NetCDF data at the 4TU.Centre for Research Data - a review of compliance with the FAIR principles

Maria Cruz and Egbert Gramsbergen, TU Delft, July 2018

(with thanks to Jasmin Böhmer for her input)

This review follows on a previous review of the FAIR principles in the context of 4TU.ResearchData written in 2017 by Alastair Dunning, the Head of the 4TU.Centre for Research Data (short version: 4TU.ResearchData). Here, we modify and update that review, and extend it to the context of netCDF[1] data deposited at the 4TU.ResearchData archive. As the previous review, our review focuses mainly on the metadata that describe each dataset rather than the data sitting within each dataset.

For the purposes of this review, we consider that the netCDF files are self-describing, i.e. with rich internal metadata describing the data they contain, and that they comply with the CF (Climate and Forecast) conventions[2]. This is not the case for all netCDF datasets currently deposited at 4TU.ResearchData. However, our intention here is to highlight the potential for full FAIR compliance for netCDF data deposited at 4TU.ResearchData that include extensive metadata, comprising both intrinsic and contextual metadata, and comply with domain-relevant standards, such as the CF conventions.

The FAIR data principles, published in 2016, are a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable. We use the FAIR Data Principles explained by the Dutch Techcentre for Life Sciences (DTL) as a guide to interpret the FAIR principles.

---

[1] As described in https://www.unidata.ucar.edu/software/netcdf/ [last accessed 19 July 2018], the Network Common Data Form, or 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 in the late 1980s and is maintained by Unidata. The format is mainly and widely used in oceanography, climate and atmospheric sciences, and although it is not an intrinsically geospatial format, netCDF is widely used for geospatial data.

[2] The CF conventions are community-developed standards for use with climate and forecast data, and atmosphere, surface and ocean model-generated data and comparable observational datasets. As described in http://cfconventions.org/[last accessed 19 July 2018], they define metadata that provide a definitive description of what the data in each variable on a netCDF file represents, and the spatial and temporal properties of the data.

To be Findable:

Principle	4TU.ResearchData Policy	netCDF data
(meta)data are assigned a globally unique and eternally persistent identifier.	Yes, a DOI is minted for each dataset.	Yes, the same applies to all netCDF datasets deposited at 4TU.ResearchData.
data are described with rich metadata.	Yes, each dataset is allocated multiple metadata fields according to international standards (RDF, Dublin Core).	Yes. Besides the standard RDF and Dublin Core metadata, netCDF files also include their own metadata fields (within the files). These metadata typically include descriptive information about the context, quality, and condition, or specific characteristics of the data.
(meta)data are registered or indexed in a searchable resource.	Yes, data are crawlable and metadata can be harvested through OAI-PMH (e.g., for NARCIS). We also push metadata to DataCite where it can be searched and harvested by third parties. There is also a SPARQL Endpoint, which is however hidden, because it also reveals non-public administrative data.	The additional metadata that are internal to netCDF files are fully machine readable and can be indexed and searchable through the OPeNDAP protocol.
metadata specify the data identifier.	Yes	This is not valid for the metadata within the netCDF files, but it applies to the standard metadata

		associated with netCDF datasets as part of being deposited at 4TU.ResearchData.

To be Accessible:

Principle	**4TU.ResearchData Policy**	**netCDF**
(meta)data are retrievable by their identifier using a standardized communications protocol.	Yes, http(s) is used.	Yes, netCDF files can also be accessed via the OPeNDAP protocol.
the protocol is open, free, and universally implementable.	Yes, http(s) is open etc.	Yes, OPeNDAP is open, etc.
the protocol allows for an authentication and authorization procedure, where necessary.	Yes, data contributors need to authenticate themselves. Data users can access the data without registration.	NetCDF files can be accessed without registration via http(s) or OPeNDAP.
metadata are accessible, even when the data are no longer available.	Yes, this is part of our policy.	This is not valid for the metadata that is included within the netCDF files, but it applies to the standard metadata associated with netCDF datasets as part of being deposited at 4TU.ResearchData.

To be Interoperable:

Principle	4TU.ResearchData Policy	netCDF data
(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Yes, metadata uses elements of Dublin Core, and can also be exposed as ORE RDF/XML.	Yes, netCDF provides a common model for concepts such as variables and dimensions, and the CF conventions add a vocabulary for many specific variables, units, and general metadata.
(meta)data use vocabularies that follow FAIR principles.	We use established vocabularies/ontologies where we can. For missing properties and classes, we developed our own ontology. All follow the relevant FAIR principles, although not all the documentation is necessarily resolvable using globally unique and persistent identifiers.	The CF conventions and vocabularies are well documented and the documentation is findable and accessible. However, the documentation is not necessarily resolvable using globally unique and persistent identifiers.
(meta)data include qualified references to other (meta)data.	Yes, because we use RDF, all references are qualified. There are many references to publications, ORCID records, and places on geonames and wikipedia.	Yes, the CF conventions allow for general links to publications as well as specialized identifiers, e.g. for observation stations.

To be Re-usable:

Principle	**4TU.ResearchData Policy**	**netCDF**
meta(data) have a plurality of accurate and relevant attributes	Yes, we employ many fields.	Yes, CF has a really long list of attributes.
(meta)data are released with a clear and accessible data usage license.	Yes, each dataset is published under a clearly specified licence. Data contributors can choose between different Creative Common and other licences, the default being CC0, or even use their own licence if they wish. Older datasets were published under our own bespoke licence.	The same applies to netCDF datasets.
(meta)data are associated with their provenance.	The source of data is included in the metadata records, but does not display the file processing and how the final data was created. After ingest, however, all changes are documented in an audit trail that is an integral part of the dataset itself. The audit trail is not public. Major events like the release of a new version of the data produce a new dataset with its own metadata and a link to the	Yes, the CF conventions allow for provenance metadata (where and how the data was produced, history of processing operations, etc.).

	previous version, which remains accessible.	
(meta)data meet domain-relevant community standards	Partially. Difficult to have subject specific metadata when we cover so many different subjects. However, some data formats are tailored for particular domains.	Yes. NetCDF is a well-established and sustainable format widely used in the climate, atmospheric and ocean sciences. The CF conventions are community standards for use with climate and forecast data; atmosphere, surface and ocean model-generated data and comparable observational datasets.