Data management

Data Management Plans
Data Formats
Metadata
Repositories

Acknowledgements: Raymond Roman (UCT-SEA)

IOC Guide 73:
Guidelines for a Data Management Plan

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

File organization: Naming conventions (Active research phase)

Make life easier

Naming conventions should be:

- Descriptive
- Consistent

Consider including:

- Unique identifier (ie. Project Name or Grant # in folder name)
- Project or research data name
- Conditions (Lab instrument, Solvent, Temperature, etc.)
- Run of experiment (sequential) • Date (in file properties too)
- Version # 29

Eg. A2003135.L2_LAC_SST.data

20040527127O3-NCEI-L3C_GHRSST-SSTskin-AVHRR_Pathfinder-PFV5.3_NOAA17_G_2004148_day-v02.0-fv01.0.nc

File organization: Naming conventions

Make life easier

Naming conventions should be:

- Descriptive
- Consistent

Eg.

YYYYMMDD

MMDDYYYY

YYMMDD

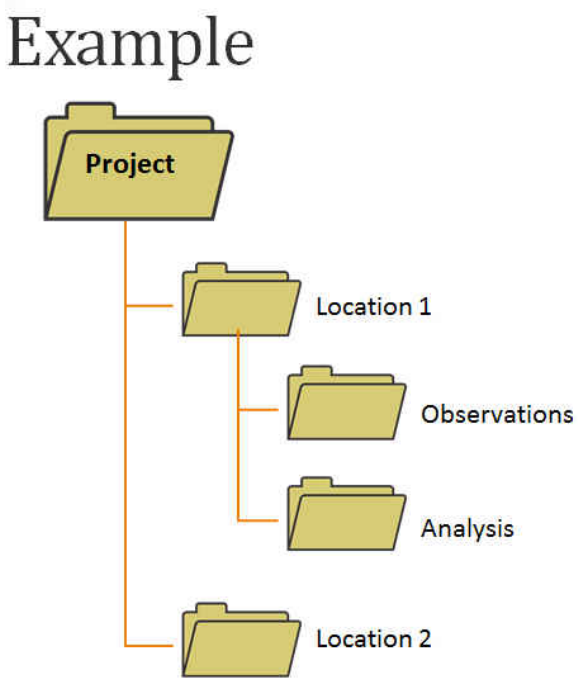
- Use date format ISO 8601: YYYYMMDD

File organization: File structure

Hierarchical file structures can add additional organization to your files.

Some considerations are:

1. Project
2. Date
3. Analysis
4. Location



File organization: versioning

Track versions of:

- ❖Analysis/program/scripts while keeping the original version of the data file the same or
- ❖Data files themselves

Things to document:

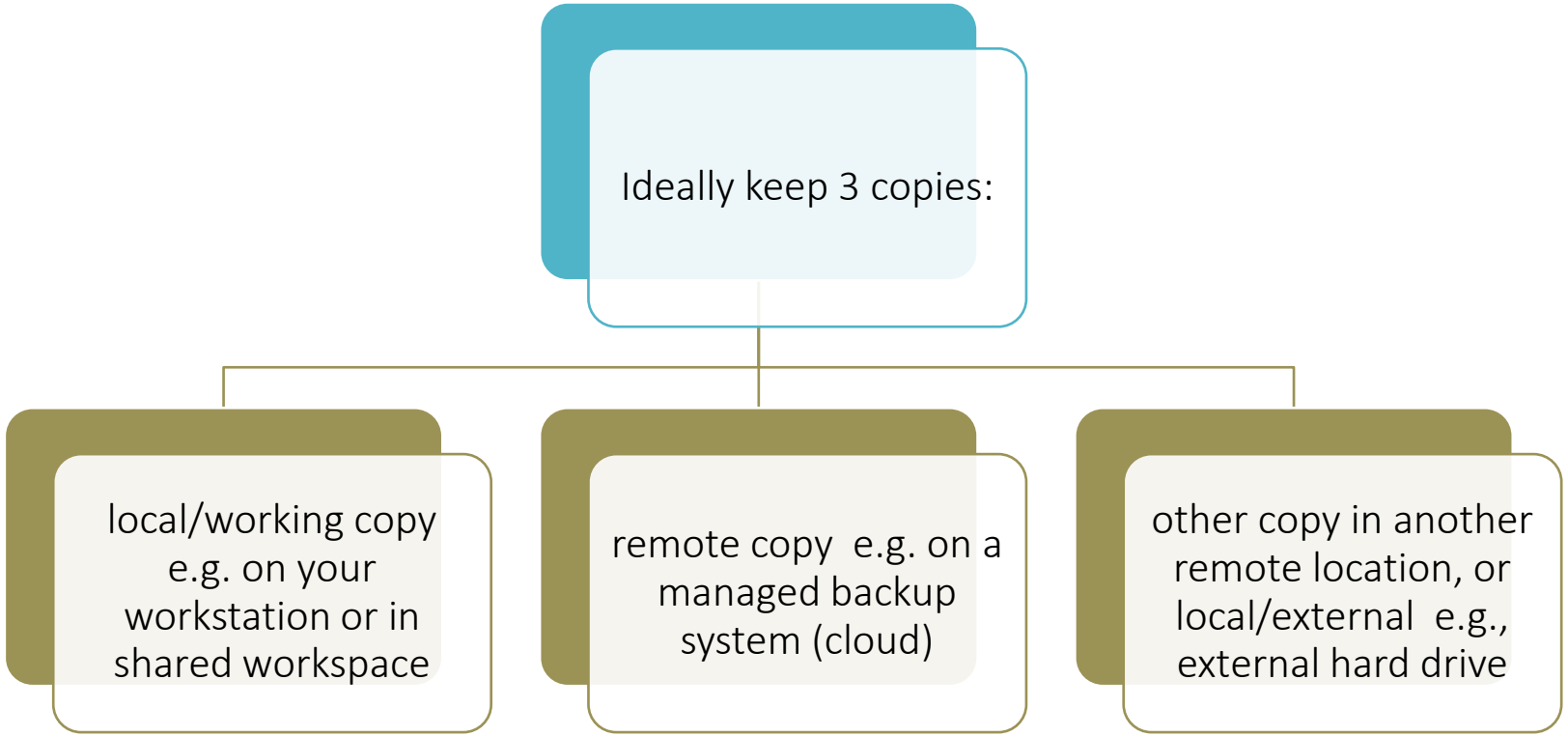
- •What was changed?
- •Who is responsible?
- •When did it happen?
- •Why?

IOC Guide 73:
Guidelines for a Data Management Plan

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

Storage (Active phase backup)



IOC Guide 73:
Guidelines for a Data Management Plan

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

What do we mean by data?

- General

1)Images

2)Videos

3)Numerical measurements

- Social Sciences

1)survey responses

2)focus group and individual interview transcripts

3)opinion polling

- Natural Sciences

1)measurements generated by sensors/laboratory instruments

2)computer modelling

3)Simulations

4)observations and/or field studies

5)specimen

File formats

- Best case is using non-proprietary (open) formats
- ✓Preferred formats

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

File formats: ASCII

- ASCII
  - American Standard character for Information Interchange

➔**PROS**: ASCII files is easily exchangeable and human readable. ASCII files make it simple to quickly open files written from acquisitions and view data immediately as well as to easily share the data.

➔**CONS**:

1)ASCII files, however, have several drawbacks, including a large disk footprint, which can be an issue when storage space is limited

2)Reading and writing data from an ASCII file can be significantly slower compared to other formats

File formats: Binary

- Binary

→**PROS**:

1)In contrast to ASCII files, binary files have a significantly smaller disk footprint

2)It can be streamed to disk at extremely high speeds, making them ideal for high-channel-count and real-time applications

→**CONS:**

1) It is not human-readable and therefore difficult to exchange between users

2)To share the files with colleagues, you must provide them with an application that interprets your specific binary file correctly

File formats : xml

- XML
  - **PROS**: XML format has been gaining in popularity due to its ability to store complex data structures
  - With XML files, you can store data and formatting along with the raw measurement values
  - Using the flexibility of the XML format, you can store additional information with the data in a structured manner
  - XML is also relatively human-readable and exchangeable
  - **CONS**: in its raw form, XML includes tags within the file that describe the structures which somewhat limits the readability
  - The weakness of the XML file format is that it has an extremely large disk footprint compared to other files and cannot be used to stream data directly to disk
  - a downside to being able to store these complex structures is that they may require considerable planning when you design the layout, or schema, of the XML structures

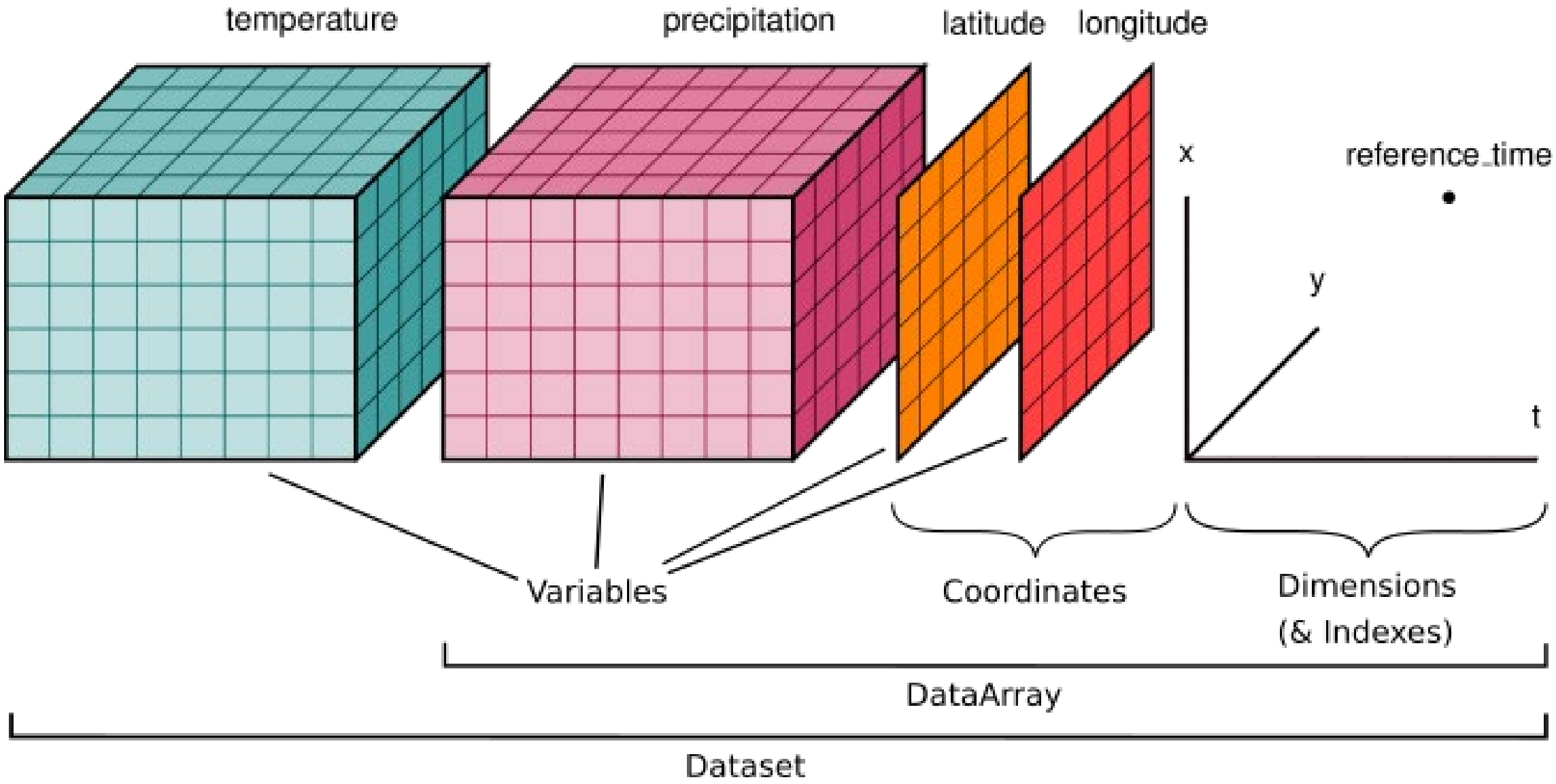
File formats: NetCDF

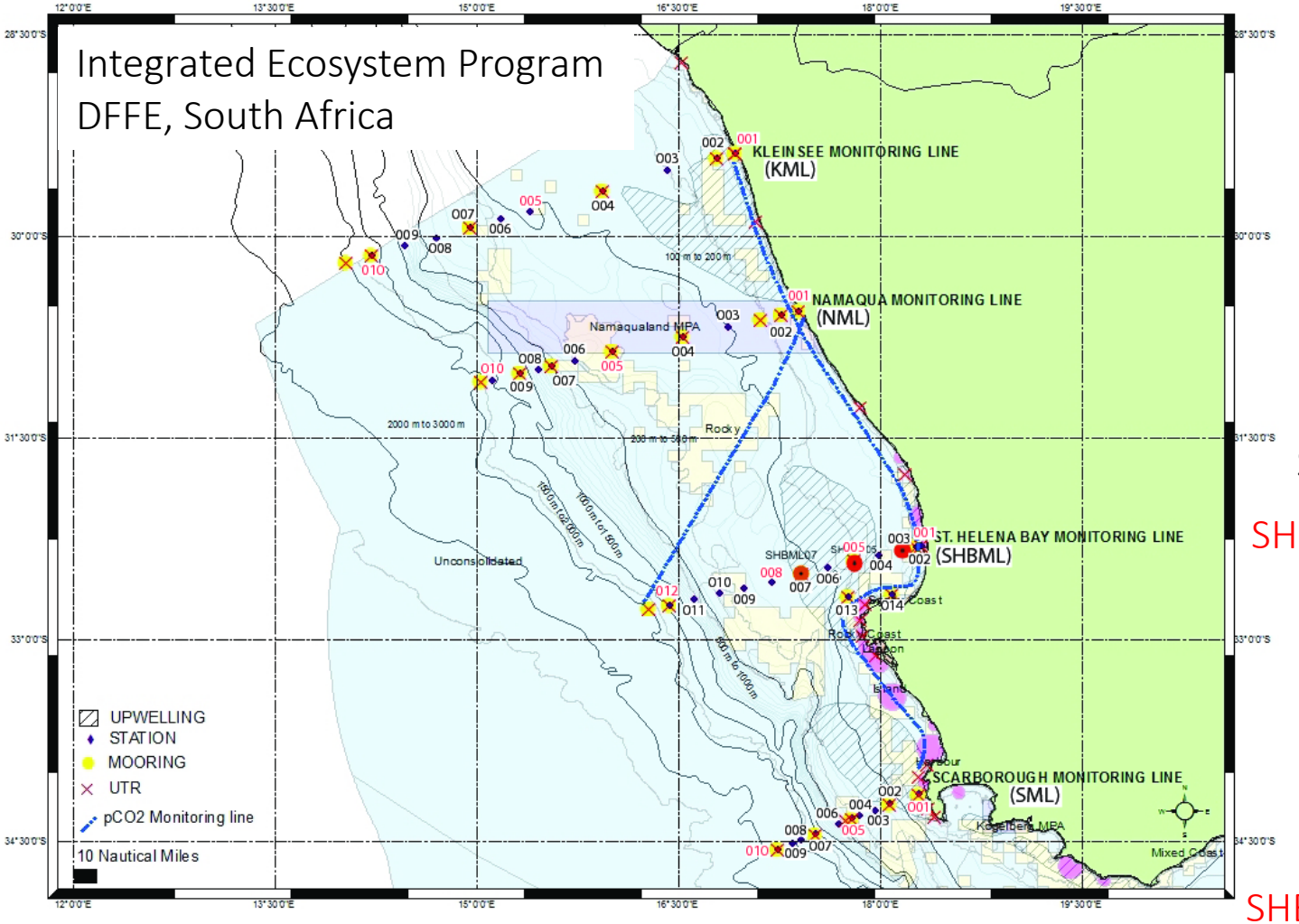
- NetCDF (Network Common Data Format)
  - NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.
  - The classic NetCDF data model consists of **variables**, **dimensions**, and **attributes**.
  - NetCDF is a subset of the Hierarchical Data Format HDF5

NetCDF data is:

- Self-Describing. A NetCDF file includes information about the data it contains.
- Portable. A NetCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- Scalable. A small subset of a large dataset may be accessed efficiently.
- Appendable. Data may be appended to a properly structured NetCDF file without copying the dataset or redefining its structure.
- Sharable. One writer and multiple readers may simultaneously access the same NetCDF file.
- Archivable. Access to all earlier forms of NetCDF data will be supported by current and future versions of the software.

The NetCDF data model (and xarray's)





	Lat	Lon	Temperature	…
SML-001				
…				
SML-010				
SHBML-001				
…				
SHBML-012				
NML-001				
…				

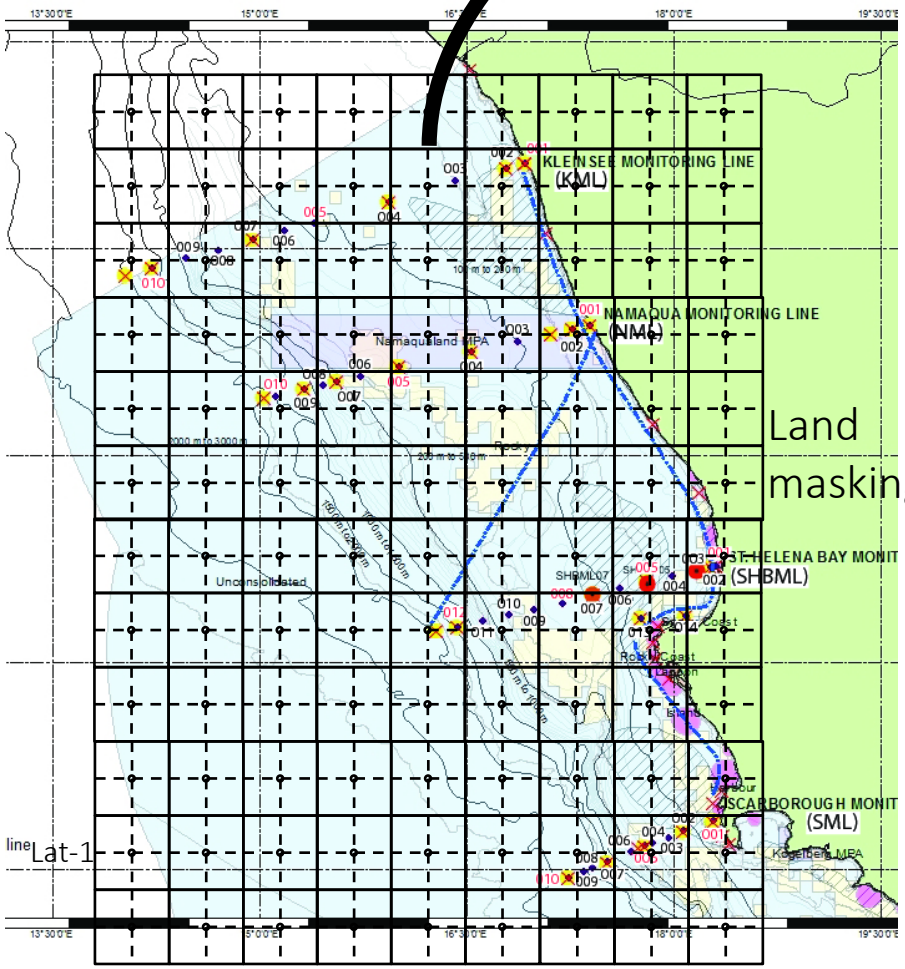
Tabular representation of station data

This is a 2D array, with a give ordering of the stations

Essential raw data, but:

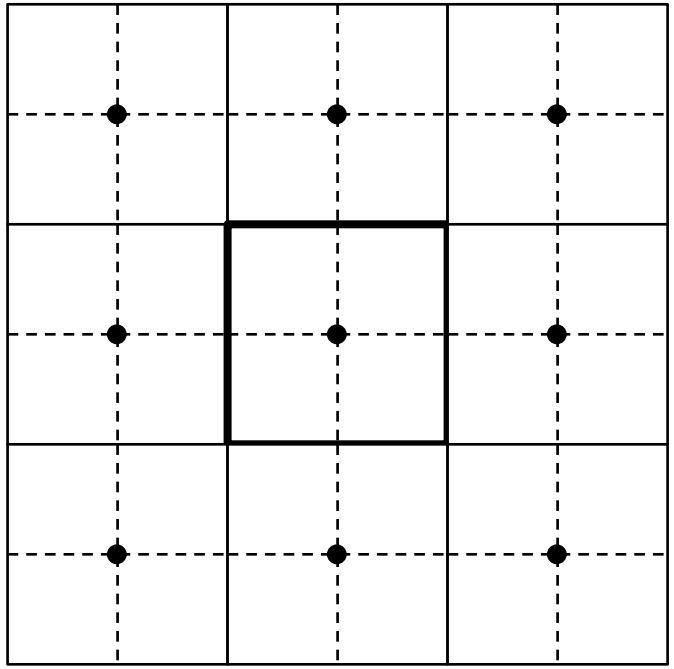
- require consistent structure (time, vertical levels)
- difficult to slice and group (but we have pandas)

GRIDDING



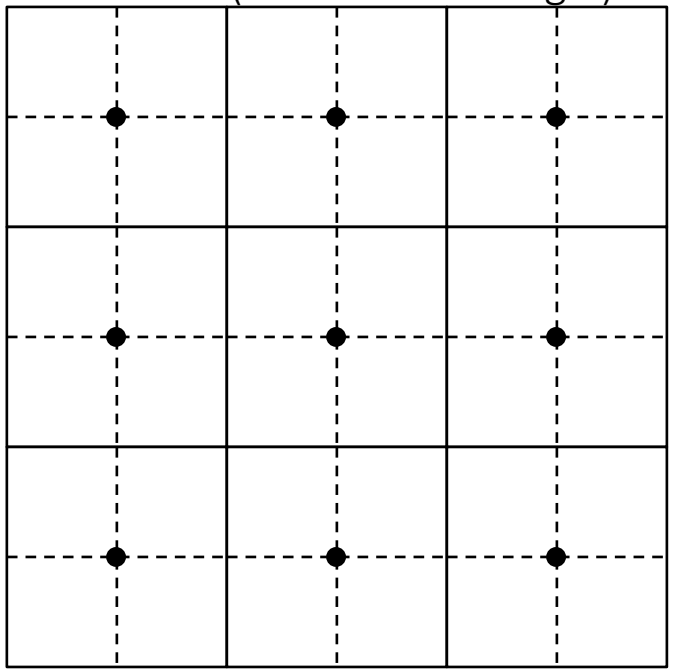
Land masking

Sea Surface Temperature (SST)

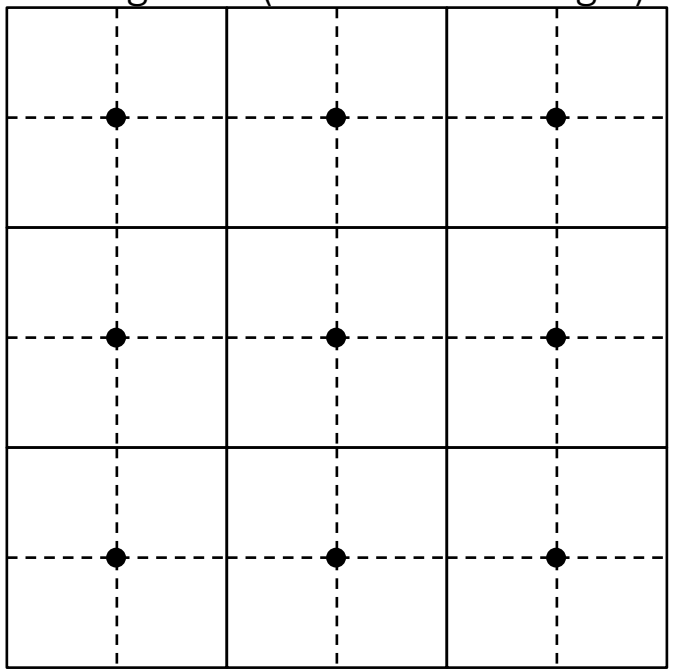


Land grid point
= missing value

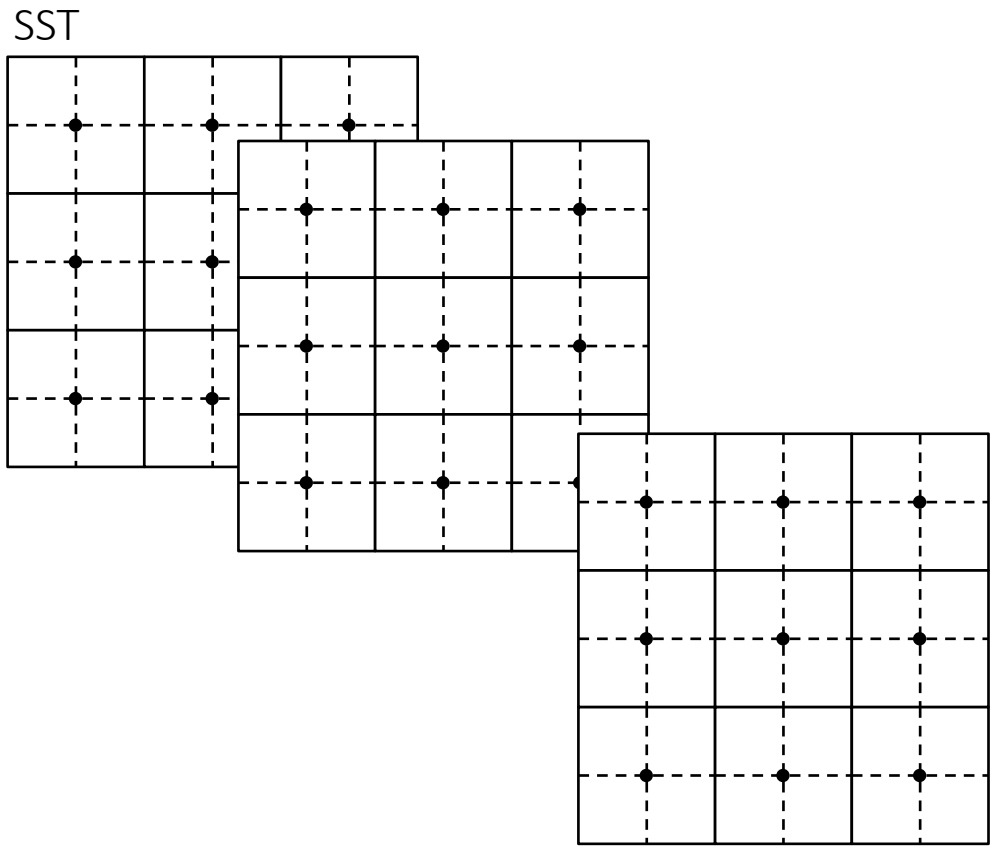
Latitude (or location along Y)

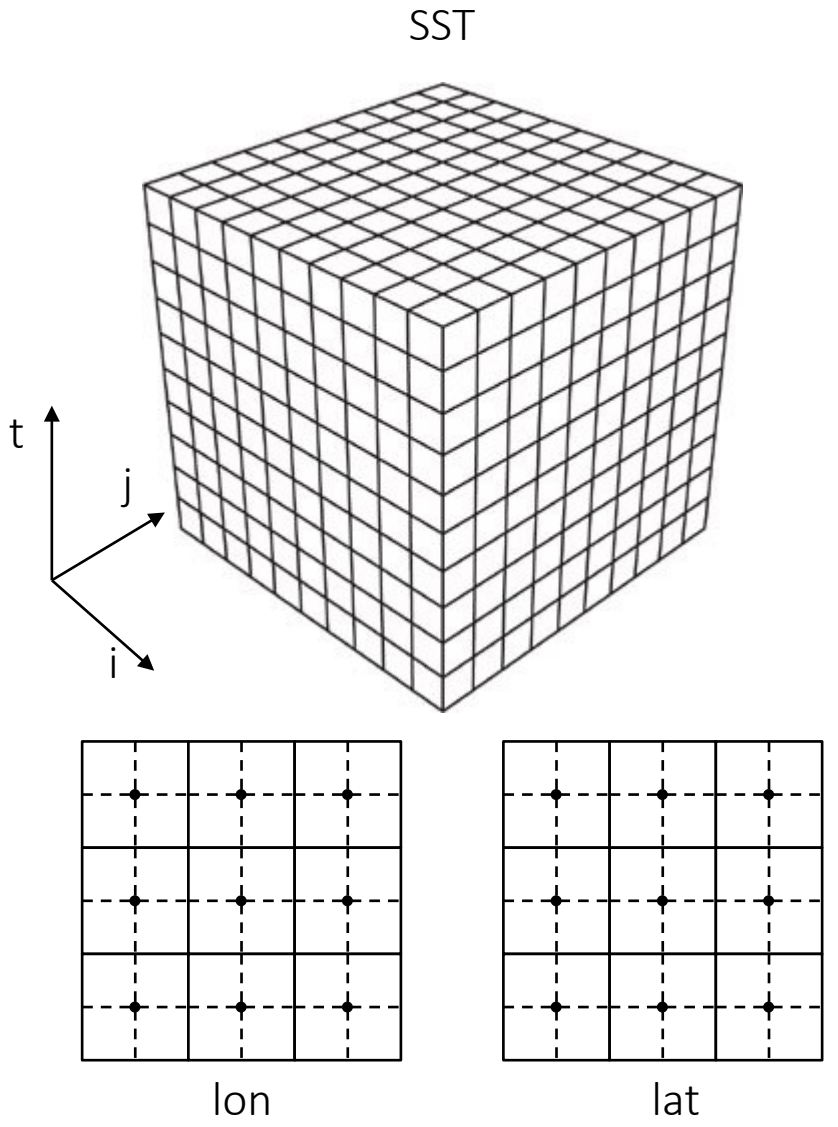


Longitude (or location along X)

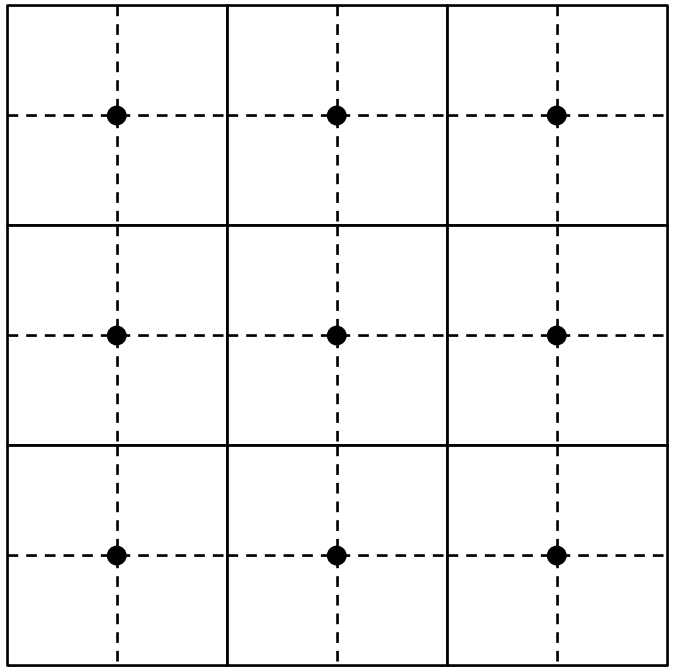


Multidimensional arrays

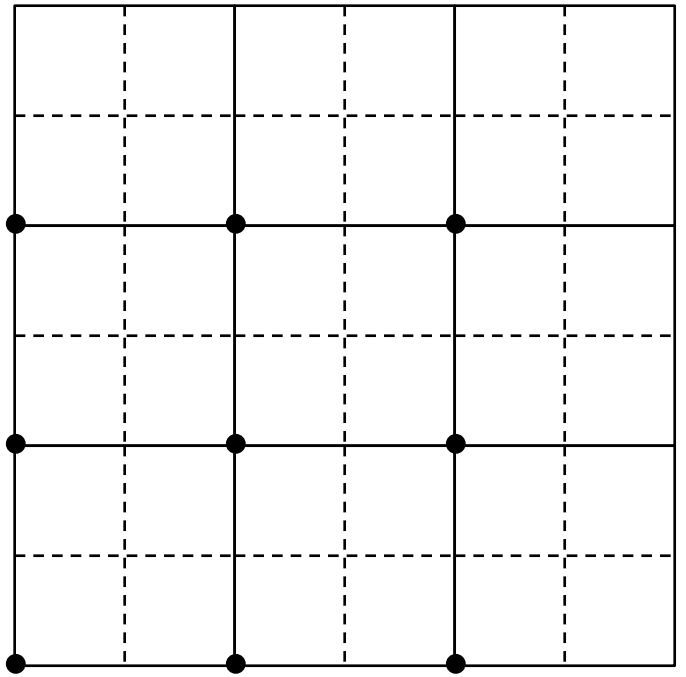




Understanding grids and pseudocolor mesh



Numerical grid



Pseudocolor grid (python, matlab, etc.)

Note: in matlab, the uppermost row and rightmost column are not shown when showing a pseudocolor mesh

File formats: PDF-A

- PDF-A
  - ISO-standardized version of the Portable Document Format (PDF) specialized for use in the archiving and long-term preservation of electronic documents
  - PDF/A differs from PDF by prohibiting features ill-suited to long-term archiving, such as font linking (as opposed to font embedding) and encryption
  - The ISO requirements for PDF/A file viewers include color management guidelines, support for embedded fonts, and a user interface for reading embedded annotations
  - A PDF/A document must embed all fonts in use; accordingly, a PDF/A file will often be larger than an equivalent PDF file that does not include embedded fonts

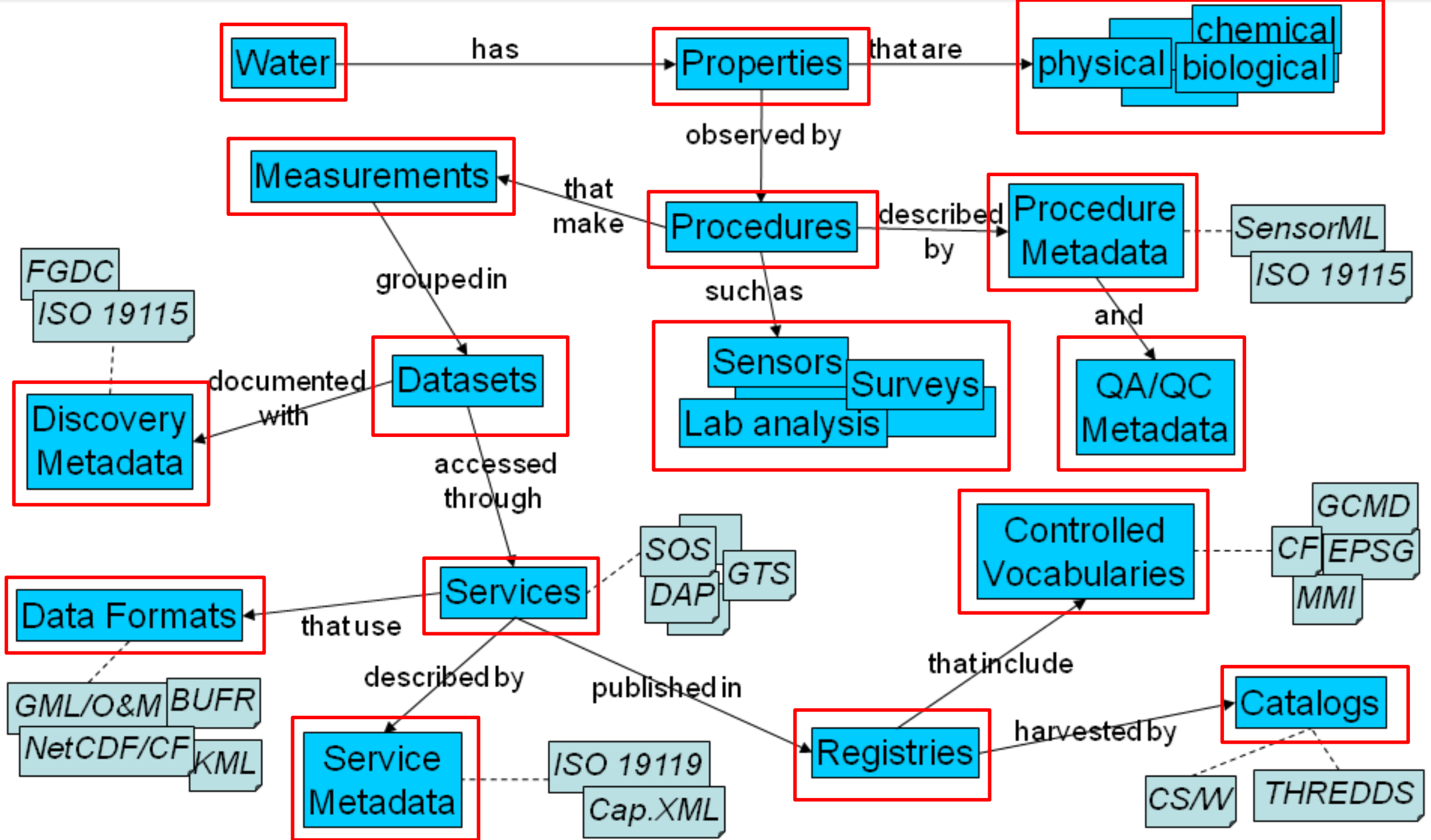
IOC Guide 73:
Guidelines for a Data Management Plan

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

Applied Ocean Science

Importance of Metadata


Water
has
Properties
that are
physical
chemical
biological
observed by
Measurements
that make
Procedures
described by
Procedure Metadata
SensorML
ISO 19115
grouped in
such as
and
FGDC
ISO 19115
Datasets
documented with
Discovery Metadata
Sensors
Surveys
Lab analysis
QA/QC Metadata
accessed through
Controlled Vocabularies
GCMD
CF
EPSG
MMI
SOS
GTS
DAP
Services
Data Formats
that use
GML/O&M
BUFR
NetCDF/CF
KML
described by
published in
that include
Service Metadata
ISO 19119
Cap.XML
Registries
harvested by
Catalogs
CSW
THREDDS

Data, finally…

_, _, _, _, _, _, _, _, _, _, _, _, _, _,
-179, -180, -180, -180, -180, -180, -180, -179, -179, -179, -179, -179,
-180, -180, -180, -180, -180, -180, -180, -179, -179, -179, -179, -179,
-179, -179, -179, -179, -179, -179, -180, -180, -180, -180, -180, -180,
-180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180,
-180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180,
-180, -180, -180, -180, -180, -180, -180, -180, -179, -179, -179, -179,
-178, -177, -176, -176, -175, -174, -173, -174, -175, -176, -175, -174,
-173, -172, -171, -169, -166, -165, -165, -166, -168, -169, -170, -171,
-171, -171, -171, -171, -169, -167, -165, -164, -163, -163, -162, -160,
-159, -160, -162, -162, -162, -160, -157, -154, -153, -153, -158, -163,
-165, -165, -165, -166, -170, -171, -172, -171, -167, -163, -161, -160,
-158, -158, -158, -158, -157, -158, -161, -161, -156, -153, -154, -157,
-159, -158, -161, _, _, _, _, _, _, _, _, _, _, _, _, _, _, -168,
-169, -170, -170, -169, -168, -167, _, _, _, _, _, _, _, _, _, _, _, _, _,

WHAT IS IT ?!?!

Data with Metadata

```
netcdf \20131010-NCDC-L4LRblend-GLOB-v01-fv02_0-AVHRR_OI {
dimensions:
        lat = 320 ;
        lon = 320 ;
        time = 1 ;
variables:
        float lat(lat) ;
                lat:_FillValue = NaNf ;
                lat:long_name = "latitude" ;
                lat:standard_name = "latitude" ;
                lat:axis = "Y" ;
                lat:units = "degrees_north" ;
                lat:comment = "uniform grid from -89.875 to 89.875 by 0.25" ;
        float lon(lon) ;
                lon:_FillValue = NaNf ;
                lon:long_name = "longitude" ;
                lon:standard_name = "longitude" ;
                lon:axis = "X" ;
                lon:units = "degrees_east" ;
                lon:comment = "uniform grid from -179.875 to 179.875 by 0.25" ;
int time(time) ;
                time:long_name = "reference time of sst field" ;
                time:standard_name = "time" ;
                time:axis = "T" ;
                time:units = "seconds since 1981-01-01" ;
                time:calendar = "proleptic_gregorian" ;
short analysed_sst(time, lat, lon) ;
                analysed_sst:_FillValue = -32768s ;
                analysed_sst:long_name = "analysed sea surface temperature" ;
                analysed_sst:standard_name = "sea_surface_temperature" ;
analysed_sst:units = "kelvin" ;
                analysed_sst:valid_min = -300s ;
                analysed_sst:valid_max = 4500s ;
                analysed_sst:add_offset = 273.15f ;
                analysed_sst:scale_factor = 0.01f ;
 byte mask(time, lat, lon) ;
                mask:_FillValue = -128b ;
                mask:long_name = "sea/land field composite mask" ;
                mask:flag_values = 1b ;
                mask:flag_meanings = "sea land lake ice" ;
                mask:comment = "b0:1=grid cell is open sea water b1:1=land is present in this grid
cell b2:1=lake surface is present in this grid cell b3:1=sea ice is present in this grid cell b4-
b7:reserve for future grid mask data" ;
 byte sea_ice_fraction(time, lat, lon) ;
                sea_ice_fraction:_FillValue = -128b ;
                sea_ice_fraction:long_name = "sea ice area fraction" ;
                sea_ice_fraction:standard_name = "sea ice area fraction" ;
                sea_ice_fraction:units = "percent" ;
                sea_ice_fraction:valid_min = 0b ;
                sea_ice_fraction:valid_max = 100b ;
                sea_ice_fraction:add_offset = 0.f ;
                sea_ice_fraction:scale_factor = 0.01f ;
```

// global attributes:
:Conventions = "CF-1.0" ;
:title = "Daily-OI-V2, Final, Data (Ship, Buoy, AVHRR: NOAA19, METOP, NCEP-ice)" ;
:DSD_entry_id = "NCDC-L4LRblend-GLOB-AVHRR_OI" ;
:references = "Reynolds, et al.(2007) Daily High-resolution Blended Analyses. Available at ftp://eclipse.ncdc.noaa.gov/pub/OI-daily/daily-sst.pdf" ;
:institution = "NOAA/NESDIS/NCDC" ;
:contact = "Richard.W.Reynolds@noaa.gov & Chunying.liu@noaa.gov" ;
:GDS_version_id = "v1.0-rev1.7" ;
:netcdf_version_id = "3.6.0-p1 of Jul 4 2005 16:41:16 $" ;
:creation_date = "2013-10-25" ;
:product_version = "Version 2.0" ;
:history = "Version 2.0" ;
:spatial_resolution = "0.25 degree" ;
:start_date = "2013-10-10 UTC" ;
:start_time = "00:00:00 UTC" ;
:stop_date = "2013-10-11 UTC" ;
:stop_time = "00:00:00 UTC" ;
:westernmost_longitude = -179.875f ;
:easternmost_longitude = 179.875f ;
:southernmost_latitude = -89.875f ;
:northernmost_latitude = 89.875f ;
:file_quality_index = "0" ;
:source_data = "NCEP GTS, AVHRR19, METOP, NCEP ice" ;
:comment = "WARNING Some applications are unable to properly handle signed byte values. If values are encountered > 127, please subtract 256 from this reported value" ;

data:

analysed_sst =
_, _,
_, _,
_, _,
_, _,
_, _,
_, _,
_, _,
_, _,
_, _, _, _, _, _, _, _, _, _, _, _, _, _,
-179, -180, -180, -180, -180, -180, -180, -179, -179, -179, -179, -179,
-180, -180, -180, -180, -180, -180, -180, -179, -179, -179, -179, -179,
-179, -179, -179, -179, -179, -179, -180, -180, -180, -180, -180, -180,
-180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180,
-180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180, -180,
-180, -180, -180, -180, -180, -180, -180, -180, -179, -179, -179, -179,
-178, -177, -176, -176, -175, -174, -173, -174, -175, -176, -175, -174,
-173, -172, -171, -169, -166, -165, -165, -166, -168, -169, -170, -171,
-171, -171, -171, -171, -169, -167, -165, -164, -163, -163, -162, -160,
-159, -160, -162, -162, -162, -160, -157, -154, -153, -153, -158, -163,
-165, -165, -165, -166, -170, -171, -172, -171, -167, -163, -161, -160,
-158, -158, -158, -158, -157, -158, -161, -161, -156, -153, -154, -157,
-159, -158, -161, _, _, _, _, _, _, _, _, _, _, _, _, _, _, _, -168,
-169, -170, -170, -169, -168, -167, _, _, _, _, _, _, _, _, _, _, _, _,

Metadata

Definition – Data about data (structured information that tells us who, what, when, where and how)

Why?

- Make data understandable (through keywords and attributes)
- Enables data sharing now and in the future
- Facilitates long-term archival preservation of data
- Helps others discover, access, use, repurpose and cite your data in the long term
- Various metadata standards tailored to specific disciplines

Metadata (cont.)

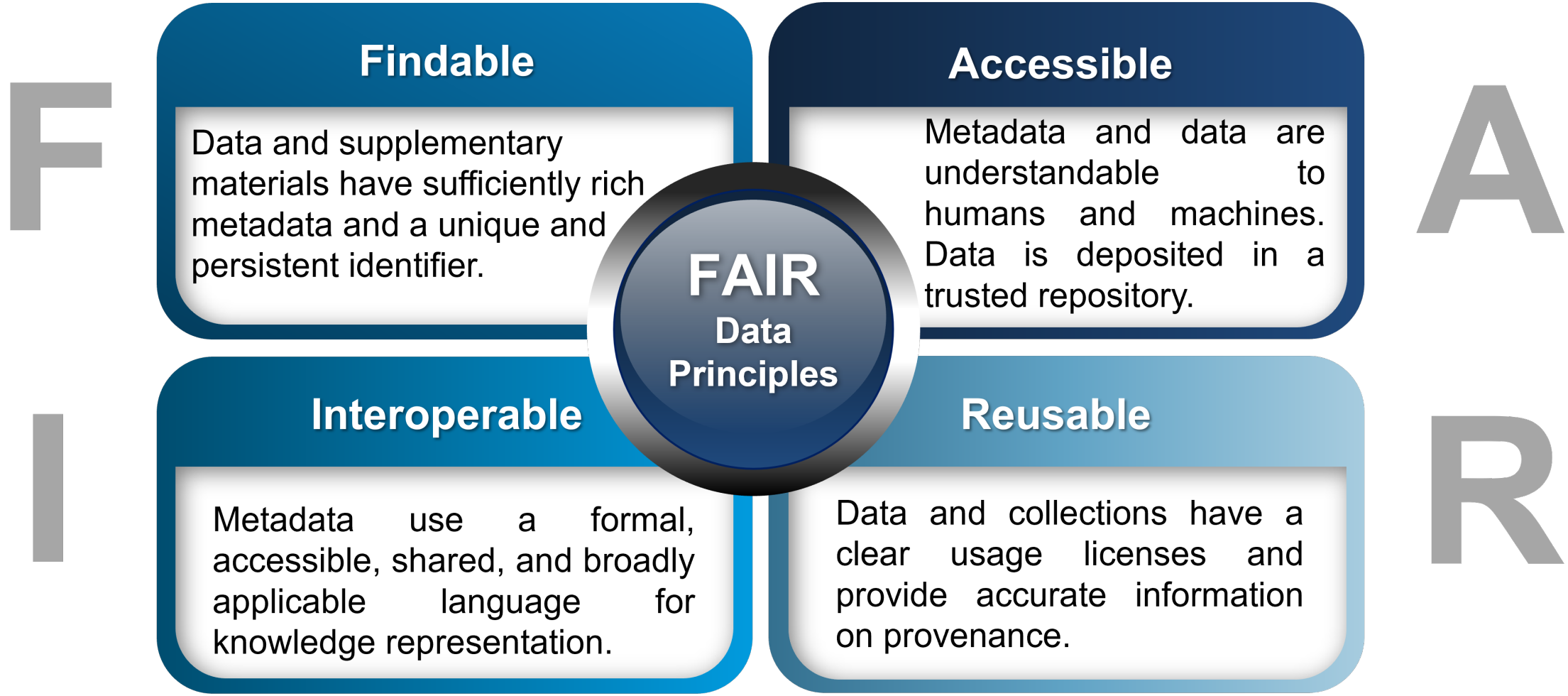
Metadata standards:

- Dublin Core- https://www.dublincore.org
- Darwin Core- https://www.tdwg.org/standards/dwc
- ABCD (Access to Biological Collection Data) https://abcd.tdwg.org
- AVMS (Astronomy Visualization Metadata Standard)
- CSDGM (Content Standard for Digital Geospatial Metadata)
- CF (Climate and Forecasting) Metadata Conventions http://cfconventions.org

Advantage of using these standards:

- Ensure you have a complete, standard set of information about each part of your data
- Enable your dataset to be organized with other datasets

https://www.go-fair.org/fair-principles/



IOC Guide 73:
Guidelines for a Data Management Plan

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

Data sharing: Copyright and Licensing, DOIs

- **Copyright** is a legal right created by the law of a country that grants the creator of an original work exclusive rights for its use and distribution.
  Copyright implications, intellectual property and other legal restrictions on use: An Archive will honor all applicable legal restrictions. These issues occur when the OAIS acts as a custodian. An OAIS should understand the intellectual property rights concepts, such as copyrights and any other applicable laws prior to accepting copyrighted materials into the OAIS. It can establish guidelines for ingestion of information and rules for dissemination and duplication of the information when necessary.

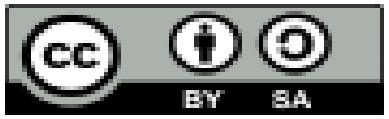
- A **licence** is an official permission or permit to do, use, or own something (as well as the document of that permission or permit).

- A **Digital Object Identifier** or **DOI** is a persistent identifier or handle used to uniquely identify objects, standardized by ISO-26324. DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports and data sets, and official publications.
  https://www.doi.org/

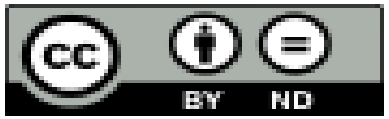
Extracts from Wikipedia

Regularly Used Licenses

Freeing content globally without restrictions CC0



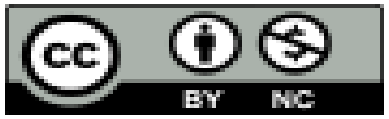
Attribution + ShareAlike BY-SA



Attribution + NoDerivatives BY-ND

Attribution + Noncommercial + ShareAlike BY-NC-SA

Attribution + Noncommercial + NoDerivatives BY-NC-ND



Attribution + Noncommercial BY-NC

Attribution alone BY

Wikipedia

Licensing

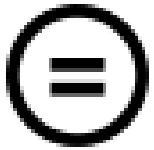
SAEON recommends Creative Commons Licenses

- Choice may be subject to DEA / National policy

Attribution (BY) Licensees may copy, distribute, display and perform the work and make derivative works and remixes based on it only if they give the author or licensor the credits (attribution) in the manner specified by these.

Share-alike (SA) Licensees may distribute derivative works only under a license identical ("not more restrictive") to the license that governs the original work. (See also copyleft.) Without share-alike, derivative works might be sublicensed with compatible but more restrictive license clauses, e.g. CC BY to CC BY-NC.)

Non-commercial (NC)
Licensees may copy, distribute, display, and perform the work and make derivative works and remixes based on it only for non-commercial purposes.



No Derivative Works (ND)
Licensees may copy, distribute, display and perform only verbatim copies of the work, not derivative works and remixes based on it.

IOC Guide 73:
Guidelines for a Data Management Plan

A data management plan should address the questions:

- What data will be generated by the activity?
- What procedures will be used to manage the data?
- Which file format(s) will be used for the data?
- How will changes in the data files be tracked?
- Where will the data be stored?
- Who will have access to the data?
- How will the data be documented (metadata)?
- Will the data be available in a repository?
- What archive and long-term retention solutions are planned?

DOIs: Digital Object Identifiers

https://doi.org/10.7289/V5WD3XHB

https://www.datacite.org

https://geo-ide.noaa.gov/wiki/index.php?title=DOI_Minting_Procedure

https://zenodo.org/communities/scale_south_africa

Example: http://www.digitalservices.lib.uct.ac.za/dls/rdm-policy

Data publication sites and DOIs

https://www.pangaea.de/

https://www.datacite.org

https://data.aad.gov.au/

https://www.bodc.ac.uk/

https://zenodo.org

https://www.epfl.ch/campus/library/services-researchers/data-publication/data-code-journals/