Network Common Data Form (CF-NetCDF) data standard

Øystein Godøy

Outline

- NetCDF
- CF and CF Discrete Sampling Geometries
- Briefly on the Attribute Convention for Dataset Discovery (ACDD)

Types of metadata for datasets

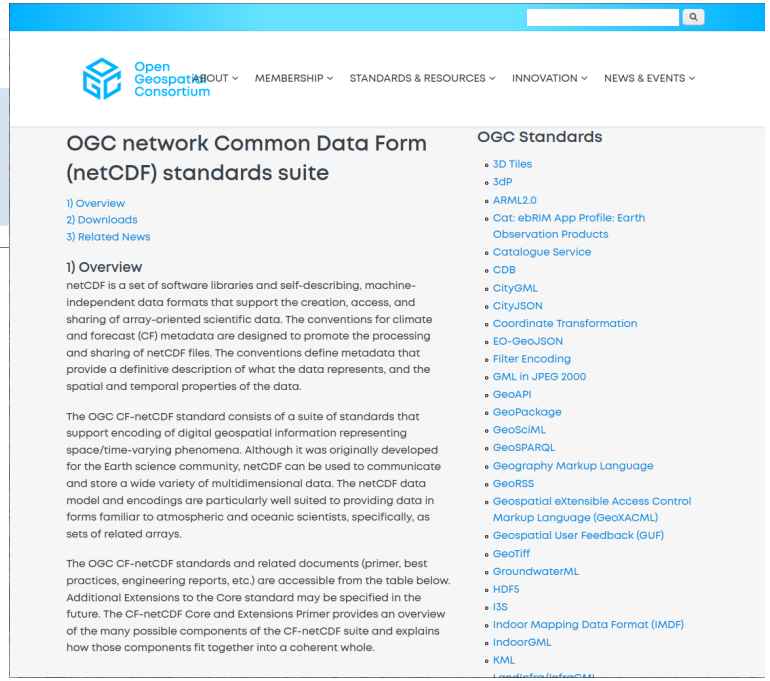
Type	Purpose	Description	Examples
Discovery metadata	Used to find relevant data	Discovery metadata are also called index metadata and are a digital version of the library index card. It describes who did what, where and when, how to access data and potential constraints on the data. It shall also link 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

Network Common Data Form (NetCDF)

- What is NetCDF?
  - 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.
  - NetCDF was developed and is maintained at Unidata.
- The Unidata Program Center supports and maintains netCDF programming interfaces for
  - C, C++, Java, and Fortran
- Programming interfaces are also available for
  - Python, IDL, MATLAB, R, Ruby, and Perl
- https://www.unidata.ucar.edu/software/netcdf/docs/faq.html

More on NetCDF

- Data in NetCDF format are
  - 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
    - Small subsets of large datasets in various formats may be accessed efficiently through NetCDF interfaces, even from remote servers.
  - 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.



ARCTIC PASSION INTER⇄ACT SIOS

NetCDF files are containers for Dimensions, Variables, and Global Attributes.

Is NetCDF a binary format only?

- NetCDF can be converted to/from ASCII representations using ncdump/ncgen
  - The ASCII representation of a NetCDF file is known as CDL (Common Data Language)
- In CDL comments Are indicated using "//"
  - They are not part of netCDF data.
  - To store comments about a variable or file, use a variable attribute or global attribute.
- NcML is an XML-based notation similar to CDL for netCDF data

```
netcdf example {// example of CDL notation
  dimensions:
      lon = 3 ;
      lat = 8 ;
  variables:
      float rh(lon, lat) ;
          rh:units = "percent" ;
          rh:long_name = "Relative humidity" ;
  // global attributes
      :title = "Simple example, lacks some
conventions" ;
  data:
   rh =
    2, 3, 5, 7, 11, 13, 17, 19,
    23, 29, 31, 37, 41, 43, 47, 53,
    59, 61, 67, 71, 73, 79, 83, 89 ;
  }
```

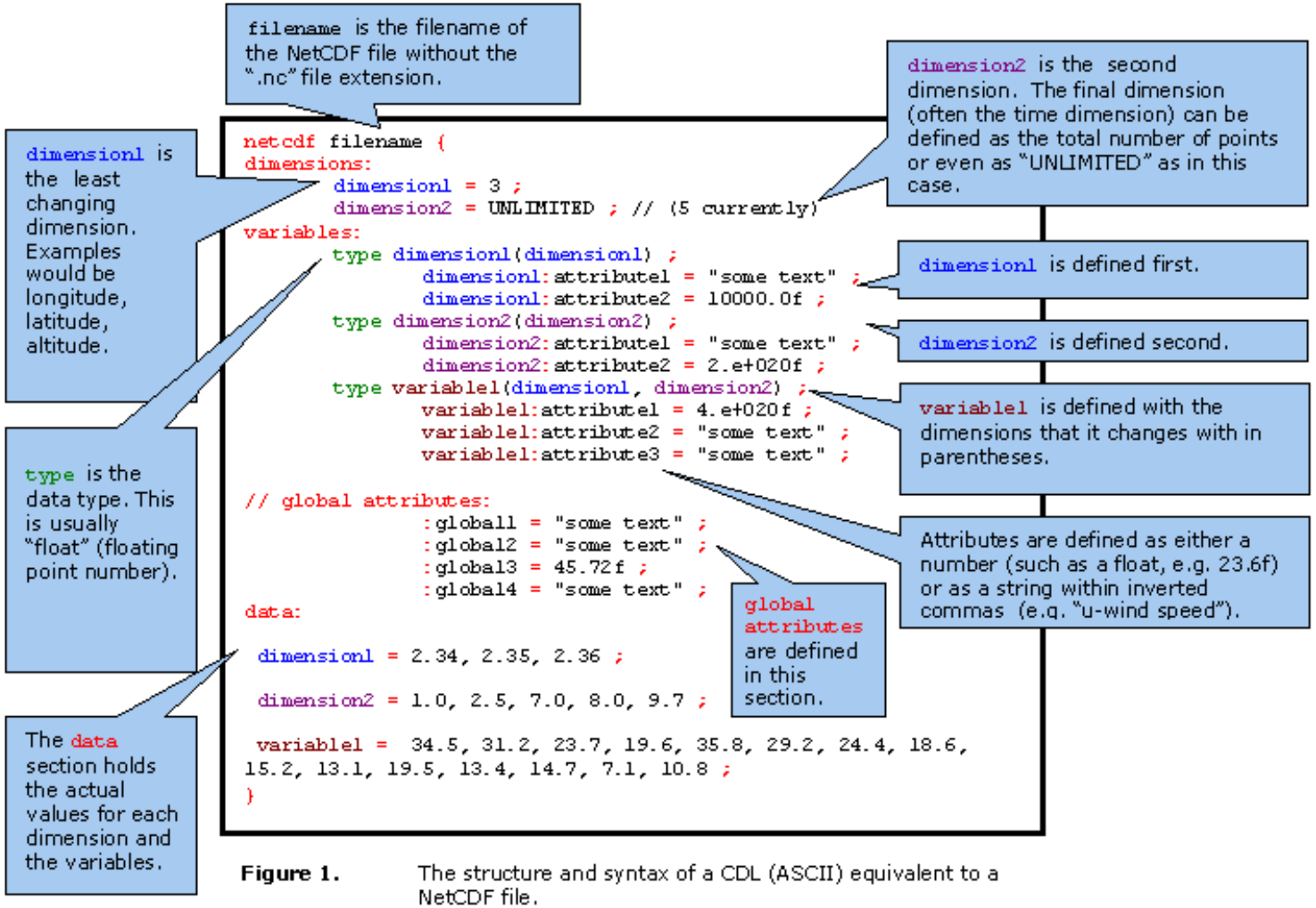


Figure 1. The structure and syntax of a CDL (ASCII) equivalent to a NetCDF file.

http://artefacts.ceda.ac.uk/formats/netcdf/index_cf.html

What are the Climate and Forecast Conventions?

- The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created with the NetCDF API.
- The CF conventions are increasingly gaining acceptance and have been adopted by a number of frameworks, projects and groups as a primary standard.
- **The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data.**
  - This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.
- The CF conventions generalize and extend the COARDS conventions.

- The ambition is to
  - **Locate data in space–time and as a function of other independent variables, to facilitate processing and graphics.**
  - **Identify data sufficiently to enable users of data from different sources to decide what is comparable, and to distinguish variables in archives**
- Framed as a NetCDF standard, but most CF ideas relate to metadata design in general and not specifically to NetCDF, and hence can be contained in other formats such as XML.

https://cfconventions.org/

Initially CF was developed for gridded data from climate and forecast models (hence “CF”) of the atmosphere and ocean, but its use has subsequently extended to other geosciences, and to observations as well as numerical models.

CF principles

- Data should be self-describing—no external tables needed to interpret it
- Conventions have been developed only for things we know we need
- Avoid being too onerous for data-writers and data-readers
- Metadata readable by humans as well as easily parsed by programs
- Minimise redundancy and possibilities for silly mistakes

- Avoid multiplicity of attributes
- Information is generally provided per-variable, not per-file
- Nothing depends on the names of variables (except the Unidata coordinate variable convention)

CF discovery level

- CF provides some basic “discovery” metadata in global attributes
  - title
    - What’s in the file
  - *institution
    - Where it was produced
  - *source
    - How it was produced e.g. model version, instrument type
  - history
    - Audit trail of processing operations
  - *references
    - Pointers to publications or web documentation
  - *comment
    - Miscellaneous

- *These ones can also be attributes of variables containing data
- But in general it is recommended to use ACDD elements instead for discovery information...

CF use metadata

- **units** is mandatory for all variables containing data other than dimensionless numbers.
  - The units do not identify the physical quantity.
    - units can be udunits strings e.g. 1, degC, Pa, mbar, W m-2, kg/m2/s, mm day^-1
    - or COARDS specials layer, level, sigma level.
    - Udunits doesn't support ppm, psu, dB, Sv.
- **standard name** identifies the quantity.
  - Units must be consistent with standard name and any statistical processing e.g. variance.
  - Standard name does not include coordinate or processing information.
- **long name** is not standardised.
- ancillary variables is a pointer to variables providing metadata about the individual data values e.g. standard error or data quality information.
- Numeric data variables may have _FillValue, missing value, valid max, valid min, valid range.
  - CF deprecates missing value.
- Variables containing "flag" values need flag values and flag meanings to make them self-describing.

NetCDF Climate and Forecast (CF) Metadata Conventions

Brian Eaton · Jonathan Gregory · Bob Drach · Karl Taylor · Steve Hankin · Jon Blower · John Caron · Rich Signell · Phil Bentley · Greg Rappa · Heinke Höck · Alison Pamment · Martin Juckes · Martin Raspaud · Randy Horne · Timothy Whiteaker · David Blodgett · Charlie Zender · Daniel Lee · David Hassell · Alan D. Snow · Tobias Kölling · Dave Allured · Aleksandar Jelenak · Anders Meier Soerensen · Lucile Gaultier · Sylvain Herlédan – Version 1.9, 10 September, 2021

Table of Contents

ARCTIC PASSION INTER

featureType	Description of a single feature with this discrete sampling geometry		Link
	Form of a data variable containing values defined on a collection of these features	**Mandatory space-time coordinates for a collection of these features**	
point	a single data point (having no implied coordinate relationship to other points)		
	data(i)	x(i) y(i) t(i)	Section H.1, "Point Data"
timeSeries	a series of data points at the same spatial location with monotonically increasing times		
	data(i,o)	x(i) y(i) t(i,o)	Section H.2, "Time Series Data"
trajectory	a series of data points along a path through space with monotonically increasing times		
	data(i,o)	x(i,o) y(i,o) t(i,o)	Section H.4, "Trajectory Data"
profile	an ordered set of data points along a vertical line at a fixed horizontal position and fixed time		
	data(i,o)	x(i) y(i) z(i,o) t(i)	Section H.3, "Profile Data"
timeSeriesProfile	a series of profile features at the same horizontal position with monotonically increasing times		
	data(i,p,o)	x(i) y(i) z(i,p,o) t(i,p)	Section H.5, "Time Series of Profiles"
trajectoryProfile	a series of profile features located at points ordered along a trajectory		
	data(i,p,o)	x(i,p) y(i,p) z(i,p,o) t(i,p)	Section H.6, "Trajectory of Profiles"

CF Standard Name Table

Version 79, 19 March 2022

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_name attribute must have 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 attribute." Furthermore, 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 discussed in Section 3.1, "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 range of supported units, refer to 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

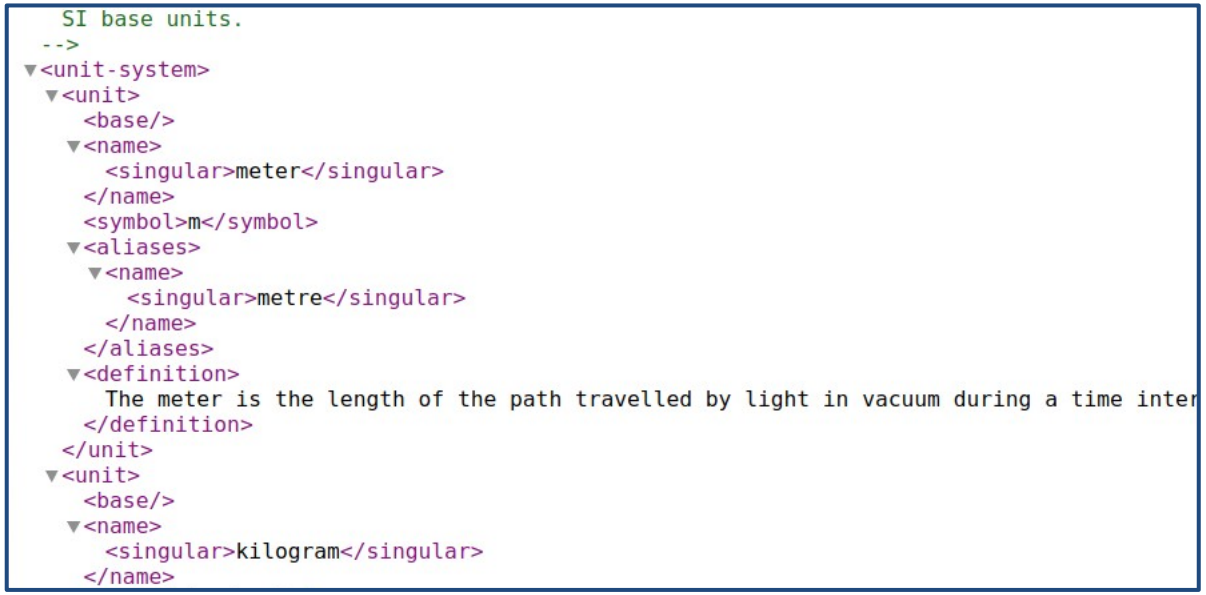
Atmospheric Chemistry	Atmosphere Dynamics	Carbon Cycle	Cloud	Hydrology
Ocean Dynamics	Radiation	Sea Ice	Surface	

Standard Name	Canonical Units	AMIP	GRIB
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	theta	13
air_pressure	Pa	plev	1
air_pressure_anomaly	Pa		26

UDUNITS

The database for the UDUNITS-2 package comprises

- one XML file containing unit prefixes
  - SI unit prefixes
- four XML files containing unit definitions:
  - SI base units
  - SI derived units
  - Units accepted for use with the SI
  - Non-SI units



It is important to check that you are using valid units in your data set

ARCTIC PASSION INTER=ACT SIOS

Some recommendations/requirements

- COARDS requires CDL dimension order tzyx. CF recommends it.
  - Not following this can cause issues downstream...
- The Unidata NetCDF standard associates one coordinate variable with each dimension, by identity of name e.g. lat(lat).
- Time (year, month, day, hour, minute, second) is encoded with units “time unit since reference time”.
- The cell methods attribute of a data variable indicates how variation within the cells is represented. The method may be different for each axis.
  - point for intensive quantities e.g. temperature
  - sum for extensive quantities e.g. precipitation amount in time
- See https://cfconventions.org/Data/cf-documents/overview/viewgraphs.pdf for details

Attribute Convention for Dataset Discovery (ACDD)

- For details check
  - https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3
- This will help systems and users locate and use data efficiently.
- THREDDS and other tools can use these attributes to extract metadata from datasets, and exporting to Dublin Core, DIF, ADN, FGDC, ISO 19115 and other metadata formats.

- Mappings for ACDD
  - https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_Mappings
- List of elements that should be included as global attributes in the SIOS context
  - https://adc.met.no/node/4

File "osisaf_bar_sie_monthly.nc"

File type: Open-source Project for a Network Data Access Protocol

```
netcdf dods://thredds.met.no/thredds/dodsC/osisaf/met.no/ice/index/v2p1/bar/osisaf_bar_sie_monthl
  dimensions:
    time = 521;
    nv = 2;
    nchar = 3;
  variables:
    int time(time=521);
      :standard_name = "time";
      :long_name = "time of the observation (centered)";
      :coverage_content_type = "auxiliaryInformation";
      :units = "days since 1970-01-01 00:00:00";
      :axis = "T";
      :bounds = "time_bounds";

    int time_bounds(time=521, nv=2);

    float sic_threshold;
      :long_name = "threshold for the sea ice extent computation";
      :coverage_content_type = "auxiliaryInformation";
      :standard_name = "sea_ice_area_fraction";
      :units = "1";

    float lat;
      :bounds = "lat_bounds";
      :coverage_content_type = "auxiliaryInformation";
      :standard_name = "latitude";
      :units = "degrees_north";

    float lat_bounds(nv=2);

    float lon;
      :standard_name = "longitude";
      :units = "degrees_east";
      :bounds = "lon_bounds";
      :coverage_content_type = "auxiliaryInformation";

    float lon_bounds(nv=2);
```

```
        :coordinates = "lat lon sic_threshold";
        :units = "1e6 km^2";
        :coverage_content_type = "physicalMeasurement";
        :comment = "Lake ice is not included";
        :cell_method = "time: mean over months";

    // global attributes:
    :title = "Monthly Mean Barents Sea Sea Ice Extent from EUMETSAT OSI SAF (v2p1)";
    :product_id = "OSI-420";
    :product_name = "OSI SAF Sea Ice Index";
    :product_status = "demonstration";
    :version = "v2p1";
    :summary = "Time series of Monthly Mean Sea Ice Extent (SIE) for Barents Sea, computed from the
    :keywords = "Earth Science > Cryosphere > Sea Ice > Sea Ice Extent,Earth Science > Oceans > Sea
    :keywords_vocabulary = "GCMD Science Keywords";
    :geospatial_lat_min = 72.0; // double
    :geospatial_lat_max = 82.0; // double
    :geospatial_lon_min = 10.0; // double
    :geospatial_lon_max = 60.0; // double
    :time_coverage_start = "1979-01-16T00:00:00Z";
    :time_coverage_end = "2022-05-16T00:00:00Z";
    :Conventions = "CF-1.7, ACDD-1.3";
    :history = "2022-05-30T102432Z creation";
    :source = "EUMETSAT OSI SAF Sea Ice Concentration V2 CDR (OSI-450) and ICDR (OSI-430-b)";
    :processing_level = "Climate Indicators derived from Sea Ice Concentration maps";
    :date_created = "2022-05-30T102432Z";
    :creator_type = "institutions";
    :creator_institution = "Norwegian Meteorogical Institute";
    :creator_name = "Norwegian Meteorogical Institute";
    :creator_email = "info@met.no";
    :creator_url = "met.no";
    :institution = "Norwegian Meteorogical Institute";
    :publisher_name = "Norwegian Meteorogical Institute";
    :publisher_email = "info@met.no";
    :publisher_url = "met.no";
    :project = "EUMETSAT OSI SAF";
    :distribution_statement = "Free";
    :copyright_statement = "Copyright 2022 EUMETSAT";
    :references = "Product User Manual for OSI-420, Lavergne et al., v1.0, November 2020";
    :featureType = "timeSeries";
}
```

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

Additional slides

Attribute	Description
id	An identifier for the data set, provided by and unique within its naming authority. The combination of the "naming authority" and the "id" should be globally unique, but the id can be globally unique by itself also. IDs can be URLs, URNs, DOIs, meaningful text strings, a local key, or any other unique string of characters. The id should not include white space characters.
naming_authority	The organization that provides the initial id (see above) for the dataset. The naming authority should be uniquely specified by this attribute. We recommend using reverse-DNS naming for the naming authority; URIs are also acceptable. Example: 'edu.ucar.unidata'.
title	A short phrase or sentence describing the dataset. In many discovery systems, the title will be displayed in the results list from a search, and therefore should be human readable and reasonable to display in a list of such names. This attribute is also recommended by the NetCDF Users Guide and the CF conventions.
summary	A paragraph describing the dataset, analogous to an abstract for a paper.
keywords	A comma-separated list of key words and/or phrases. Keywords may be common words or phrases, terms from a controlled vocabulary (GCMD is required), or URIs for terms from a controlled vocabulary (see also "keywords_vocabulary" attribute). If keywords are extracted from e.g. GCMD Science Keywords, add keywords_vocabulary='GCMD'.
geospatial_lat_min	Describes a simple lower latitude limit; may be part of a 2- or 3-dimensional bounding region. Geospatial_lat_min specifies the southernmost latitude covered by the dataset. Must be decimal degrees north.
geospatial_lat_max	Describes a simple upper latitude limit; may be part of a 2- or 3-dimensional bounding region. Geospatial_lat_max specifies the northernmost latitude covered by the dataset. Must be decimal degrees north.
geospatial_lon_min	Describes a simple longitude limit; may be part of a 2- or 3-dimensional bounding region. geospatial_lon_min specifies the westernmost longitude covered by the dataset. See also geospatial_lon_max. Must be decimal degrees east.
geospatial_lon_max	Describes a simple longitude limit; may be part of a 2- or 3-dimensional bounding region. geospatial_lon_max specifies the easternmost longitude covered by the dataset. Cases where geospatial_lon_min is greater than geospatial_lon_max indicate the bounding box extends from geospatial_lon_max, through the longitude range discontinuity meridian (either the antimeridian for -180:180 values, or Prime Meridian for 0:360 values), to geospatial_lon_min; for example, geospatial_lon_min=170 and geospatial_lon_max=-175 incorporates 15 degrees of longitude (ranges 170 to 180 and -180 to -175). Must be decimal degrees east.

Attribute	Description
time_coverage_start	Describes the time of the first data point in the data set. Use the ISO 8601:2004 date format, preferably the extended format as recommended in the Attribute Content Guidance section. I.e. YYYY-MM-DDTHH:MM:SSZ (always use UTC).
time_coverage_end	Describes the time of the last data point in the data set. Use ISO 8601:2004 date format, preferably the extended format as recommended in the Attribute Content Guidance section. I.e. YYYY-MM-DDTHH:MM:SSZ (always use UTC).
Conventions	A comma-separated string of the conventions that are followed by the dataset. For files that follow this version of ACDD, include the string 'ACDD-1.3'. (This attribute is described in the NetCDF Users Guide.)
history	Provides an audit trail for modifications to the original data. This attribute is also in the NetCDF Users Guide: 'This is a character array with a line for each invocation of a program that has modified the dataset. Well-behaved generic netCDF applications should append a line containing: date, time of day, user name, program name and command arguments.' To include a more complete description you can append a reference to an ISO Lineage entity; see NOAA EDM ISO Lineage guidance.
source	The method of production of the original data. If it was model-generated, source should name the model and its version. If it is observational, source should characterize it. This attribute is defined in the CF Conventions. Examples: 'temperature from CTD #1234'; 'world model v.0.1'.
processing_level	A textual description of the processing (or quality control) level of the data.

Attribute	Description
date_created	The date on which this version of the data was created. (Modification of values implies a new version, hence this would be assigned the date of the most recent values modification.) Metadata changes are not considered when assigning the date_created. The ISO 8601:2004 extended date format is recommended, as described in the Attribute Content Guidance section. E.g. 2020-10-20T12:35:00Z.
creator_type	Specifies type of creator with one of the following: 'person', 'group', 'institution', or 'position'. If this attribute is not specified, the creator is assumed to be a person. If multiple persons are involved, please list these as a comma separated string. In such situation please remember to add a comma separated string for creator_institution and creator_email as well. Consistency between these fields are done from left to right.
creator_institution	The institution of the creator; should uniquely identify the creator's institution. This attribute's value should be specified even if it matches the value of publisher_institution, or if creator_type is institution. See last paragraph under creator_type.
creator_name	The name of the person (or other creator type specified by the creator_type attribute) principally responsible for creating this data. See last paragraph under creator_type.
creator_email	The email address of the person (or other creator type specified by the creator_type attribute) principally responsible for creating this data. See last paragraph under creator_type.
creator_url	The URL of the person (or other creator type specified by the creator_type attribute) principally responsible for creating this data. See last paragraph under creator_type.
institution	The name of the institution principally responsible for originating this data. This attribute is recommended by the CF convention. If provided as a string ending with a keyword in parantheses (), the main text will be interpreted as the long name and the keyword in the parantheses as the short name. E.g. 'Norwegian Meteorological Institute (MET)'

Attribute	Description
publisher_name	The name of the person (or other entity specified by the publisher_type attribute) responsible for publishing the data file or product to users, with its current metadata and format.
publisher_email	The email address of the person (or other entity specified by the publisher_type attribute) responsible for publishing the data file or product to users, with its current metadata and format.
publisher_url	The URL of the person (or other entity specified by the publisher_type attribute) responsible for publishing the data file or product to users, with its current metadata and format.
project	The name of the project(s) principally responsible for originating this data. Multiple projects can be separated by commas, as described under Attribute Content Guidelines. Examples: 'PATMOS-X', 'Extended Continental Shelf Project'. If each substring includes a keyword in parantheses, this is interpreted as the short name for the project while the rest is the long name. E.g. 'Nansen Legacy (NLEG)'.