Meteorol. Z. (Contrib. Atm. Sci.), Vol. **29**, No. 6, 483–491 (published online August 24, 2020)
© 2020 The authors

Numerical Modelling

A short guide to increase FAIRness of atmospheric model data

Anette Ganske[1]*, Daniel Heydebreck[2], Heinke Höck[2], Angelina Kraft[1], Johannes Quaas[3] and Amandine Kaiser[2]

[1]Technische Informationsbibliothek (TIB), Hanover, Germany
[2]German Climate Computing Center, Hamburg, Germany
[3]Institute for Meteorology of the University of Leipzig, Leipzig, Germany

(Manuscript received May 26, 2020; in revised form June 24, 2020; accepted June 24, 2020)

Abstract

The generation, processing and analysis of atmospheric model data are expensive, as atmospheric model runs are often computationally intensive and the costs of 'fast' disk space are rising. Moreover, atmospheric models are mostly developed by groups of scientists over many years and therefore only few appropriate models exist for specific analyses, e.g. for urban climate. Hence, atmospheric model data should be made available for reuse by scientists, the public sector, companies and other stakeholders. Thereby, this leads to an increasing need for swift, user-friendly adaptation of standards.The FAIR data principles (Findable, Accessible, Interoperable, Reusable) were established to foster the reuse of data. Research data become findable and accessible if they are published in public repositories with general metadata and Persistent Identifiers (PIDs), e.g. DataCite DOIs. The use of PIDs should ensure that describing metadata is persistently available. Nevertheless, PIDs and basic metadata do not guarantee that the data are indeed interoperable and reusable without project-specific knowledge. Additionally, the lack of standardised machine-readable metadata reduces the FAIRness of data. Unfortunately, there are no common standards for non-climate models, e.g. for mesoscale models, available. This paper proposes a concept to improve the FAIRness of archived atmospheric model data. This concept was developed within the AtMoDat project (Atmospheric Model Data). The approach consists of several aspects, each of which is easy to implement: requirements for rich metadata with controlled vocabulary, the landing pages, file formats (netCDF) and the structure within the files. The landing pages are a core element of this concept as they should be human- and machine readable, hold discipline-specific metadata and present metadata on simulation and variable level. This guide is meant to help data producers and curators to prepare data for publication. Furthermore, this guide provides information for the choice of keywords, which supports data reusers in their search for data with search engines.

Keywords: AtMoDat, FAIR, DOI, Metadata, Controlled Vocabulary

1 Introduction

Scientific outcome of publicly financed projects should be published in a way, that the results are reusable and comprehensible. This includes the data on which the results are based unless the publication of the data conflicts with the right of personality, current legislation or similar restrictions. Other exceptions for publication are data of parameter testing studies or similar works that might mislead reusers. In many atmospheric research projects the produced amount of data is so huge, that the whole outcome cannot be analysed in detail by the data producers within the funding period.

Given these large amounts of raw data that are produced as output by atmospheric models, two options emerge:

1. Projects may store exact model codes, run- and post-processing scripts, rather than the data: since CPU time for re-running models is cheaper than storing data on accessible devices for long time, it may be recommendable to store the programme code and scripts and re-run the model in case a reproduction is required.
2. Publish the output data following a flexible and easy-to-use standardisation, in order to foster the reuse of data by other scientists and stakeholders.

When using option 1 the data would only be (re-)usable if the computer environment remained unchanged (no change of compilers etc.) and if the user has a licence for the source code to rerun the model. When using option 2 a long-term storage of the versioned model code should also be included to ensure reproducibility. It has been more common during the last years to publish the output data. Therefore, this study addresses the second approach.

Researchers are encouraged to open their data to other research groups, see e.g. the guidelines for the

*Corresponding author: Anette Ganske, Technische Informationsbibliothek (TIB), Welfengarten 1B, 30167 Hannover, Germany, e-mail: anette.ganske@tib.eu

DOI 10.1127/metz/2020/1042

Gebrüder Borntraeger Science Publishers, Stuttgart, www.borntraeger-cramer.com

EU programme Horizon 2020 (European Commission, 2016). However, even if the data are stored together with their mandatory metadata in a repository, it is often not reusable, e.g. if

- detailed discipline-specific metadata are missing,
- no information is provided, whether and how data and metadata are checked for accuracy and completeness,
- machine-readability and associated software information is not given,
- data are saved in proprietary and undocumented file formats,
- file formats are depending on the version of the writing program, and/or
- the rights for the reuse of the data are not specified.

Hence, the data should be published in a way, such that they are findable, accessible, interoperable, and reusable – following the FAIR data principles, see e.g. Wilkinson et al. (2016). Even though the FAIR data principles are often meant for increasing machine readability and interoperability, we interpret them to be guidelines for human-readability and usability of data as well.

The first principle of FAIR is that "(meta)data are assigned a globally unique and persistent identifier" (Wilkinson et al., 2016). Atmospheric model data are often published with a DataCite Digital Object Identifier (DOI, DataCite Metadata Working Group, 2019). Within this study we assume that a DataCite DOI is assigned to the data. Therefore, we only use the DataCite Metadata Schema in the following. Nevertheless, the principles can also be used for the metadata of other Persistent Identifiers (PIDs). However, a PID is necessary for FAIRness, but not sufficient (Mons et al., 2017).

Publishing FAIR data is a combined effort of scientists and repositories which store the data and make it accessible. Even though the FAIR data principles are well known and author guidelines (Enabling FAIR Data Community et al., 2018) and different assessment tools exist (Bahim et al., 2019), it is often not obvious, how they can be applied in practice. Several projects and initiatives exist which develop practical FAIRness rules for different disciplines (FAIRplus[1]; AtMoDat[2]), guidelines for FAIRness metrics (FAIR Data Maturity Model Working Group, 2020), and FAIRness assessment metrics (FAIRsFAIR[3]; FAIRmetrics[4]).

This publication presents instructions to FAIRly publish results of atmospheric models. These instructions were developed in the project AtMoDat (**At**mospheric **Mo**del **Dat**a). We used the experience of the World Data Centre for Climate (WDCC) with the curation of climate model results and transferred the methods to other atmospheric model data, e.g. of global/regional cloud models or microscale models.

In this paper, we describe the results of our investigations into the improvement of the reusability of atmospheric model data, which should enable FAIRer datasets. Even if the datasets are published by a repository, not only data curators do need to know these methods. This text is also interesting for data producers, as they have to deliver all necessary information about their data to the data curators. In addition, data reusers can learn from this text, which information about data can be searched for with search engines.

This paper is organised as follows: In section two we present definitions and methods. In section three we recommend how to provide rich and machine readable metadata, how the human- and machine readable parts of the landing page should be constructed, and which file formats and standards shall be used, followed by a short summary and discussion.

2 Materials and methods

2.1 Definitions

Within this paper we will use the following definitions:

Data file: a digital file which stores data to be used by a computer application or system. The digital results of a simulation with an atmospheric model are stored in data files[5].

Metadata: contains descriptive, contextual and provenance assertions about the properties of data[6].

For simplicity we additionally define the following terms:

Dataset: contains *both* the metadata *and* the data themselves.

Dataset collection: a collection of several datasets, which also might contain other dataset collections. If a dataset collection e.g. consists of the results of an Earth System Model and the corresponding metadata, then one dataset collection might contain all datasets with the results for the atmosphere, another one all datasets with oceanic data.

Landing page: a web page to which a resource identifier resolves.

Maturity: describes the degree of the formalisation and standardisation of a dataset with respect to FAIRness, completeness and accuracy of the (meta-)data. Both data and metadata mature as they pass through the different data post-production steps, which are performed by the repository. The higher the maturity, the easier it is to reuse the data. How the maturity of a dataset is specified in detail depends on the respective repository or might be defined by a community.

[1] https://fairplus-project.eu/
[2] https://www.AtMoDat.de
[3] https://fairsfair.eu/
[4] https://www.fairmetrics.org
[5] modified definition of https://en.wikipedia.org/wiki/Data_file
[6] according to the definition in https://www.rd-alliance.org/sites/default/files/DFT20Core20Terms-and20model-v1-6.pdf

2.2 Metadata

Metadata are attached to data, publications, files and other things for different purposes, e.g. bibliographic citation or administrative tasks. Metadata of data are describing characteristic aspects of the particular data. This might be information needed for citation (author(s), title, publication year) but also detailed information on how the data was generated (model version, compiler options, creation date).

A metadata schema is a collection of mandatory, recommended and optional metadata fields. Different metadata schemas exist, such as schema.org (Schema.org Steering Group, 2019), Metadata Terms of the Dublin Core Metadata Initiative (DCMI Usage Board, 2020) or Data Catalog Vocabulary (DCAT, Albertoni et al., 2019). In addition to these general purpose metadata schemata exist some domain specific ones, such as the Climate and Forecast Metadata Conventions (CF Conventions, Eaton et al., 2019).

Metadata fields storing the same information might be named differently between different schemata. Or, one field might exist in one schema but not in the other one. Hence, it is necessary to create a mapping between different metadata schemata if one holds metadata in one format but wants to import/export metadata in another format. It is reasonable to use standardised public mappings between the two schemata or make the own mapping publicly available. Jacobsen et al. (2020) also recommend the use of a knowledge representation language such as RDF[7] or OWL[8], so that the metadata can also be analysed by machines.

In most cases, one aggregation of metadata is attached to the PID of a dataset, another one to the landing page and a third aggregation is part of the dataset itself. Even if these three aggregations coincide in large parts, they are all needed. Their purposes will be explained in the next chapters. In any case, the process of publishing data can be facilitated, if a repository extracts as many information as possible from the metadata in the datasets. Nevertheless, additional metadata will always be needed for the DOI metadata.

The DataCite Metadata Schema 4.3 (DataCite Metadata Working Group, 2019), which is in the focus of this publication, is associated with DataCite DOIs and is meant for general purpose. It evolved from Dublin Core, but changed by the time due to requests from DataCite community members. The DataCite schema has 19 metadata properties. A property consists of one top level metadata field, e.g. *creator*, and, possibly, of additional subordinate metadata fields (subproperties) in a hierarchical tree structure, e.g. *creatorName*, *nameType*, *affiliation*, ... DataCite metadata can be stored as XML and/or JSON files. An example for a JSON metadata file can be found in the Supplement.

The DataCite metadata schema is quite extensive compared to the schemata of other DOI registration agencies. Other persistent identifiers, e.g. handles, are not associated to a metadata schema. Vice versa, only a few metadata schemata are associated with PIDs.

2.3 File format

The two most commonly used file formats in the atmospheric modelling community are netCDF and GRIB.

The *Network Common Data Format* (netCDF) is a self-describing and openly documented file format. The current version of the netCDF library (netCDF4) allows compression of data. NetCDF files are widely used in the whole Earth system modelling community, for the publication of the model results of phases 5 and 6 of CMIP (Taylor et al., 2012; Eyring et al., 2016 and Juckes et al., 2020), and for other applications as described in Signell et al. (2008).

The *General Regularly-distributed Information in Binary form* (GRIB) format is standardised by the World Meteorological Organization (WMO) and allows higher compression rates than netCDF. GRIB by default is not interoperable without extra attention, as the external GRIB tables do not deliver explained standard names but require the knowledge of the model to understand what is meant by the given 'long name'.

2.4 The granularity of archived data

Results of atmospheric simulations are usually available in various levels of granularity (= levels of detail). As the atmosphere is a complex system with many degrees of freedom, the output of an atmosphere model simulation consists of many variables. It includes 1-dimensional (1D) properties and 2D and 3D fields, also time-dependent and time-constant variables. Variables can also be stored in several temporal aggregations, e.g. hourly, monthly or annual values. Therefore, the model output usually consists of several large files. Specific diagnostics may additionally store statistics with further dimensions at each grid-box and time step.

Often, various temporal aggregations, such as annual and monthly averages, are written into separate files. During post-processing, data may be redistributed into a different file structure. For example it was specified within the Coupled Model Intercomparison Project (CMIP, Taylor et al., 2012 and Eyring et al., 2016), that all model results are stored in a way, such that each netCDF file contains only one single variable. Alternatively, multiple variables can be stored in a joint netCDF file if no conformance with the CMIP standard is aimed at.

In any case, the results of a simulation can be represented in different levels of detail, denoted as granularity. Defining the levels of granularity determines a hierarchical structure of the data, e.g. for their archiving. The lowest level of granularity could be the simulation itself, or only parts of a simulation (e.g. only the atmospheric results of an earth system model), or even

[7] https://www.w3.org/RDF/
[8] https://www.w3.org/TR/owl2-overview/

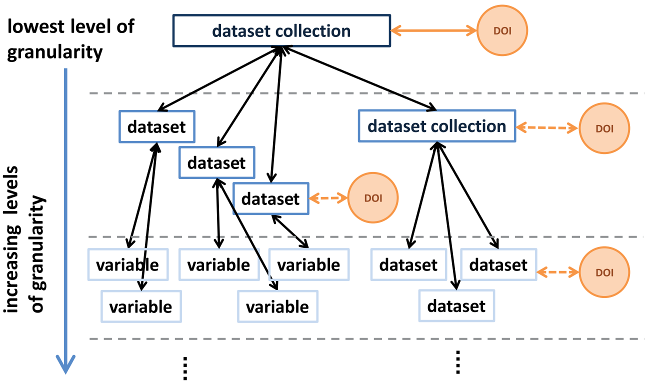

Figure 1: Example for a possible structure of a dataset collection, which contains the results of an atmospheric model simulation and the corresponding metadata. If this is defined as the lowest level of granularity and if a DOI is assigned to this dataset collection, then the orange arrow represents the link from and to the DOI. The black arrows represent possible references, which can be addressed by links on the web pages leading from the description of the dataset collection to the individual datasets and from the descriptions of the datasets to those of the individual variables. Nevertheless, additional DOIs could also be assigned to the higher levels of granularity (dashed arrows), e.g. to individual datasets.

a whole project in which several simulations were performed. As there are several possibilities to define a hierarchical structure, it should first be defined which group of data forms the lowest level of granularity. Then, the higher levels of granularity are determined.

Several datasets with results in different dimensions and temporal resolutions and the corresponding metadata are e.g. grouped into one dataset collection, see Figure 1. In this case the whole dataset collection could represent the low granular level (upper part of Figure 1). The next higher level of granularity could be the individual datasets (as denoted in the middle part of Figure 1). Each dataset could contain several variables, which then belong to the next higher level of granularity (lower part of Figure 1).

2.5 DOI assignment

The repository applies for the assignment of a DOI to a dataset. Often, a DOI is assigned to a lower granularity level of the data, e.g. to a dataset collection representing one model experiment whose output consists of several datasets as described in Figure 1 (upper part). Additional DOIs might be assigned to the elements of higher granularity, e.g. the individual datasets of a simulation (lower part of Figure 1). Thus, one could have one parent DOI for the whole simulation and one or more child DOIs for individual datasets. In any case, the metadata of the DOI describe that level of granularity, to which the DOI is attached. In most cases, only one DOI is assigned to the lowest granularity level and then only a fraction of the metadata of the DOI coincides with the metadata of the individual datasets.

2.6 Landing page

In many cases, a DOI is connected to a HTML landing page, which is created by the repository. Resolving a DOI in a conventional web browser will redirect to this respective landing page. This does not only apply to DataCite DOIs but to most DOIs[9]. The landing page contains metadata, which might be both human- as well as machine-readable. However, the DOI concept does not include any mandatory human-readable nor machine-readable formats. For every DataCite DOI a landing page is provided, that is as well human- and machine readable and searchable, e.g. with the DataCite Search[10].

3 Recommendations

The following recommendations aim to improve the FAIRness of atmospheric model datasets and follow the precondition, that the dataset is provided with a DataCite DOI, as this is the case within the AtMoDat project. It should be noted that other PID systems and metadata schema can also be used to improve dataset FAIRness; this is discussed in Section 4.

3.1 The importance of metadata for FAIRness

Metadata should describe the research data in such a way that the possible reusers are able to decide whether these data are useful for their application. Also, the process of the data generation and used input data should be documented. All necessary information must be given by the data producer. It can either be written directly into the metadata properties of the DataCite DOI or can be given by links to external documents – preferably via persistent identifiers (PIDs). Such an external document could be the documentation of the numerical model, which was used to calculate the data.

It is strongly recommended that metadata are machine-readable (for example using XML files and provision of metadata according to schema.org), such that they can be used to create automated lists, e.g. for evaluations of institutions. External documents should also be machine-readable.

3.1.1 FAIR principles for the Metadata itself

Machine-readability and therefore FAIRness of the metadata can be increased with the following principles:

- All information about persons and institutions should be complemented with a PID: e.g. an ORCID[11] for persons or a ROR[12] for institutions (if applicable).

[9] DOI Handbook, Chapter 5: https://www.doi.org/doi_handbook/5_Applications.html

[10] https://search.datacite.org/

[11] https://orcid.org/

[12] https://ror.org/

- All links to documents, homepages etc. should be provided as PIDs: documents e.g. with a DOI and homepages e.g. with an URN.
- All temporal information as e.g. dates, should be given in a standardised way, e.g. according to ISO 8601 (ISO8601, 2019) and ISO 19108 (ISO19108, 2002).
- Keywords and subjects should preferably be taken from controlled vocabularies (CVs), as much as possible, e.g in atmospheric modelling from CVs of ES-DOC[13], variable names from CF-conventions[14] or other keywords e.g. from the United Nations Terminology Database [15], the Global Change Master Directory (GCMD)[16], or the Climate Tagger Thesaurus[17].
- Geographic information should include the *spatial reference system* (also named *coordinate reference system*), e.g. WGS84 (NIMA, 2000). If no reference system is provided, WGS84 is assumed. If geographic coordinates are based on a geographic projection, then it is strongly recommended to include the projection, e.g. Lambert Conformal Conic, including relevant parameters. Geographic names should be chosen according to geonames[18], if possible.
- Usage rights/licence must always be described, preferably as a standard machine-readable licence, e.g. Creative Commons[19].

3.1.2 Metadata of the lowest level of granularity for atmospheric model data

The metadata of the lowest level of granularity is attached to the DataCite DOI, see e.g. the example in Figure 1. As we consider only DataCite DOIs in this paper, we restrict our recommendations to the properties of DataCite Metadata Schema 4.3 (DataCite Metadata Working Group, 2019). In order to increase the FAIRness of atmospheric model data, **all** reasonable DataCite metadata properties should be used. Datasets metadata that do have no corresponding DataCite metadata properties should be placed in general DataCite metadata properties, e.g. in *Description*.

The DataCite metadata properties *Contributor*, *Creator* and *Funder* should always be included, so that synopses about the publication of a single researcher, all researchers in an institution or all publications within a project can be automatically compiled.

All dates connected to the dataset are important information for the reuser and therefore should always be mentioned via the metadata properties *Date*. Except from the Publication Year all other temporal information is added with different values for the subproperty *dateType*. The length of time series or the period, for which a simulation is valid, is noted with the DataCite Metadata property *Date* and the subproperty *valid*.

The description of the geographical location of the model area (*GeoLocation*) is very important for the reusability of atmospheric model data, as well as the temporal coverage.

Also, there are many possibilities to mention related sources by using *RelatedIdentifier* and values for the subproperty *relationType*. Thereby the documentation of the model, the boundary conditions, or a publication about the data are noted. In order to increase the interoperability, all possible references to other data should also be listed; e.g. if data was calculated in an model intercomparison project (MIP), the data of other models included in the MIP should be mentioned.

As future reuse of a dataset cannot be predicted, as much information about the data as possible should been written into the metadata and therefore all applicable DataCite properties should be filled. Unfortunately, many archived datasets do not have sufficient information to be reusable, even if they have DataCite DOIs associated. Moreover, it is helpful for a reuser if information about performed maturity checks are added to the metadata. This can be done with a documentation of the maturity control that was made by the repository. The PID of the documentation of this maturity control can be added to the metadata with the property *RelatedIdentifier* and the subproperty *relationType="IsReviewedBy"*.

An example for the DataCite DOI's metadata of an atmospheric dataset of (Neumann, 2017), written in JSON, can be found in the Supplement (example _json_metadata.json).

3.1.3 Metadata of the higher levels of granularity of atmospheric model data

DataCite metadata properties are only describing the lowest level of granularity. Therefore, metadata of the higher level of granularity contain all necessary information about the data that couldn't be given with the DOI's metadata. If a DOI is assigned to a dataset collection with several datasets (as shown in Figure 1) and there are no additional DOIs for the individual datasets, then the descriptions of the datasets and variables (higher levels of granularity) can only be written on the landing page and into the metadata of the data files itself, see the tables in the Appendix. For atmospheric model data, the properties in Table 3 in the Appendix are necessary, if applicable.

Nevertheless, if there are additional DOIs assigned for each individual dataset, this information has to be included in the respective DOI's metadata.

[13] https://specializations.es-doc.org/

[14] http://cfconventions.org/Data/cf-standard-names/72/build/cf-standard-name-table.html

[15] https://unterm.un.org/

[16] https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords

[17] https://www.climatetagger.net/climate-thesaurus/

[18] https://www.geonames.org/

[19] https://creativecommons.org/choose/

3.2 The importance of the landing page

Beside the data repository as a whole, the landing page is the first "point of contact" between most of the reusers and their dataset of interest. As such, certain requirements for the design of a landing page have to be met, if the datasets provided with the DOI should comply with the FAIR principles. The landing page always both contains the human-readable and machine-readable metadata. According to *Best Practices for DOI Landing Pages of DataCite*[20], the landing page should always include a complete citation of the dataset in human-readable format including the DOI itself, so that the data record can be uniquely identified by humans. Also, the DOI should be stored in the machine-readable part of the landing page so that search engines also can find it. In addition, a landing page should always contain information on how to access the data. If the data record itself is no longer available, this should be noted on the landing page (Tombstone Page).

All metadata fields that are in the metadata record of the DataCite DOI should be listed on the landing page. Nevertheless, metadata fields on the landing page might have different names than the DataCite metadata properties. This e.g. enables to name maturity assurance information that is stored in the DataCite metadata schema under *Related Identifier:IsDerivedFrom*, directly as "Maturity Information". Also, it is strongly recommended that the landing page contains additional and needed information, for which no fields exist in the metadata schema of the DOI, see Tables 2 and 3 in the Appendix.

If the DOI was assigned to a dataset collection with several datasets or to datasets with several variables, the descriptions of the individual dataset or variables (higher levels of granularity) must also be written on the landing page or sub-landing pages. The top level of the landing page shall, however, always expose the metadata of the low level granularity and additionally a list of all available files/variables. The metadata of each file/variable (higher levels of granularity) can be provided on further web pages (sub-landing pages) linked to the landing page. On these web pages, individual metadata fields of the lower level of granularity, such as *Licence, Contributor*, . . . , can be repeated.

Both the landing page and the subordinate web pages have to be publicly and permanently available. The layout might change and metadata might be added. The landing page itself should have a PID which is in our case the DataCite DOI. The machine-readable part of the landing page should be provided compliant to schema.org[21] in order to enable search engines such as Google Dataset Search or Bing to extract relevant information for their search algorithms from the individual web pages. Alternatively, an equivalent structure based on the W3C DCAT format (Data Catalog Vocabulary, Albertoni et al. (2019)) might be used, see e.g. the data description from Google[22]. A mapping of the DataCite DOI metadata schema to schema.org is provided by DataCite[23].

3.3 File formats and standards for interoperability and reusability

A condition for the interoperability of research data is that they are stored in a self-describing open file format and that the structure of the file adheres to a convention that can be understood by humans and machines. Therefore, it is necessary that in addition to the data also metadata are stored in the file e.g. names of variables, units, licence information or contact details.

For atmospheric model data, the netCDF[24] format seems to be appropriate for this purpose because netCDF files are self-describing. However, the user is free to decide, in which detail and with which wording the variables, dimensions and attributes are described. This restricts reusability and machine readability significantly. Therefore, there are several standards that provide specifications for names, units and other parameters in order to standardise the file contents and make them automatically processible. The Climate and Forecast Conventions (CF Conventions, Eaton et al., 2019) are one of the most widely used standards in the atmosphere modelling community, which are also used for CMIP6 data (Juckes et al., 2020). This metadata standard gives specifications for descriptions of variables and coordinates in netCDF files. The CF Conventions have been developed for atmospheric model data, but they are becoming more and more popular in other fields of Earth System science and have been extended to some of these fields, e.g. atmospheric and oceanic measurement data, ocean model data, satellite data.

An example for a netCDF header for the same dataset as in section 3.1.2 by Neumann et al. (2017) is shown in the Supplement (example_netcdf_header.cdl).

4 Conclusion and discussion

The aim of the paper is to provide a concept for the enhancement of reusability of archived atmospheric model data, which can be achieved by following the FAIR data principles. Therefore, a PID is necessary but not the only precondition. As atmospheric model data are often published with a DataCite DOI, we referred to the respective metadata schema in this text. Nevertheless, our recommendations are also applicable for other PIDs if they are associated with metadata.

If data are published with a DataCite DOI, FAIRness can be achieved by several means:

- All reasonable metadata parameters of the DOI should be filled.

[20] https://support.datacite.org/docs/landing-pages

[21] https://schema.org/

[22] https://developers.google.com/search/docs/data-types/dataset

[23] DOI: 10.5438/0000-00cc

[24] DOI: 10.5065/D6H70CW6

- PIDs should be used to link metadata to external sources, e.g. a publication in ESSD, documentations, citations, persons, organisations, etc.
- All metadata are listed on the landing page (human- and machine-readable).
- If one DOI is given for a dataset collection with several datasets, then both the dataset collection (low granular level) and the respective datasets (higher granular level) are described on the landing page or on sub-pages.
- Data are stored in files with a self-descriptive and open format, e.g. in netCDF files. The header of each file contains rich metadata, which are consistent with the given metadata on the landing pages.

FAIRness of datasets can be tested with several tools, which were collected and described by Bahim et al. (2019) and Wilkinson et al. (2018). Nevertheless, there are other principles which enhance the accessibility, usability and interoperability of data, e.g. the 5 Stars Linked Open Data (LOD) principles by Berners-Lee (2010). A comparison of FAIR and LOD was done by Hasnain and Rebholz-Schuhmann (2018). One important difference between the two principles is that FAIR data do not have to be openly available – only the licence agreement must be specified. As the reusability of data with restricted access is diminished, this should be avoided for atmospheric model data.

In theory, the FAIRness is a very useful concept. In practice, metadata fields have to be filled with correct information appropriate in detail and wording. The data producer, as the expert on the data, has to provide this information. However, the data curators and repository staff should ensure that the metadata are properly filled by checking the metadata. This ensures the FAIRness of the stored data. Information about these maturity controls can be added to the metadata by linking a description of the maturity control via the *Related Identifier* parameter. Nevertheless, it would be advantageous if the maturity of a dataset could be indicated explicitly in the DOI's metadata. Therefore, we suggest that a new DataCite metadata property for the maturity of the dataset should be introduced (Heydebreck et al., 2020). The maturity property should only refer to the individual datasets and differ from repository certifications.

Altogether, atmospheric model data are valuable, so that its producers should aim for their reusability. This can be reached by the FAIR principles. We hope that this guide helps to make atmospheric data FAIRer. Nevertheless, FAIRification of data is an open process and following these recommendations does not mean, that atmospheric model data will reach the maximum of FAIRness, especially as we have to assume that other needs will show up in future. As well, there are further possible actions to ensure better machine readability and processibility, see e.g. Jacobsen et al. (2020), and preferably DataCite metadata should be published in RDF as Open Linked Data, see Peroni et al. (2016). This will be investigated within the project AtMoDat and described in future publications.

Acknowledgments

The AtMoDat project is funded within the framework of "Forschungsvorhaben zur Entwicklung und Erprobung von Kurationskriterien und Qualitätsstandards von Forschungsdaten" by the German Federal Ministry of Education and Research (BMBF; FKZ: 16QK02A, 16QK02B, 16QK02D). We thank the DataCite team, the CF Conventions community and the other AtMoDat partners for fruitful discussions. The publication of this article was funded by the Open Access Fund of Technische Informationsbibliothek (TIB).

Appendix

Most DataCite Metadata properties are easy to understand, see DataCite Metadata Working Group (2019). The properties for the description of the dataset itself are DOI, AlternateIdentifier, Title, ResourceType, Summary, Size, Format,Version and Licence. All persons and institutions connected to the production and publication of the dataset are mentioned with the properties Creator, Publisher, Contributor and Funding. When the dataset was published, updated, created and when its is available is stated in PublicationYear and Date with its subproperties Created, Updated, Issued, and Available.

Additional information can be given with the properties and subproperties in the following tables:

Table 1: DataCite Property RelatedIdentifier for all information about the production of the data and connections to other data.

Landing Page	DataCite Property	Comment
Boundary Conditions	RelatedIdentifier: relationType = "IsDerivedFrom"	Link or DOI describing the boundary conditions applied
Model Documentation	RelatedIdentifier: relationType = "IsDescribedBy"	Either PID of the model or PID of the description of the model, which was used to calculate the data
References	RelatedIdentifier: relationType = "IsCitedBy"	Citation and PID of the publication, for which the data was used
Simulation is part of	RelatedIdentifier: relationType = "IsPartOf"	Only if applicable – PID of the MIP for which the simulation was made.
Related Simulations	RelatedIdentifier relationType = "IsVariantFormOf"	Only if applicable – PIDs of the descriptions of other simulations, that were made for the same MIP.
Maturity Check	RelatedIdentifier: relationType = "IsReviewedBy"	Link to the documentation of the performed quality checks.

Table 2: Information about the dataset, which are necessary but for whom there are no obvious DataCite properties. Most of this information must be written on the landing page.

Landing Page	**DataCite Property**	**Comment**
Model	–	Name of the model, which was used for the calculation of the data
Model version	–	Version of the model, which was used for the calculation of the data
Horizontal Resolution	–	Horizontal resolution of the data
grid	–	Grid specification
Projection	–	Used geographic projection
Vertical Coordinate	–	Vertical coordinate of the model, e.g. height, sigma, pressure, …
Temporal Coverage	Date: dateType = “Valid”	Temporal coverage of a time series
Spatial Coverage	GeoLocation	Lon/lat Box and/or name of a place or a region
Basic Approximations	–	Basic approximations used, e.g. hydrostatic, non-hydrostatic, …

Table 3: Information about each single file or variable: if only one DOI is assigned to a dataset collection including many datasets, this information is given on the landing page. Otherwise it is included in the DOI’s metadata.

Landing Page	**DataCite Property**	**Comment**
Variable/Dataset Name	Part of Subject or Description	Name of the dataset or variable
Temporal Aggregation	Part of Description	Temporal aggregation of the data, e.g. hourly, daily, monthly means, …
Spatial Aggregation	Part of Description	Spatial mean over regions, e.g. Europe, North Sea, …
Dimension	Part of Description	Dimension of the data: 1D, 2D, 3D, 4D
Valid Range	Part of Description	Only if applicable – valid range for specific variables, e.g. wind directions (0°–360°) or temperature given in K (> 0 K), …
Size	Size	File size
Spatial Coverage	Part of Description	Precise information about the model area, if it is so small that it can’t be specified with lon/lat box

References

Albertoni, R., D. Browning, S. Cox, A. Gonzalez Beltran, A. Perego, P. Winstanley, 2019: Data catalog vocabulary (DCAT). – Data Catalog Vocabulary (DCAT) – Version 2, https://www.w3.org/TR/vocab-dcat-2/.

Bahim, C., M. Dekkers, B. Wyns, 2019: Results of an analysis of existing FAIR assessment tools. – Research Data Alliance, DOI: 10.15497/RDA00035.

Berners-Lee, T., 2010: Is your linked open data 5 star?. – Linked Data, https://www.w3.org/DesignIssues/LinkedData.html.

DataCite Metadata Working Group, 2019: Datacite metadata schema documentation for the publication and citation of research data. – DataCite Metadata Schema 4.3, https://schema.datacite.org/meta/kernel-4.3/.

DCMI Usage Board, 2020: DCMI metadata terms. – DCMI Usage Board, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/. accessed 2020-03-09.

Eaton, B., J. Gregory, B. Drach, K. Taylor, S. Hankin, J. Blower, J. Caron, R. Signell, P. Bentley, G. Rappa, H. Höck, A. Pamment, M. Juckes, M. Raspaud, R. Horne, T. Whiteaker, D. Blodgett, C. Zender, D. Lee, 2019: NetCDF climate and forecast (CF) metadata conventions. – Published online, http://cfconventions.org/cf-conventions/cf-conventions.pdf.

Enabling FAIR Data Community, de A. Waard, H. Cousijn, J. Heber, B. Hanson, M. Bradford, M. Friedman, S. Hou, P. Jones, E. Kavanagh, K. Perry, H. Smith, J. VanDecar, J. Yeston, 2018: Author guidelines for enabling FAIR data in the earth, space, and environmental science. – Zenodo, DOI: 10.5281/ZENODO.1447108.

European Commission, 2016: Guidelines on FAIR data management in horizon 2020. – Published online, https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Eyring, V., S. Bony, G.A. Meehl, C.A. Senior, B. Stevens, R.J. Stouffer, K.E. Taylor, 2016: Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. – Geosci. Model Develop. **9**, 1937–1958, DOI: 10.5194/gmd-9-1937-2016.

FAIR Data Maturity Model Working Group, 2020: FAIR data maturity model: specification and guidelines. – Technical report, Research Data Alliance, DOI: 10.15497/RDA00045.

Hasnain, A., D. Rebholz-Schuhmann, 2018: Assessing FAIR data principles against the 5-star open data principles. – In: Gangemi, A., A.L. Gentile, A.G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J.Z. Pan, M. Alam (Eds.): The Semantic Web: ESWC, volume 11155. – Springer International Publishing, 469–477, DOI: 10.1007/978-3-319-98192-5_60.

Heydebreck, D., A. Ganske, A. Kraft, A. Kaiser, 2020: The data maturity indicator concept. – Webinar, DOI: 10.5446/47696.

ISO19108, 2002: ISO 19108:2002(en) geographic information – temporal schema. – ISO, https://www.iso.org/.

ISO8601, 2019: ISO 8601-1:2019(en) date and time – representations for information interchange – part 1: basic rules. – ISO https://www.iso.org/.

Jacobsen, A., de R. Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C.T. Evelo, C. Goble, G. Guizzardi, K.K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R.W.W. Hooft, M. Imming, K.G. Jeffery, R. Kaliyaperumal, M.G. Kersloot, C.R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S.A. Sansone, L.O.B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M.D. Wilkinson, E.L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, 2020: FAIR Principles: Interpretations and Implementation Considerations. – Data Intelligence **2**, 10–29, DOI: 10.1162/dint_r_00024.

Juckes, M., K.E. Taylor, P.J. Durack, B. Lawrence, M.S. Mizielinski, A. Pamment, J.Y. Peterschmitt, M. Rixen, S. Sénési, 2020: The CMIP6 data request (DREQ, version 01.00.31). – Geosci. Model Develop. **13**, 201–224, DOI: 10.5194/gmd-13-201-2020.

Mons, B., C. Neylon, J. Velterop, M. Dumontier, L.O.B. da Silva Santos, M.D. Wilkinson, 2017: Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the european open science cloud. – Information Services & Use **37**, 49–56, DOI: 10.3233/ISU-170824.

Neumann, D., V. Matthias, J. Bieser, A. Aulinger, 2017: Concentrations of gaseous pollutants and particulate compounds over northwestern Europe and nitrogen deposition into the North and Baltic Sea in 2008. – World Data Center for Climate (WDCC), DKRZ, DOI: 10.1594/WDCC/CMAQ_CCLM_HZG_2008.

NIMA, 2000: Department of defense world geodic system 1984 – its definition and relationships with local geodic systems. – Technical Report 3rd Edition, Amendment 1, Geodesy and Geophysics Department, National Imagery and Mapping Agency (NIMA), Department of Defense, NIMA(GIMG), USA, NSN: 7643-01-402-0347.

Peroni, S., D. Shotton, J. Ashton, A. Barton, E. Gramsbergen, M.C. Jacquemot, 2016: DataCite2RDF: Mapping datacite metadata schema 3.1 terms to RDF. – Figshare, DOI: 10.6084/m9.figshare.2075356.

Schema.org Steering Group, 2019: Schema.org. – Schema.org, http://schema.org/.

Signell, R.P., S. Carniel, J. Chiggiato, I. Janekovic, J. Pullen, C.R. Sherwood, 2008: Collaboration tools and techniques for large model datasets. – J. Marine Sys. **69**, 154–161, DOI: 10.1016/j.jmarsys.2007.02.013.

Taylor, K.E., R.J. Stouffer, G.A. Meehl, 2012: An overview of CMIP5 and the experiment design. – Bull. Amer. Meteoro. Soc. **93**, 485–498, DOI: 10.1175/BAMS-D-11-00094.1.

Wilkinson, M.D., M. Dumontier, I. Aalbersberg, J. Jsbrand, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, 't P.A.C. Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, 2016: The FAIR guiding principles for scientific data management and stewardship. – Scientific data **3**, 160018, DOI: 10.1038/sdata.2016.18.

Wilkinson, M.D., M. Dumontier, S.A. Sansone, L.O.B. da Silva Santos, P. Prieto, M. McQuilton, J. Gautier, D. Murphy, M. Crosas, E. Schultes, 2018: Evaluating FAIR-compliance through an objective, automated, community-governed framework. – bioRXiv, the preprint server for Biology, DOI: 10.1101/418376.

The pdf version (Adobe Java Script must be enabled) of this paper includes an electronic supplement:
JSON Example for DOI Metadata