

FAIR data principles, metadata and data standards, documentation and formatting tools

Lara Ferrighi, Norwegian Meteorological Institute

Ocean Data Dojo Workshop, 2022-11-01



The value of sharing data


Data sharing requires

- Effort
- Planning
- Resources

Always think about being a data provider, but also a data consumer




Goals

- 
- Support/expect transparency and open science
 - Support/expect integration and reuse of data
 - Support/expect visibility and recognition (both of the institute and the researchers)



Who benefits from it?

- 
- The research community
 - The researchers
 - The general public
 - The funding entities

The **FAIR** guiding principles

- To be **Findable**:

- **F1. (meta)data are assigned a globally unique and persistent identifier**

- F2. data (defined)

- F3. meta identifier

- **F4. (meta) a search**

- To be **Access**

- **A1. (meta) identifier**

- **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**

Remember:

- Don't take it as a MUST
- Something is better than nothing
- You need to start somewhere
- Do as much as you can
- Be aware

- **(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

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**

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. <u>It describes who did what, where and when, how to access data and potential constraints on the data.</u> 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 <u>the actual content of a dataset</u> 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 <u>to describe the context of observational data.</u> 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:

“Not ready yet”

“Done”

“Still acquiring data”

“Continuously updating”

“???”

Data provider B can use:

“Not finished”

“Finished and stored”

“unknown”

“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.

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

- | | | |
|---|---|---|
| <ul style="list-style-type: none">• 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) | <ul style="list-style-type: none">• 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) | <ul style="list-style-type: none">• 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 name has a canonical unit which is 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." [Section 1.3 of the CF conventions](#) states: "The values of the units attributes are character strings that are recognized by UNIDATA's Udonits package [UDONITS], (with exceptions allowed as discussed in [Section 1.3 of the CF conventions](#))." 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 Udonits. For the full details, refer to section 6 of the [Udonits documentation](#). Refer to the [CF conventions](#) for full details of the units attribute.

Search

☒ 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 environment, equipment or remote
4	Environment	Describes the geographic location. 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 multiple instruments are used to generate the observation, then this category should
6	Sampling	Specifies how sampling and/or analysis are used to generate the 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)

Example: Poor Data Entry

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010		1 DIPU	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10		2 Pero	13.22	j		DW		1 y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS		2 j	PERO	escaped <15		
5	rioSladu	"	1+	CleGap	18.92	gut away		RS		3 ri	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well		1 y	woodrat	13									
15	riosalado		2 y	PERO	24.5									
16	riosalado		3 y	Clegap	91									
17														
18														
19														
20														


Sheet1

Inconsistency between data collection events

- Location of **date** information
- Inconsistent **date** format
- **Column names**
- Order of columns

Format conversion: Rosetta tool

- Web-based application
- Can read metadata from header(s), and add user defined metadata
- Profiles, time series, trajectory
- Saves “setup” as templates for future use
- Open source tool, written in Java
- <http://tomcat.nersc.no/rosetta/>



Rosetta

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

Welcome to Rosetta, a data transformation tool. Rosetta is a web-based service that provides an easy, wizard-based interface for data collectors to transform their datalogger generated ASCII output into Climate and Forecast (CF) compliant netCDF files. These files will contain the metadata describing what data is contained in the file, the instruments used to collect the data, and other critical information that otherwise may be lost in one of many dreaded README files.




In addition, with the understanding that the observational community does appreciate the ease of use of ASCII files, methods for transforming the netCDF back into a user defined CSV or spreadsheet formats is planned to be incorporated into Rosetta.

We hope that Rosetta will be of value to the science community users who have needs for transforming the data they have collected or stored in non-standard formats.

Rosetta is currently under continued further development, and ready for beta testing.

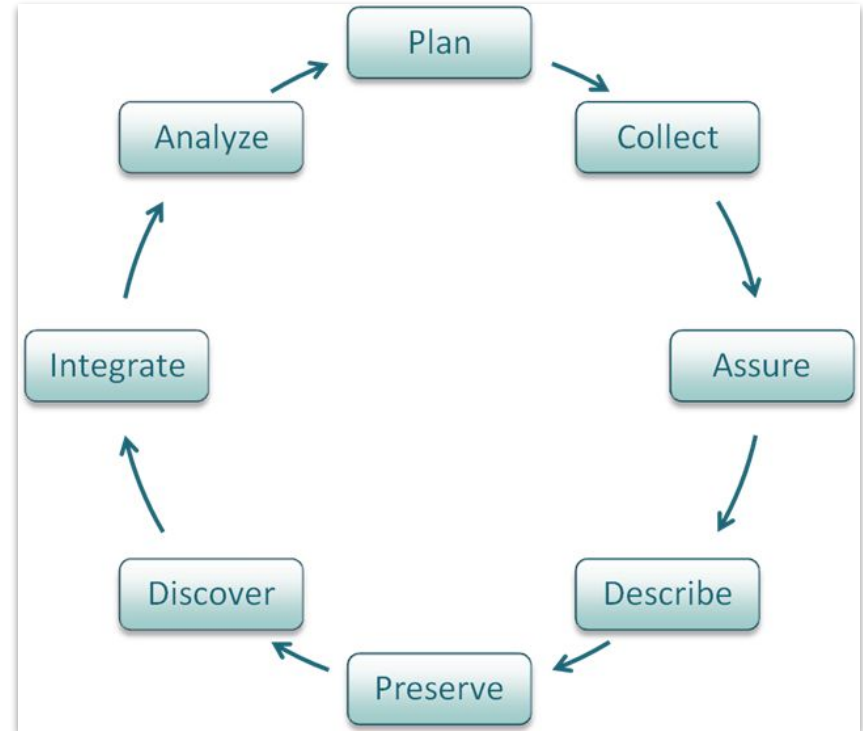
What would you like to do?

[Rosetta User Manual](#)



Data Life Cycle and its relation to the researchers workflow

- Plan
- Generate or collect data
- Quality control and document data
- Analyse data
- Prepare and publish data through a mandated archive
 - Remember to get a DOI!!
- Write a data paper
 - Cite your datasets using the provided DOI
- Write your scientific paper
 - Cite your data paper describing the context of the data



Data Citation

- **Data authors** get credit for the data publication and subsequent citations
- Can be used to show the impact of research projects
- Shows the scientific impact of **data center**

Goal: citations become a normal part of the scientific process

- Publishing data products is recognized as a part of scholarly research
- Authors cite the data products they use
- Journals require data to be published when related papers are published

Scientist overview:

- Develop appropriate citations for your data
- Assign identifiers to data
- Find citations for existing data
- Make citing data a regular practice

Data License

A license is an official permission or permit to do, use, or own something (as well as the document of that permission or permit). (Cambridge English Dictionary)

Make sure to use standards when you declare your license.

- Include the license in your metadata
- Whenever possible use machine readable licenses
- Pick from the creative commons license list (or spdx)
- Avoid using free text

Research data: “as open as possible, as close as necessary” to ensure transparency and openness.

- FAIR data does not mean Open Data
 - Embargo
 - Sensitive data



CC BY 4.0: This license lets others distribute, remix, adapt, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

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>
- <https://datacite.org/>
- <https://support.datacite.org/docs/doi-basics>
- <https://www.doi.org/faq.html>
- <https://commons.esipfed.org/node/726>
- <https://commons.esipfed.org/node/1428>
- https://dataoneorg.github.io/Education/lessons/08_citation/index.html
- <https://creativecommons.org/licenses/>

Example: Poor Data Entry

The screenshot shows an Excel spreadsheet with two data tables. The first table (rows 1-6) has columns: Site, Date, Plot, Species, Weight, and Acult. The second table (rows 12-16) has columns: Site, Plot, Adult, Species, grams, and Comments. Various data entry errors are highlighted with colored boxes: blue for site names, yellow for a mean value, and green for a weight column containing both text and numbers.

Site	Date	Plot	Species	Weight	Acult
DeepWell	2/13/2010		1 DIPO	12.1	j
Deep Well	Feb-10		2 Pero	13.22	j
rioSalado	2/13/2010	1a	pero	16	N
riuSladu	"	1*	CleGap	18.92	gul away
			Mean1	15.06	

Site	Plot	Adult	Species	grams	Comments
deep well	1	y	woodrat	13	
riosalado	2	y	PERO	24.5	
riosalado	3	y	Clegap	91	

Inconsistency between data collection events

- Different **site** spellings, capitalization, spaces in **site** names = hard to filter
- Codes used for site names for some data, but spelled out for others
- **Mean1** value is in Weight column
- **Text and numbers** in same column – what is the meaning of 12, “escaped < 15”, and 91?

Best practices

- Columns of data are consistent: only numbers, dates, or text
- Consistent names, codes, formats (date ISO8601) used in each column
- Data are all in one table, which is much easier for a statistical program to work with than multiple small tables which each require human intervention



	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

Best practices

- Identify missing values with dedicated code
- Indicate reason for missing value in code
- In numeric fields, use a distinct value such as 9999 to indicate a missing value
- In text fields, use NA (“Not Applicable” or “Not Available”)

Date	Time	NO3_N_Conc	NO3_N_Conc_Flag
20081011	1300	0.013	
20081011	1330	0.016	
20081011	1400		M1
20081011	1430	0.018	
20081011	1500	0.001	E1

M1 = missing; no sample collected

E1 = estimated from grab sample

Validation

datavalidation.xls

1	A	B	C	D	E	F	G	H	I
2	Date	Site	Plot	Species	Height				
3	1/12/2011	Deep Well	N	BOGR2	12.00				
4				BOGR2					
5				BOHI2					
6				BOIN					
7				BOPU					
8				BO5A					
9				BO5P					
10				BRAN					
11				BRBA2					
12									
13									
14									
15									

Sheet1 / Sheet2 / Sheet3

datavalidation.xls

Microsoft Excel

You have entered an illegal value.

Retry Cancel

C	D	E	F	G	H	I
t	Species	Height				
	BOGR2	20				
6						
7						
8						
9						
10						

Data Validation

Settings Input Message Error Alert

Validation criteria

Allow: Decimal ☒ Ignore blank

Data: between

Minimum: 11

Maximum: 15

☐ Apply these changes to all other cells with the same settings

Clear All OK Cancel

Data Validation

Settings Input Message Error Alert

Validation criteria

Allow: List ☒ Ignore blank ☒ In-cell dropdown

Any value
Whole number
Decimal
List
Date
Time
Text length
Custom

☐ Apply these changes to all other cells with the same settings

Clear All OK Cancel