                time:long_name = "time of the observation" ;
                time:short_name = "time" ;
                time:standard_name = "time" ;
                time:units = "seconds since 1970-01-01 00:00:00 UTC" ;
                time:axis = "T" ;
        char stationid(strlen25) ;
                stationid:long_name = "name and/or stationnumber used as identifier" ;
                stationid:cf_role = "timeseries_id" ;
        float latitude ;
                latitude:long_name = "latitude" ;
                latitude:short_name = "latitude" ;
                latitude:standard_name = "latitude" ;
                latitude:units = "degree_north" ;
                latitude:valid_min = -90.f ;
                latitude:valid_max = 90.f ;
        float longitude ;
                longitude:long_name = "longitude" ;
                longitude:short_name = "longitude" ;
                longitude:standard_name = "longitude" ;
                longitude:units = "degree_east" ;
                longitude:valid_min = -180.f ;
                longitude:valid_max = 180.f ;
        float ssi(time) ;
                ssi:long_name = "shortwave irradiation at the surface" ;
                ssi:short_name = "ssi" ;
                ssi:standard_name = "surface_downwelling_shortwave_flux" ;
                ssi:_FillValue = -999.f ;
                ssi:units = "watts/meter2" ;
                ssi:cell_method = "time: mean (last minute)" ;
        float ssisenstemp(time) ;
                ssisenstemp:long_name = "temperature of the surface shortwave irradiation sensor" ;
                ssisenstemp:short_name = "ssisenstemp" ;
                ssisenstemp:_FillValue = -999.f ;
                ssisenstemp:units = "degC" ;
                ssisenstemp:cell_method = "time: mean (last minute)" ;
        float dli(time) ;
                dli:long_name = "difference between downward atmospheric longwave irradiation and emitted CGR4 irradiance" ;
                dli:short_name = "dli" ;
                dli:standard_name = "surface_net_downward_longwave_flux" ;
                dli:_FillValue = -999.f ;
                dli:units = "watts/meter2" ;
                dli:cell_method = "time: mean (last minute)" ;
        float dlisenstemp(time) ;
                dlisenstemp:long_name = "temperature of the surface longwave irradiation sensor" ;
                dlisenstemp:short_name = "dlisenstemp" ;
                dlisenstemp:_FillValue = -999.f ;
                dlisenstemp:units = "degC" ;
                dlisenstemp:cell_method = "time: mean (last minute)" ;
        float battery(time) ;
                battery:long_name = "minimum battery voltage" ;
                battery:short_name = "battery" ;
                battery:_FillValue = -999.f ;
                battery:units = "V" ;
                battery:cell_method = "time: min (last minute)" ;

// global attributes:
                :Conventions = "CF-1.6" ;
                :history = "2009-11-03 creation\n",
                        "2016-01-01 revision" ;
                :title = "Downwelling surface radiative fluxes at Hopen" ;
                :abstract = "Downwelling surface radiative fluxes observed at the meteorological station at Hopen Island in the Barents Sea.
Measurements are made using Kipp and Zonen CMP21 and CGR4 pyranometers and pyrgeometers. Daily maintenance is performed by the
meteorological personnel at the station. Data are averaged over the last minute and the time is set to UTC. This data set has been collected
with support from the Norwegian Research Council. The quality control performed is by visual inspection and by comparison of clear sky
values against RTM simulations. Originally this station was started as an IPY station funded through iAOOS-Norway and IPY-THORPEX, currently
it is continued by METNO." ;
                :topiccategory = "ClimatologyMeteorologyAtmosphere" ;
                :keywords = "Radiative Flux" ;
                :gcmd_keywords = "Atmosphere > Atmospheric Radiation > Shortwave Radiation\n",
                        "Atmosphere > Atmospheric Radiation > Longwave Radiation" ;
                :area = "Barents Sea" ;
                :activity_type = "Land station" ;
                :PI_name = "Øystein Godøy" ;
                :contact = "o.godoy@met.no" ;
                :institution = "Norwegian Meteorological Institute" ;
                :url = "http://www.met.no/" ;
                :product_name = "radiative fluxes" ;
                :Platform_name = "Hopen" ;
                :project_name = "iAOOS-Norway" ;
                :start_date = "2015-12-01 00:00 UTC" ;
                :stop_date = "2015-12-16 06:50 UTC" ;
                :distribution_statement = "Restricted to iAOOS-Norway" ;
                :southernmost_latitude = 76.5 ;
                :northernmost_latitude = 76.5 ;
                :westernmost_longitude = 25.07 ;
                :easternmost_longitude = 25.07 ;
                :quality_statement = "Quality controlled by visual inspection and RTM comparison" ;
                :featureType = "timeSeries" ;

data:
```

Documentation standards relevant for SIOS

- Biological data
  - Darwin Core Archive
- Geophysical data
  - NetCDF adhering to the Climate and Forecast convention
- Else a non proprietary file format with an comprehensive product manual
  - Identifying variables, units, missing values, aggregation levels, positions, reference frames, etc.

Supporting tools

- Existing
  - Rosetta for conversion between spreadsheet data and NetCDF-CF
  - NetCDF-CF validator
  - Excel spreadsheet generator
- Required
  - Should we e.g. look into MeteoIO?
  - ??



tomcat.nersc.no/rosetta/
Rosetta
NMDC
NorDataNet
SIOS
https://sios-svalbard.org/dataset_validation/form
Dashboard Content Structure Appearance People Modules Configuration Reports Help
Hello steingod Log out
Search
HOME ABOUT S
Home / Services / Da
https://www.sios-svalbard.org/cgi-bin/darwinsheet/?setup=aen
aen
Select setup
Create template
Browse...
Submit

Rosetta

This specific version of Rosetta has been tailored for NMDC, NorDataNet and SIOS.

Dataset V

Webform for validatio

Select the test you w

☑ CF-1.6

☑ ACDD

Upload Your File *

You can only upload a si

SIOS Knowledge C
PO Box 156
N - 9171 Longyearbyen

Nansen Legacy Excel Template Generator

Check the boxes next to the terms you want to include in your template and click the **Create template** button.

Darwin Core Terms: A quick reference guide

Back to main page

REQUIRED
- ☑ Event Date
- ☑ Event ID
- ☑ Event Remarks
- ☑ Event Time
- ☑ Parent Event ID
- ☑ PI email
- ☑ PI Institution
- ☑ Principal Investigator (PI)
- ☑ Recorded By
- ☑ Sample Location
- ☑ Sample Type
- ☑ Sampling Protocol

EVENT
- □ Bottle Number
- □ End Date
- □ Gear Type
- □ Intended Method
- □ Maximum Depth In Meters
- □ Minimum Depth In Meters
- □ Record Number
- □ Sample Depth (m)
- □ Ship Speed (m/s)
- □ Start Date
- □ Local Station ID

RECOMMENDED
- □ Bottom Depth (m)
- □ Cruise number
- □ Decimal Latitude
- □ Decimal Longitude
- □ Station Name

STORAGE
- □ Fixative
- □ Storage temp

DESCRIPTION
- □ Colour
- □ Description
- □ Filtered volume (mL)
- □ Sample volume (mL)
- □ Smell

TRAWL OR NET OR TRAP
- □ Ectoparasites
- □ End Date
- □ End Time
- □ End Latitude
- □ End Longitude
- □ Endoparasites
- □ Gonad Weight (g)
- □ Liver Weight (g)

Colour [colour]

Validation info:

Colour

Also we need to agree on a mechanism to identify SIOS Core Data when harvesting information from contributing data centres

Preferably by adding a specific project to these datasets?