FAIR Data Guidelines and Measures - Mapping best practices for climate services data output to GO-FAIR metrics

Introduction to FAIR data

The impact of climate services depends on the trustworthiness of its output datasets including, for example, being able to trace back the full data flow, as well as processing software and its settings, by the end-user. Here, we focus on the data aspect. In order to provide optimised access to, and (re)use of, data it is important that the data are as FAIR as possible. FAIR data means that data is Findable, Accessible, Interoperable and Reusable for humans **and machines**, which can only be achieved if the data is readily available through machine accessible services, supported by complete machine interpretable metadata, and follows international metadata and data standards. This means that best practices for metadata and data will need to be adopted as early as possible in the data life cycle.

A well-established data and metadata management system at the source is key to providing access to FAIR data throughout the data life cycle, which implies not only for first use, often national/regional, but also later in re-use via international aggregators. Increased FAIRness can be achieved by developing new data management best practices for climate services data, combining the WMO metadata requirements, with more detailed practices on how to enable machine-to-machine accessibility regarding each of these aspects. Best practices include metadata formats, creation of new vocabularies (increases the I and R), references to services and software and their settings (increases the R, if referred to in the metadata).

Mapping best practices for climate services metadata to GO-FAIR metrics

In order to determine how optimised FAIRness could be achieved, we start by relating the GO-FAIR metrics to the current existing practices for climate data and metadata. A set of 14 principles to determine the level of FAIRness of data has been developed by GO-FAIR (https://www.go-fair.org/fair-principles/). Each of the principles contains a reference to recommendations of how they could be met in practice.

The path to FAIR data can be seen as a “movement”, a route that parties involved in a certain domain agree on how to achieve it. Especially for aspects such as Interoperability and Reusability, the path depends a lot on agreed solutions (e.g. vocabularies) in the domain. But of course, it helps to have a common target and means to validate how far a data provider is in the process. The FAIR principles themselves can be used as the first generic metrics to evaluate the FAIRness of data from the climate services.

To support increasing FAIRness of data in the Climateurope2 community, guidelines and measures can be formulated on top of existing best practices in the field of climate services. Based on a review of documentation from the World Meteorological Organization (WMO), the Global Framework for Climate Services (GFCS), the Open Geospatial Consortium (OGC), the Group on Earth Observations (GEO), and other actors in the metadata sphere for climate services, best practices for climate services metadata in connection with data have been summarised. Using the abovementioned GOFAIR metrics these existing best practices are now grouped per metric, as seen below, in order to find the matches, challenges and gaps. The metrics that could not be mapped to the best practices are listed as identified gaps.

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101056933.

Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

- File Naming Conventions: Use consistent and informative file naming conventions for data files. Include relevant timestamps, variables, and other metadata in the file names.
    - **The convention is a good starting point, but it is not a globally unique identifier (Digital Object Identifier - DOI) or Persistent Identifier (PID) as such, which might lead to inconsistencies.**
- Data Citation: Encourage data citation by providing a standardised citation format. This helps acknowledge the data source and promotes proper attribution.
    - **Especially for citation, a persistent identifier is much needed.**

F2. Data are described with rich metadata (defined by R1 below)

- Clear and Comprehensive Descriptions: Provide detailed information about the dataset. Include titles, abstracts, keywords, and descriptions that explain the purpose, content, and context of the data. This helps users understand what the data is about.
    - **Important to keep in mind the machine user, keywords should be annotated from common vocabularies as much as possible.**
- Temporal and Spatial Information: Clearly specify the temporal and spatial coverage of the data. This should include start and end times, as well as geographical coordinates, resolution, and projection information.
- Data Quality and Uncertainty: Include information on data quality, accuracy, precision, and any known uncertainties. This helps users assess the reliability of the data.
    - **Also important to keep in mind the machine user, information on quality, accuracy and precision should be annotated via common vocabularies as much as possible.**
- Units and Parameters: Clearly define the units of measurement for variables and parameters in the dataset. Include standard abbreviations and conversion factors where applicable.
    - **Using agreed common code lists/vocabularies as much as possible, e.g. Global Change Master Directory (GCMD), or Climate and Forecast (CF) conventions**
- Review and Quality Assurance: Implement a review and quality assurance process for metadata to ensure consistency and accuracy.

F4. (Meta)data are registered or indexed in a searchable resource

- Metadata Accessibility: Make metadata discoverable and accessible through metadata catalogues, data repositories, or data portals. Use standardised metadata search and discovery mechanisms.

Accessible

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

- Data Access and Distribution: Specify how users can access the data, including download links, APIs, or other methods.
    - **Important that the metadata contains the PID's of the datafiles it describes, and vice versa.**
- User Support: Provide contact information for users who have questions or need assistance with the data. Offer user support and FAQs when possible.

- Training and Documentation: Offer documentation, tutorials, and training materials to help users understand and use the data effectively.
    - **In case of data access API's document these for developers (e.g. OpenAPI standard/Swagger).**

A1.1. The protocol is open, free, and universally implementable

- Ensure that data are available in open and machine-readable formats.
    - **For FAIRness "free" does not mean that the data should be free of charge and open, but rather the protocol to access it should be open and free and available to all. This means that anyone with a computer and an internet connection can access at least the metadata and clearly know how to access the data. It could then also be possible to include restricted data while still remaining 'FAIR'.**

Interoperable

I2. (Meta)data use vocabularies that follow FAIR principles

- Interoperability: Ensure that metadata conforms to international standards and can be easily integrated with other climate data sources and tools.
    - **This FAIR metric goes one step further than the syntax (metadata format), but also puts emphasis on agreed international vocabularies to support the semantics in the metadata.**

I3. (Meta)data include qualified references to other (meta)data

- Metadata Cross-Referencing: Link related datasets, publications, and documentation. This helps users find additional information and resources.
    - **This works best if persistent unique identifiers are used for each metadata resource.**

Reusable

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes

- Repeated from F2:
    - Clear and Comprehensive Descriptions: Provide detailed information about the dataset. Include titles, abstracts, keywords, and descriptions that explain the purpose, content, and context of the data, e.g. the experimental protocols. This helps users understand what the data is about.
        - **Important to keep in mind the machine user, keywords should be annotated from common vocabularies as much as possible.**
    - Temporal and Spatial Information: Clearly specify the temporal and spatial coverage of the data. This should include start and end times, as well as geographical coordinates, resolution, and projection information.
    - Data Quality and Uncertainty: Include information on data quality, accuracy, precision, and any known uncertainties. This helps users assess the reliability of the data.
        - **Important to keep in mind the machine user, quality, accuracy, precision and uncertainties, should be described via agreed common vocabularies as much as possible.**

- Plus:
    - Version Control: Maintain version control for datasets. Document changes, updates, and revisions to ensure users are aware of the dataset's history.
    - Metadata Updates: Ensure that metadata is regularly updated to reflect changes in the dataset, project, or data source.
        - **CE2 recommendation on this point is to be stricter: When updating a dataset (reprocessing, or e.g. changing the collection of data), it is recommended to update the related DOI as well to make sure that end-users are not citing the wrong version. If there are only changes made to the metadata, for the sake of simplicity the DOI can be maintained.**

R1.1. (Meta)data are released with a clear and accessible data usage licence

- License and Usage Terms: Clearly state the terms of use, licensing, and any restrictions on data usage. Make it explicit if the data are open access or require permissions.
    - **Important to keep in mind the machine user, licensing and usage (and e.g. authentication systems) should be specified via common vocabularies as much as possible.**

R1.2. (Meta)data are associated with detailed provenance

- See also the elements under R1.
- Data Provenance: Document the data source, including the instruments used, data collection methods, and organisations responsible for data collection. This helps establish data credibility and traceability.
    - **Important to keep in mind the machine user, instruments, methods (including, where applicable, how the data were combined to achieve a data product or result), organisations, persons, all should be annotated from common vocabularies or agreed directories (e.g. ORCID) as much as possible. Here it is also important to include machine accessible/interpretable quality assessment information, such as the accuracy and/or resolution of the measurement devices and datasets in order to assess the Reusability.**

R1.3. (Meta)data meet domain-relevant community standards

- Standardised Metadata Formats: Use established metadata standards such as WMO Core Metadata Profile (WCMP) 2019 for weather and climate data, or ISO 19115-1 or DataCite Metadata Schema 4.4 (2021) for datasets more generally. These standards provide a structured way to document information.
    - **Community standards listed are a good starting point, especially when including community vocabularies. Metadata and data formats could be reviewed for machine actionability, and extended where needed (see gaps/challenges).**
- Engage Stakeholders: Involve stakeholders, including the scientific community and end-users, in the development of metadata standards and requirements to ensure they meet user needs.
    - **The Climateurope2 community could play an important role here.**

Identified Gaps

Several of the FAIR metrics could not be mapped (yet) to WMO requirements and as such, associated recommendations have been identified:

- FAIRness should apply to the whole life cycle of the data and software, and that preservation of the metadata (even if the data and software are not preserved) is essential;
- the link between data and metadata must be always preserved and this is not always the case when having the metadata stored separately from the files containing the data;
- readability of the metadata and FAIRness must be ensured both for humans and machines.

They are listed in more detail below.

F3. Metadata clearly and explicitly include the identifier of the data they describe

In the case that the dataset and the accompanying metadata are separate files, the association between the two should be made explicit by mentioning the dataset's globally unique and persistent identifier in the metadata.

A1.2. The protocol allows for an authentication and authorisation procedure, where necessary

The 'A' in FAIR does not necessarily mean 'open' or 'free', but rather that the exact conditions should be clear under which the data is accessible, meaning that even restricted data can be FAIR. Preferably, the accessibility is specified such that a machine can understand the requirements, and then either automatically execute the requirements or alert the user to the requirements. By having repository users create an account, the owner (or contributor) of each dataset can be authenticated and user-specific rights can be set-up.

A2. Metadata are accessible, even when the data are no longer available

Principle A2 states that metadata should persist even when the data are no longer sustained. Datasets tend to degrade or disappear over time because there is a cost to maintaining an online presence for data resources. When this happens, links become invalid and users waste time hunting for data (e.g. which is used in a climate service product) that might no longer be there. Storing the metadata generally is much easier and cheaper. Even if the original data are missing, tracking down people, institutions or publications associated with the original research can be extremely useful.

I1. (Meta)data uses a formal, accessible, shared, and broadly applicable language for knowledge representation

Humans and **machines** should be able to exchange and interpret each other's data. The main goal of this principle is to provide a "common understanding" of digital objects by means of a language for knowledge representation to be used to represent these objects. The chosen language (RDF, OWL, JSON LD, etc.) should have a formal specification and the specifications should be shared and accessible so others can learn the language.

Outlook on next steps for FAIR Guidelines and Measures

In order to ensure the right level of trust in climate services output datasets it is important that the data is published in accordance with FAIR principles. This means that not only the output itself is FAIR, but also that the underlying dataflow is FAIR. As we have seen in previous sections this puts additional emphasis on the data management of climate data. The WMO guidelines and existing metadata and

data standards are an excellent starting point, but when it comes to FAIRness the guidelines could be “tightened”, and we can conclude additional guidance in the form of Best Practices are needed.

Some first directions that need further attention when it comes to Best Practices for climate service data management:

- FAIR is focused on machine-2-machine interaction to the data and metadata. This puts additional focus on data management, e.g. using as little as possible free text in metadata and data, but instead using vocabularies for these attributes. Not only for parameters, units, but also for data access restrictions, usage policy, quality information, organisations, persons, etc.
- To achieve optimum interoperability the vocabularies used should be community supported and FAIR by itself in order to be widely used and accessible by machines (e.g. to make mappings, request ontology).
- The best practices for File Naming Conventions mentions the use of consistent and informative file naming conventions for data files, including the relevant timestamps, variables and other metadata in the file names. Looking at the GO-FAIR metrics, this relates and differs to what Metric F1 states, which is that (Meta)data should be assigned a globally unique and persistent identifier.
- Processing services involved should be clearly indicated in the metadata with a link to the service and version, as well as to documentation (OpenAPI or software) where possible.