Standardization of data and metadata

Lara Ferrighi, Norwegian Meteorological Institute

Data Interoperability Workshop, 2022-05-31

Data and Metadata

The concept of data and metadata can be at times unclear

- Data is a collection of information, such as observations, measurements, computations of models etc…
  - It can be used to analyze trends and patterns, to extract and visualize actual values of variables
- Metadata, on the other hands, is data about the data, i.e. they extensively provide description about the data they are attached to
  - It gives the necessary context to the user to be able to access, understand and use the data correctly
  - Several types of metadata can and should be used to describe the data

Types of metadata

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. GCW is required to expose this information through WMO Information System as well. Discovery metadata are thus WIS metadata, although the GCW portal can translate to WIS for those not using WMO standards directly.	ISO19115 GCMD DIF ACDD
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 Darwin Core Archive
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. This information is maintained by the GCW portal and is not covered by discovery or use metadata standards.	Used locally by data centres
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

What is in a metadata standard?

- A metadata standard is made up of defined elements, including the type of information the user should enter (e.g. text, numbers, date).
- Examples of elements include Title, Abstract, Keyword, Online Link
- Multiple standard exists and they are linked to the type of metadata they address and the communities they target
- Terminology for the same concepts may vary across standards (values of mapping)

Adopting 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 (meta)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 using 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.

Why using community standards?

- Consider your **archive**:
  - Do they have any recommendations?
- Consider your **users**:
  - Who wants this data? Why do they want it?
  - What do they want to do with it?
  - Will they be using your data in concert with other data?
- Consider **heritage**:
  - What worked well for similar data in the past?
  - What could be done better for newly created data?
- Consider **tools**:
  - Try to use data formats supported by the software you intend to use it with.

Filling in Metadata Standards

I need to fill in a metadata element about:

Dataset Production Status: Describes the production status of the data set regarding its completeness.

What should I put in there?

Metadata file:

<Title>The title of the dataset</Title>
<Abstract> This dataset collects…</Abstract>
<Dataset_Production_Status>XXX</Dataset_Production_Status>
<Start_Date>2020-01-20</Start_Date>

Filling in Metadata Standards

I need to fill in a metadata element about:

Dataset Production Status: Describes the production status of the data set regarding its completeness.

Data provider A can use:	**Data provider B can use:**
“Not ready yet”	“Not finished”
“Done”	“Finished and stored”
“Still acquiring data”	“unknown”
“Continuously updating”	“Not started yet”
“???”	“See www.mydataset.com”

They will all pass and we are left with unmanageable information

Controlled *vocabulary*

controlled vocabularies are a **source of authoritative terms** to be entered for values of certain elements

Label	**Description**
Planned	Refers to data sets to be collected in the future and are thus unavailable at the present time.
In Work	Refers to data sets currently undergoing production or data that is continuously being collected or updated.
Complete	Refers to data sets in which no updates or further data collection will be made.
Obsolete	A new version of the dataset has been generated. The new version should be used, this is kept for back tracing.

Controlled *vocabulary*

controlled vocabularies are a source of authoritative terms to be entered for values of certain elements

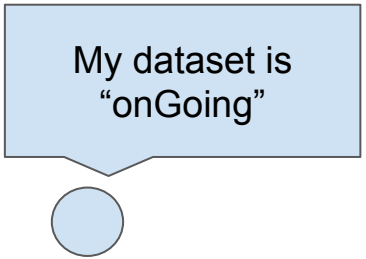
<Title>The title of the dataset</Title>
<Abstract> This dataset collects…</Abstract>
<Dataset_Production_Status>Complete</Dataset_Production_Status>
<Start_Date>2020-01-20</Start_Date>

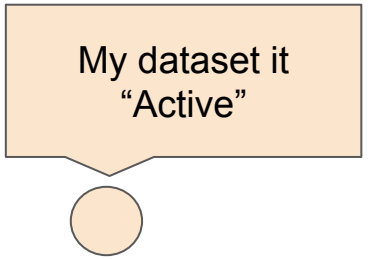
Label	
Planned	
In Work	
Complete	Refers to data sets in which no updates or further data collection will be made.
Obsolete	A new version of the dataset has been generated. The new version should be used, this is kept for back tracing.

Mapping between standards

Enumeration/Code List Mapping

Std A	Translation Direction	Std B
planned	↔	PLANNED
underDevelopment	→	PLANNED
onGoing	↔	ACTIVE
completed	↔	COMPLETE
historicalArchive	→	COMPLETE
obsolete	→	COMPLETE
retired	→	DEPRECATED
deprecated	→	DEPRECATED
NOT APPLICABLE a string is used instead of the defined codes. The codeList="" and codeListValue = ""	↔	NOT APPLICABLE
Blank or doesn't exist	→	NOT PROVIDED
Any other value	→	NOT PROVIDED
Don't translate	←	NOT PROVIDED





Only through the use of standards, understanding both languages is possible.

Profiles

Some standards can have different profiles, i.e. a generic standard is adapted to specific requirements of a community, leading to a specific profile of that standard.

A specific element of a standard can be within a profile:

- from optional to mandatory attributes
- from a string to a controlled vocabulary
- from a controlled vocabulary to a subset of it

Discovery Metadata - ACDD Convention

When encoding data as netCDF/CF is good practise to include discovery metadata in the file using the Attribute Convention for dataset Discovery (ACDD). Discovery metadata will then be directly connected to the data themselves and can be extracted for ingestion in the searchable catalogue.

Index by Attribute Name

- acknowledgement (Recommended)
- cdm_data_type (Suggested)
- comment (Recommended)
- contributor_name (Suggested)
- contributor_role (Suggested)
- Conventions (**Highly Recommended**)
- coverage_content_type (**Highly Recommended**) [Variable]
- creator_email (Recommended)
- creator_institution (Suggested)
- creator_name (Recommended)
- creator_type (Suggested)
- creator_url (Recommended)
- date_created (Recommended)
- date_issued (Suggested)
- date_metadata_modified (Suggested)
- date_modified (Suggested)
- geospatial_bounds (Recommended)
- geospatial_bounds_crs (Recommended)
- geospatial_bounds_vertical_crs (Recommended)
- geospatial_lat_max (Recommended)
- geospatial_lat_min (Recommended)
- geospatial_lat_resolution (Suggested)
- geospatial_lat_units (Suggested)
- geospatial_lon_max (Recommended)
- geospatial_lon_min (Recommended)
- geospatial_lon_resolution (Suggested)
- geospatial_lon_units (Suggested)
- geospatial_vertical_max (Recommended)
- geospatial_vertical_min (Recommended)
- geospatial_vertical_positive (Recommended)
- geospatial_vertical_resolution (Suggested)
- geospatial_vertical_units (Suggested)
- history (Recommended)
- id (Recommended)
- institution (Recommended)
- instrument (Suggested)
- instrument_vocabulary (Suggested)
- keywords (**Highly Recommended**)
- keywords_vocabulary (Suggested)
- license (Recommended)
- long_name (**Highly Recommended**) [Variable]
- metadata_link (Suggested)
- naming_authority (Recommended)
- platform (Suggested)
- platform_vocabulary (Suggested)
- processing_level (Recommended)
- product_version (Suggested)
- program (Suggested)
- project (Recommended)
- publisher_email (Recommended)
- publisher_institution (Suggested)
- publisher_name (Recommended)
- publisher_type (Suggested)
- publisher_url (Recommended)
- references (Suggested)
- source (Recommended)
- standard_name (**Highly Recommended**) [Variable]
- standard_name_vocabulary (Recommended)
- summary (**Highly Recommended**)
- time_coverage_duration (Recommended)
- time_coverage_end (Recommended)
- time_coverage_resolution (Recommended)
- time_coverage_start (Recommended)
- title (**Highly Recommended**)
- units (**Highly Recommended**) [Variable]

Use Metadata - Climate and Forecast (CF) convention

For proper interpretation of data

- Standardised naming of variables
- Units of variables
  - date ISO8601
- Encoding of missing values

CF Standard Name Table

Version 78, 21 September 2021

Refer to the Guidelines for Construction of CF Standard Names for information on how the names are constructed and interpreted, and how new names could be derived.

A note about units
The canonical units associated with each standard name are usually the SI units for the quantity. Section 3.3 of the CF conventions states: "Unless it is dimensionless, a variable with a standard_n units which are physically equivalent (not necessarily identical) to the canonical units, possibly modified by an operation specified by either the standard name modifier ... or by the cell_methods a Section 1.3 of the CF conventions states: "The values of the units attributes are character strings that are recognized by UNIDATA's Udunits package [UDUNITS], (with exceptions allowed as disc "Units")." For example, a variable with the standard name of "air_temperature" may have a units attribute of "degree_Celsius" because Celsius can be converted to Kelvin by Udunits. For the full refer to section 6 of the Udunits documentation. Refer to the CF conventions for full details of the units attribute.

Search

Search Standard Names | Show All Standard Names

AND OR (separate search terms with spaces)
Also search help text

View by Category

Atmospheric Chemistry	Atmosphere Dynamics	Carbon Cycle	Cloud	Hydrology
Ocean Dynamics	Radiation	Sea Ice	Surface	

Standard Name	Canonical Units
acoustic_signal_roundtrip_travel_time_in_sea_water	s
aerodynamic_particle_diameter	m
aerodynamic_resistance	m-1 s
age_of_sea_ice	year
age_of_stratospheric_air	s
age_of_surface_snow	day
aggregate_quality_flag	1
air_density	kg m-3
air_equivalent_potential_temperature *alias:* equivalent_potential_temperature	K
air_equivalent_temperature *alias:* equivalent_temperature	K
air_potential_temperature	K
air_pressure	Pa

Site Metadata - WIGOS WMO Integrated Global Observing System

#	Category	Description
1	Observed variable	Specifies the basic datasets.
2	Purpose of observation	Specifies the main programme(s) and
3	Station/platform	Specifies the enviro equipment or remot
4	Environment	Describes the geog It also provides an unstructured element for additional meta-information that is considered relevant for adequate use of the data and that is not captured anywhere else in this standard.
5	Instruments and methods of observation	Specifies the method of observation and describes instrument(s) used to make the observation. If mu generate the observation, then this category shoul
6	Sampling	Specifies how sampling and/or analysis are used t observation or how a specimen is collected.
7	Data processing and reporting	Specifies how raw data are transferred into the observed variable and reported to the users.
8	Data quality	Specifies the data quality and traceability of the observation.
9	Ownership and data policy	Specifies who is responsible for the observation and owns it.
10	Contact	Specifies where information about the observation or dataset can be obtained.

Category	ID	Name	Definition	MCO	Phase
6. Sampling	6-01	Sampling procedures	Procedures involved in obtaining a sample	O	III
	6-02	Sample treatment	Chemical or physical treatment of sample prior to analysis	O	III
	6-03	Sampling strategy	The strategy used to generate the observed variable	O*	I
	6-04	Sampling time period	The period of time over which a measurement is taken	M#	III
	6-05	Spatial sampling resolution	Spatial resolution refers to the size of the smallest observable object. The intrinsic resolution of an imaging system is determined primarily by the instantaneous field of view of the sensor, which is a measure of the ground area viewed by a single detector element in a given instance in time	M#	II
	6-06	Temporal sampling interval	Time period between the beginning of consecutive sampling periods	M#	III
	6-07	Diurnal base time	Time to which diurnal statistics are referenced	C#	I
	6-08	Schedule of observation	Schedule of observation	M#	I

Code table: 6-03
Code table title: Sampling strategy

#	Name	Definition
6-03-1	Continuous	Sampling is done continuously, but not necessarily at regular time intervals. Sampling is integrating, i.e., none of the medium escapes observations..
6-03-2	Discrete	Sampling is done at regular time intervals for certain sampling periods that are smaller than the time interval. Sampling is not integrating, i.e., parts of the medium escape observation.
6-03-3	Event	Sampling is done at irregular time intervals.

File formats

- **Always choose open file formats**
  - Remember that data are to be handled in a 50-100 year perspective
- **Use self-describing formats**
  - You will not be around to answer questions forever
  - Well accepted way of archiving and disseminating scientific data.
  - Information describing the data contents of the file are embedded within the data file itself
- NetCDF – Network Common Data Form
  - Widely used by agencies (NASA and NOAA)
  - Climate and forecast (CF) metadata conventions help standardize some things into NetCDF in a common manner.

On spreadsheets

- Spreadsheet for computer readability and computability
  - Extra information is often lost in translation
- Avoid extra formatting
  - Merged cells, bold/italic, colors
  - Anything that has to do with visual formatting is not computer-readable.
- Aim at one table (one row for variables, the other for data points)
  - This helps computing the data
  - While a human can see the layout and interpret the tables as separate, the computer doesn't have eyes and won't understand that these are separate
  - Get rid of extra information (graphs/figures)

Values of standardization - bring home message

- Make people understand and use your data
  - Find a common language and/or translation between them
  - Support broader sharing
  - Increase discoverability/reusability
- Support multidisciplinary research
  - Interoperability (discovery/data level)
- Reduce costs
  - Reuse supported standards, no maintenance
  - Free tools to analyze and visualize for standardized data
- Support preservation
- Meet FAIR principle

Resources

- https://commons.esipfed.org/
- https://dmtclearinghouse.esipfed.org/browse
- DataONE Education: https://dataoneorg.github.io/Education/
- https://www.youtube.com/watch?v=r29LTAR1-vg
- https://ecorepsci.github.io/reproducible-science/spreadsheets.html