ATMODAT Standard v3.0

Anette Ganske, Angelina Kraft,
Technische Informationsbibliothek (TIB)

Amandine Kaiser, Daniel Heydebreck, Andrea Lammert,
Heinke Höck, Hannes Thiemann,
Deutsches Klimarechenzentrum (DKRZ)

Vivien Voss, David Grawe, Bernd Leitl, K. Heinke Schlünzen,
Universität Hamburg, Meteorologisches Institut

Jan Kretzschmar, Johannes Quaas
Universität Leipzig, Institut für Meteorologie

11 December 2020

Contents

1. Introduction

It is an established procedure to publish observational atmospheric data in an accessible, standardised format. Now also for atmospheric model data, the standardised publication of output e.g. in the 5th and 6th Coupled Model Intercomparison Projects (CMIP, Taylor et al. (2012) and Eyring et al. (2016)) has proven to contribute outstandingly to successful science. AtMoDat [1] (**At**mospheric **Mo**del **Dat**a) proposes a flexible and easy-to-use standardisation for all atmospheric model data beyond the large projects such as CMIP.

In this standard, we follow the FAIR data principles, see e.g. Wilkinson et al. (2016), which demand that data should be published in a way, that makes them **f**indable, **a**ccessible, **i**nteroperable, and **r**eusable. Even though the FAIR data principles are often meant for increasing machine readability and interoperability, they can be interpreted to be guidelines for human readability and usability of data as well. Additionally, reusability of data is enhanced if it is published as open as possible.

The ATMODAT standard is part of the newly introduced **Ea**rth **Sy**stem **Da**ta **B**randing (EASYDAB)[2]. The requirements of EASYDAB are that published data

- have a DataCite DOI[3] and
- follow the corresponding rules for metadata, files and landing pages, either defined in the ATMODAT standard or a comparable standard.

Even though we focus on results of atmospheric models in this current standard, most of the principles can also be applied to model data of other realms (compartments) in the earth system. The ATMODAT standard does not only give instructions for the curators of the repositories who handle the assignment of the DOIs to the data and set up the landing pages. It also guides data producers which information they shall hand over to the repository. Moreover, this standard is also useful for data users, as they can learn which vocabulary they should use in case they are searching for data and which standards are met by the metadata and data.

In section 2 we specify the scope of this standard. In section 3 we present materials and methods. In section 4 we explain 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. A short summary follows in section 5.

Curators, who only want an overview over the required specifications, can use Tables 12 to 14. Look-up tables for data producers are Tables 15 and 14.

2. Scope

This standard provides guidelines to be followed for sucessful application for a DOI with EASYDAB by specifying

- the necessary metadata for all three: the DOI, the landing pages, and the header of the data files,
- the use of controlled vocabulary (CV), and
- the assignment of open licences and the specification of an embargo.

By following the ATMODAT standard, the FAIRness of data is considerably enhanced and data files follow standards of the climate model community. Even though the ATMODAT standard is focused on atmospheric model data, it might also be applied to results of other realms[4] of earth system models. In this case, the ATMODAT standard might need some adaptation and a realm specific standard can be provided to the community (e.g. OCMODAT for oceans or so).

[1] `https://www.atmodat.de`
[2] `https://easydab.de`
[3] `https://datacite.org`
[4] compartments

3. Materials and Methods

Within this standard we will use the following wording, which is taken from the specifications in Bradner (1997):

Must, required, shall: These words mean that the definition is an absolute requirement of the specification.

Must not, shall not: These phrases mean that the definition is an absolute prohibition of the specification.

Should, recommended: These words mean that there may exist valid reasons in particular circumstances to ignore a particular item, but the full implications must be understood and carefully weighed before choosing a different course.

3.1. Definitions

We use the following definitions:

Data: numbers, typically with physical units. Data can be structured or unstructured.

Data file: a digital file which contains data, e.g. results of a model simulation.

Metadata: data that serves to provide context or additional information about other data[5].

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

Dataset collection: a collection of several datasets.

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.

Realm: compartment of the earth system. CMIP uses the following realms: Aerosol, Atmosphere, Atmospheric Chemistry, Land Surface, Land Ice, Ocean, Ocean Biogeochemistry, and Sea Ice.

Curator: Institution/person providing the data to the public in a repository.

Data producer: Institution/person having created the data.

Data user: Institution/person using the data for own analyses, forcing of models etc.

3.2. Metadata

Metadata of data are describing characteristic aspects of the particular data. Metadata are attached to data, publications, files and other things for various purposes, e.g. bibliographic citation or administrative tasks. 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. In most cases, one aggregation of metadata is attached to the DOI of a dataset (see Chapter 4.1.2), another one to the landing page (see Chapter 4.2) and a third aggregation is part of the dataset itself (see Chapter

[5] http://www.businessdictionary.com/definition/metadata.html

3.5). These three aggregations can coincide but are all individually neccessary. 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, often there is more information needed for the DOI metadata than can be extracted from the metadata of the datasets. Further information about metadata can be found in Appendix A.

The DataCite Metadata Schema 4.3 (DataCite Metadata Working Group (2019)), which is used for this standard, is associated with DataCite DOIs and is meant for general purpose. It evolved from Dublin Core, but changed over time due to requests from DataCite community members. It 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 hierachical tree structure, e.g. *creatorName*, *nameType*, *affiliation*, ... Compliance with the DataCite Metadata Schema 4.3 has to be ensured by the curator, but the data producers have to supply all needed information. Former DataCite Metadata Schema can also be used, but the curator must ensure compliance with this standard.

3.3. DOI Assigment

Results of a model experiment have several levels of detail, ranging from the simulation itself to the values of each individual variable. This is called granularity (see Appendix B). The levels of granularity increase between the simulation and the single value. DataCite DOIs are 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. Additional DOIs might be assigned to the elements of higher granularity, e.g. the individual datasets of a simulation. 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 however, only one DOI is assigned to the lowest granularity level, like the example in Figure 1.

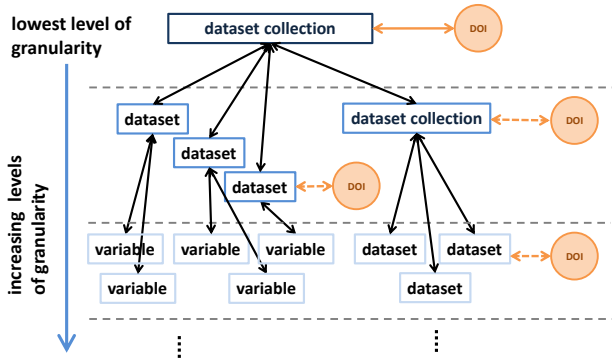

Figure 1: Example for a possible structure of a dataset collection, which contains the results of an atmospheric model simulation and the corresponding metadata. If this is defined as the lowest level of granularity and if a DOI is assigned to this dataset collection, then the orange arrow represents the link from and to its 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 lines), e.g. to individual datasets.

3.4. Landing Page

Beside the data repository as a whole, the landing page is the first "point of contact" between most of the data users and the 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. Both, the landing page and the subordinate web pages, must be publicly and permanently available. The layout of the landing page may change and metadata may be added.

DataCite DOIs are always connected to a HTML landing page[6], which must be human- and machine-readable. Resolving a DOI in a conventional web browser will redirect to this respective landing page. The machine-readable part of the landing page should be harvestable by search-engines. For every registered DOI, DataCite also provides a landing page within the DataCite Search[7], that is human- and machine-readable.

3.5. Data Files

A condition for the interoperability of data is that they are stored in a standardised, 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 to store the output data together with metadata in the data files.

For atmospheric model data, the netCDF[8] format is appropriate for this purpose because netCDF files are self-describing. More information about common data formats for atmospheric model data can be found in Appendix C.

The CF Conventions (Climate and Forecast Conventions, Eaton et al. (2019)) are one of the most widely used standards for netCDF data in the atmospheric modelling community. More information about the CF Conventions is given in Appendix D.

4. Specifications

The following guidelines aim to improve the FAIRness of atmospheric model datasets. It is assumed that the dataset is provided with a DataCite DOI. Being in compliance with these guidelines is mandatory for the ATMODAT standard:

- Machine readability is facilitated, see section 4.1.1.
- The complete DataCite DOI Metadata Schema is used, see section 4.1.2.
- All DOI metadata properties are also listed on the landing page, see section 4.2.
- Additional discipline specific metadata are enclosed in the metadata field *Description* and on the landing page, see sections 4.1.3 and 4.2.
- Open and non-proprietary file formats are used, see section 4.3.
- Controlled vocabulary is used as much as possible, see sections 4.1 and the lists with controlled vocabularies in Appendix F.
- An open licence is used. Any embargo can only be of limited duration, see section 4.4.

4.1. Specifications for Metadata

Metadata must describe the data in such a way that the possible data users are able to decide whether these data are beneficial for their application. Additionally, the process of the data generation and used input data must be documented. All necessary information 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 e.g. be the documentation of the numerical model, that was used to calculate the data.

It is strongly recommended that metadata are machine-readable so that they can be used to create automated lists, e.g. for evaluations of institutions. External documents itself, but at least their metadata, should also be machine-readable.

[6] DOI Handbook, Chapter 5: `https://www.doi.org/doi_handbook/5_Applications.html`
[7] https://search.datacite.org/
[8] https://doi.org/10.5065/D6H70CW6

4.1.1. FAIR principles for the Metadata

It is strongly recommended to increase machine-readability and therefore the FAIRness of the metadata by considering the following principles:

- According to the recommendations in the DataCite Blog by M. Fenner[9], the language of the dataset must always be specified according to ISO 639-1 (ISO639 (2002)).
- All information about persons and institutions should be complemented with a PID, i.e. an ORCID[10] for persons or a ROR[11] for institutions (if applicable). More information about PIDs can be found in Appendix G
- All links to documents, homepages etc. should be provided as PIDs (see Appendix G), e.g. documents with a DOI and homepages preferably with an URN.
- All temporal information must be given in a standardised way, e.g. according to ISO 8601 (ISO8601 (2019)) and ISO 19108 (ISO19108 (2002)).
- Keywords and subjects should be taken from controlled vocabularies (CVs). Lists of controlled vocabularies are given in Appendix F.
- Geographic information must include the *spatial reference system* (also named *coordinate reference system*), e.g. WGS84 (NIMA (2000)). If no reference system is provided, WGS84 is assumed. In any case geographic coordinates of single points or the boundaries of a modeled domain must be provided in WGS84 - if another coordinate reference system was used, coordinates must be recalculated. For details see the flow chart in Figure 3.
  Geographic names should be chosen according to geonames[12], if possible.
- Usage rights/licence must be open and always be described. Preferably, a standard machine-readable licence should be used, see Chapter 4.4.

4.1.2. DOI Metadata

The metadata of the lowest level of granularity is attached to the DataCite DOI, see the example in Figure 1. In order to increase the FAIRness of atmospheric model data, it is strongly recommended to use **all reasonable** DataCite metadata properties. Table 2 shows a list of all DataCite metadata properties, which have to be provided by the curator. Data producers find a list of all information, which they have to contribute, in Table 3.

The following principles exist for the respective properties:

- The DataCite metadata properties *Contributor* and *Creator* must 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. For the same reasons it is strongly recommended to add *Funder*, if applicable.
- All dates connected to the production or publication of the dataset are important information for the data user. Therefore, it is strongly recommended to mention them via the metadata properties *Date*. Except from the Publication Year, all other temporal information is added in the subproperty *dateType*, see Table 4.
- The length of time series or the period for which a simulation is valid, should be noted in the DataCite metadata property *Date* and the subproperty *valid* according to the flow chart in Figure 2.

[9] https://doi.org/10.5438/1dgk-1m22
[10] https://orcid.org/
[11] https://ror.org/
[12] https://www.geonames.org/

- The temporal coverage of the performed simulation should be stated according to the flow chart in Figure 2.
- The description of the dataset/dataset-collection shall briefly describe the simulation and the purpose of the simulation. It must also contain all metadata for which no DataCite metadata properties exist, see Table 7.
- The description of the geographical location of the model area (*GeoLocation*) is important for the reusability of atmospheric model data. It should be described according to the flow chart in Figure 3.
- In order to increase the interoperability, all possible references to other data should be listed; e.g. if data was calculated in an model intercomparison project (MIP), the data of other models included in the MIP, the boundary conditions, or a publication about the data. For these purposes the *RelatedIdentifier* and values for the subproperty *relationType* are used, see Table 6.
- Information about any maturity checks should be added to the metadata, if available. This can be done with a documentation of the maturity checks which were performed 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"*, see Table 6.

Only if all applicable metadata properties are provided, all the necessary information is present to enable a further reuse of the data. An example for the DataCite DOI's metadata of an atmospheric dataset of Neumann et al. (2017), written in JSON, can be found in Appendix L.

4.1.3. Additional Metadata

DataCite metadata properties are only describing the lowest level of granularity. Additional information about the datasets can only be included in the DataCite metadata parameter *Description* and on the landing page, see Table 10.

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 provided on the landing page and with the metadata of the data files itself. For atmospheric model data, it is strongly recommended to use the properties in Tables 9 and 10, if applicable.

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

4.2. Specifications for the Landing Page

The ATMODAT standard has the following requirements for the landing page:

- It must contains both the human-readable and machine-readable metadata.
- It must include a complete citation of the dataset in human-readable format and the DOI itself[13], so that the data record can be uniquely identified by humans.
- The DOI shall be stored in the machine-readable part of the landing page, so that search engines also can find it.
- It contains information on how to access the data. If the data record itself is no longer available, this shall be noted (Tombstone Page).

[13] according to *Best Practices for DOI Landing Pages of DataCite*, see https://support.datacite.org/docs/landing-pages

- All metadata fields that are in the metadata record of the DataCite DOI (see 4.1.2) shall be listed. Nevertheless, metadata fields on the landing page may have different names than the DataCite metadata properties. For example the maturity information, which is stored in the DataCite Metadata Schema under *Related Identifier:isReviewedBy*, can be directly named as "Maturity Information".
- It is strongly recommended that the landing page contains additional and needed information, for which no fields exist in the DataCite Metadata Schema, see Tables 9 and 10 in the Appendix.
- The machine-readable part of the landing page should be provided compliant to schema.org[14] 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)) may be used, see e.g. the data description from Google[15]. A mapping of the DataCite Metadata Schema to schema.org is provided by DataCite[16].

If the DOI was assigned to a dataset collection with several datasets or to datasets with several variables, then the following instructions are valid for the landing page:

- The landing page shall always expose the metadata of the low level granularity and it is strongly recommended to add 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 *License, Contributor*, ..., can be repeated. It is strongly recommended to provide additional information according to Tables 9 and 10.

4.3. Specifications for File Formats and Standards

The ATMODAT standard requires that:

- Atmospheric modeling data is archived in netCDF[17] files.
- NetCDF file headers include description of time, coordinate and vertical axes according to Appendix E.
- CF Conventions version 1.4 or higher are applied, see Appendix D.
- Global attributes from Table 11 are set in the netCDF files. The values of these attributes should be taken from controlled vocabularies, if available.

4.4. Specifications for Licencing and Embargos of Published Data

As the reusability of data with restricted access is diminished, this should be avoided for atmospheric model data. The ATMODAT standard has the following requirements:

- Data must be published under an open licence[18] that is at least available in English language.
- It is strongly recommended to choose a licence that is machine-readable and use its standard name, if it is provided by SPDX [19] (Software Package Data Exchange).
- Additionally, an embargo[20] is allowed but must be limited in time. The length of the time period itself can be freely choosen but has to be publicly visible.

[14] https://schema.org/
[15] https://developers.google.com/search/docs/data-types/dataset
[16] https://doi.org/10.5438/0000-00cc
[17] https://doi.org/10.5065/D6H70CW6
[18] open/free licence: the data are freely available with some restrictions (e.g. free non-commercial use), see e.g. `https://opendefinition.org/licences/`
[19] `https://spdx.dev/`
[20] embargo: no one is allowed to access data before the end of the embargo period

5. Summary

The aim of the ATMODAT standard is the enhancement of reusability of archived atmospheric model data. The standard requires that atmospheric model data is published with a DataCite DOI and therefore only the respective metadata schema is refered to in this document. Nevertheless, these recommendations are also applicable for other PIDs if they are associated with metadata.

The standard demands FAIRness for (meta-)data, which is achieved by several means:

- All reasonable metadata parameters of the DOI should be filled.
- The use of PIDs to link metadata to external sources, e.g. documentation, citations, persons, organisations, etc.
- All metadata for the DOI are also 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 netCDF files. The header of each file contains rich metadata.

In theory, 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, being the expert for the data, has to provide this information. However, the curators and repository staff should ensure that the metadata are properly filled by performing maturity controls of the metadata. Information about these maturity controls should be added to the metadata by linking a description of the quality control via the *Related Identifier* property. Nevertheless, FAIRification of data is an open process and following these recommendations does not ensure, that atmospheric model data will reach the maximum of FAIRness. There are further possible actions to ensure better machine-readability and -processibility (Jacobsen et al. (2020)). This will be investigated within the project AtMoDat and further versions of the ATMODAT standard will be published in future.

References

Albertoni, R., Browning, D., Cox, S., Gonzalez Beltran, A., Perego, A. and Winstanley, P. (2019), 'Data catalog vocabulary (DCAT)'.
URL: *https://www.w3.org/TR/2019/PR-vocab-dcat-2-20191119/*

Bradner, S. (1997), 'Key words for use in RFCs to indicate requirement levels'. Internet RFC 2119.
URL: *http://www.ietf.org/rfc/rfc2119.txt*

DataCite Metadata Working Group (2019), 'Datacite metadata schema documentation for the publication and citation of research data'. doi: https://doi.org/10.14454/f2wp-s162.
URL: *https://schema.datacite.org/meta/kernel-4.3/*

DCMI Usage Board (2020), 'DCMI Metadata Terms'. accessed 2020-03-09.
URL: *https://www.dublincore.org/specifications/dublin-core/dcmi-terms/*

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Rich, S., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Horne, R., Whiteaker, T., Blodgett, D., Zender, C. and Lee, D. (2019), 'NetCDF Climate and Forecast (CF) Metadata Conventions'.
URL: *http://cfconventions.org/cf-conventions/cf-conventions.pdf*

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J. and Taylor, K. E. (2016), 'Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization', *Geoscientific Model Development* **9**(5), 1937–1958. doi: https://doi.org/10.5194/gmd-9-1937-2016.

ISO19108 (2002), 'ISO 19108:2002(en) Geographic information — Temporal schema'.
URL: *https://www.iso.org/standard/26013.html*

ISO639 (2002), 'ISO 639-1:2002. Codes for the representation of names of languages — Part 1: Alpha-2 code'.
URL: *https://www.iso.org/standard/22109.html*

ISO8601 (2019), 'ISO 8601-1:2019(en) Date and time — Representations for information interchange — Part 1: Basic rules'.
URL: *https://www.iso.org/standard/70907.html*

Jacobsen, A., Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M. and Dumontier, M. (2020), 'FAIR Principles: Interpretations and Implementation Considerations', *Data Intelligence* **2**(1-2), 10–29. doi: https://doi.org/10.1162/dint_r_00024.

Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., Peterschmitt, J.-Y., Rixen, M. and Sénési, S. (2020), 'The cmip6 data request (dreq, version 01.00.31)', *Geoscientific Model Development* **13**(1), 201–224. doi: https://doi.org/10.5194/gmd-13-201-2020.
URL: *https://www.geosci-model-dev.net/13/201/2020/*

Madden, F., van Horik, R., van de Sandt, S., Lavasa, A. and Cousijn, H. (2020), 'Guides to Choosing Persistent Identifiers - Version 2'. doi: 10.5281/zenodo.3956569.
URL: *https://doi.org/10.5281/zenodo.3956569*

Neumann, D., Matthias, V., Bieser, J. and Aulinger, A. (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) at DKRZ. doi: https://doi.org/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), Mail Stop L-41, 3200 South Second Street, St. Louis, MO 63118-3399, USA. NSN: 7643-01-402-0347.
URL: *https://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html*

Schema.org Steering Group (2019), 'Schema.org'. accessed 2019-12-12.
URL: *http://schema.org/*

Signell, Richard P.and Carniel, S., Chiggiato, J., Janekovic, I., Pullen, J. and Sherwood, C. R. (2008), 'Collaboration tools and techniques for large model datasets', *Journal of Marine Systems* **69**(1), 154 – 161. doi: https://doi.org/10.1016/j.jmarsys.2007.02.013.

Taylor, K. E., Stouffer, R. J. and Meehl, G. A. (2012), 'An overview of cmip5 and the experiment design', *Bulletin of the American Meteorological Society* **93**(4), 485–498. doi: https://doi.org/10.1175/BAMS-D-11-00094.1.

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

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, 16QK02C und 16QK02D). We thank the DataCite team, the CF Conventions community and our collegues in the different institutions for fruitful discussions.

A. Additional Information about Metadata

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, some domain specfic ones, such as the Climate and Forecast Metadata Conventions (CF Conventions, Eaton et al. (2019)), have been established. Metadata fields storing the same information might be named differently between different schemata. Furthermore, one field might exist in one schema but not in others. Hence, it is necessary to create a mapping between different metadata schemata, so that metadata in one format can be transformed into 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[21] or OWL[22] so that the metadata is machine-readable.

B. The Granularity of Archived Data

Results of atmospheric simulations are usually available in various levels of granularity (i.e. levels of detail). As the atmosphere is a complex system with many degrees of freedom, the output of a simulation with an atmospheric model consists of many variables. It includes 1-dimensional (1D) properties and 2D and 3D fields, as well 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 files. Specific diagnostics may further store statistics with further dimensions at each grid-box and time step.

Various temporal aggregations, such as annual and monthly averages, are often written into separate files. Data may be redistributed into a different file structure during post-processing. For example, it was specified within CMIP that all model results are stored as such that each netCDF file only contains one single variable. Alternatively, multiple variables can be stored in a joint netCDF file, if no conformance with the CMIP standard is aimed for.

Defining a simulation's levels of granularity determines a hierarchical structure of the data, e.g. for their archivation. The lowest level of granularity could be the simulation itself, only parts of a simulation (e.g. only the atmospheric results of an earth system model), or even a whole project in which several simulations were performed. As there are several possibilities to define a hierachical structure, the curator and the data producer should first jointly define which group of data forms the lowest level of granularity. Then, the higher levels of granularity are determined. The so defined structure should balance the interests of the repository and the data user:

- The repository has to store all data in a structure, which is easy to manage and which does not need too many DOIs to be assigned to.
- The data users want the data to be stored in small partitions, because usually they don't need the whole simulation. In that case, they want to be able to download only parts of the datasets.

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

C. File Formats in Atmospheric Modeling

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

[21] https://www.w3.org/RDF/

[22] https://www.w3.org/TR/owl2-overview/

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 modeling 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, e.g. generic analysis and visualization tools, 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, as the external GRIB tables do not contain standardised variable names but require the knowledge of the model to understand what a given *long name* is referred to.

D. Climate and Forecast Conventions

The Climate and Forecast Conventions (CF Conventions, Eaton et al. (2019)) are one of the most widely used standards in the atmospheric modelling community, which are also used for CMIP6 data (Juckes et al. (2020)). They consist of a conventions document, which mainly describes formulations for coordinates axes, as well as a standard name table, which is a controlled vocabulary for for the *standard name* attribute of each respective variable in a netCDF file.

.

E. Description of Model's Axes

Providing horizontal, vertical and temporal axes is optional in the CF Conventions. The conventions just prescribe how these axes have to be described when they are provided. However, having spatial and temporal information is often required for a proper reuse of atmospheric model data. Therefore, this standard requires this information under specific conditions:

- If data are horizontally resolved (e.g. *lon* + *lat* or *x* + *y*), then horizontal coordinate axes should be provided.
- If data have reasonable vertical information (e.g. *pressure* or *height*), then the vertical axis should be described.
- If data are not static in time (e.g. via dimension and variable *time*), then the time axis should be provided.

F. Controlled Vocabulary

Controlled vocabulary for the DataCite metadata property *Subject* and keywords for the landing page and netcdf-headers can be found in

1. **Field of science[23]:** `https://www.oecd.org/science/inno/38235147.pdf` or `https://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/6BB427AB9696C225CA2574180004463` `opendocument`
2. **Realm of the model:** `https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_realm.json`
3. **Description of Model Components:** ES-Doc `https://specializations.es-doc.org/static/index.html?target=cmip6.atmos&client=esdoc-url-rewrite`, `https://www.earthsystemcog.org/projects/es-doc-models/cim_cv`

[23] see the blog by M. Fenner `https://doi.org/10.5438/1dgk-1m22`

4. **Names of Variables:** CF-Conventions http://cfconventions.org/Data/cf-standard-names/72/build/cf-standard-name-table.html
5. **Description of Data:** http://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.pdf
6. **United Nations Terminology Database:** https://unterm.un.org/
7. **Global Change Master Directory (GCMD):** https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords
8. **Climate Tagger:** https://www.climatetagger.net/climate-thesaurus/
9. **EU Vocabularies (includes continents, countries and places):** https://op.europa.eu/en/web/eu-vocabularies/home
10. **Geographic identifiers for marine regions:** https://www.marineregions.org/
11. **Geonames:** https://www.geonames.org/
12. **Marine keywords:** https://www.bodc.ac.uk/data/codes_and_formats/seavox/
13. **Environnmental keywords:** https://www.eionet.europa.eu/gemet/
14. **Glossary of the Australien Bureau of Meteorology:** http://www.bom.gov.au/lam/glossary/
15. **Australien Bureau of Meteorology - Weather Words :** http://www.bom.gov.au/info/wwords/
16. **International Glossary of Hydrology:** http://www.wmo.int/pages/prog/hwrp/publications/international_glossary/385_IGH_2012.pdf
17. **CGI Observation Method vocabulary:** https://vocabs.ardc.edu.au/viewById/89

More lists can be found in BARTOC (Basel Register of Thessauri, Ontologies and Classifications), see http://www.bartoc.org/en.

G. Persistent Identifiers

Persistent Identifiers (PIDs) are used for the enduring identification of persons, documents, funders, etc. An introduction to PIDs can e.g. be found at the homepages of the FREYA project[24]. An overview for all PIDs and their usage are given in Madden et al. (2020).

As the following PIDs can be linked to each other by using the so called PID-Graph [25], we propose using these PIDs preferably (if available):

Persons: ORCID (https://orcid.org/),

Documents: DOI (https://doi.org/),

Institutions: ROR (https://ror.org/).

Additionally, PIDs for funders can be found in https://www.crossref.org/services/funder-registry/.

There will be more PIDs invented in the future, e.g. for conferences.

[24] https://youtu.be/9G4EMJCwCw4

[25] see e.g. https://youtu.be/2HLxN8p_rpQ

H. Metadata Tables

Table 1: Additional tables for metadatda information, which should either be included in the DOI's metadata or on the landing page.

Topic	**Table/Figure Number**
List of all DOI metatada properties - Table for curators	Table 2
List of all needed metadatda properties - Table for data producers	Table 3
Dates related to publication, updates etc. of the dataset	Table 4
Horizontal discretisation	Table5
Relations to other datasets, boundary conditions, documentation, etc.	Table 6
Elements of the description	Table 7
Boundary conditions	Table 8
Additional informations about the dataset	Table9
Additional informations about the respective files	Table 10
Temporal informations of the data itself	Figure 2
Spatial information of the data	Figure 3

Table 2: Detailed table for data curators: DataCite DOI metadata properties with status for the ATMODAT standard – mandatory (*status:* M) or recommended (R). Further explanations of the DataCite properties and subproperties can be found in DataCite Metadata Working Group (2019). Examples are enriched metadata taken from the dataset collection of Neumann et al. (2017)

DataCite ID	**Property**	**ATMODAT Status**	**Example**	**Description**
1	Identifier (with mandatory type sub-property)	M	`https://doi.org/10.1594/wdcc/cmaq_cclm_hzg_2008`	the DOI itself
2	Creator (with optional family name, given name, name identifier and affiliation sub-properties)	M	Neumann, Daniel, ..., `https://orcid.org/0000-0001-8574-9093`	It is strongly recommended to use ORCID for persons and ROR for affiliation, see Appendix G.
3	Title (with optional type sub-properties)	M	Concentrations of gaseous pollutants and particulate compounds over Northwestern Europe and nitrogen deposition into the North and Baltic Sea in 2008	Dataset title
4	Publisher	M	World Data Center for Climate (WDCC) at DKRZ	The name of the entity that holds archives, publishes prints, distributes, releases, issues, or produces the resource.
5	PublicationYear	M	2017	Year of publication
6	Subject	M	EASYDAB, ATMODAT, meteorology and atmospheric sciences, atmosphere	Always use several keywords, which must at least include: EASYDAB, ATMODAT, the field of science and the realm of the model, which must be taken from controlled vocabularies (CVs). More than one realm is possible.
		R	atmospheric chemistry, climate,....	It is strongly recommended to add further keywords, which also should be taken from CVs, if applicable.
	Subject scheme sub-properties	R	for "atmospheric chemistry" vocabulary= GEMET, `https://www.eionet.europa.eu/gemet/en/concept/623`	Name and URI of the controlled vocabulary
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	M	Matthias, Volker	All Contributors with names and PIDs, see Appendix G.
8	Date with date-type Created or Updated	M	dateType =created: 2017-06-08	Date of creation or update of the dataset/dataset-collection
	Date with other date-types than created/updated	R	dateType=valid:20080101/20081231	All dates for publication, updates etc., see Table 4. Always use ISO 8601 for dates. For specifying all dates connected to the data itself (e.g. length of time series), follow the flow chart in Figure 2.
9	Language	M	en	The language of the dataset must always be specified according to ISO 639-1 (ISO639 (2002)).
10	ResourceType (with mandatory general type description sub-property)	M	ResourceType=Digital, resourceTypeGeneral=Dataset	Further explanations in Table 5.
...	...	...	...	...

DataCite ID	Property	ATMODAT Status	Example	Description
...	...	...	...	...
11	AlternateIdentifier (with type sub-property)	R		If there is an alternative Identifier, like a second DOI, it should be specified.
12	RelatedIdentifier (with type and relation type sub-properties)	R	relationType=IsDocumentedBy: http://doi.org/10.1029/2001JD001409, relationType=isDerivedFrom: http://doi.org/10.5194/gmd-4-47-2011, relationType=isReviewedBy: https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=quality_checking_report relationType=References: http://doi.org/10.5194/acp-16-739-2016	It is strongly recommended to add related information about documentation, boundary conditions (see Table 8), publications citing the dataset, other related datasets (e.g. from older simulations or made in the same project), and maturity tests. Use PIDs, if applicable, see Appendix G.
13	Size	R	12459482227 Bytes	Size of the dataset or dataset collection, to which the DOI is assigned to.
14	Format	M	application/x-netcdf	Data format
15	Version	R	1	It is strongly recommended to versionize the datasets.
16	Rights	M	Creative Commons Attribution 4.0 International (CC-BY-4.0)	The ATMODAT standard requires an open licence that is at least available in English language (see chapter 4.4).
17	Description (with type sub-property)	M	Project: coastDat-2 - CoastDat-2 comprises a compilation of coastal analyses (that is hindcasts and reconstructions) and scenarios for the future obtained from numerical models. ...	Description/Abstract of the dataset/dataset collection, to which the DOI is assigned to.
18	GeoLocation(with point, box and polygon sub-properties)	R	geoLocationBox:eastBoundLongitude=36.9, northBoundLatitude=67, southBoundLatitude=42.6, westBoundLongitude=18.9, geoLocationPlace=Northwestern Europe	For specifying all metadata connected to the location of the data (e.g. model area), follow the flow chart in Figure 3.
19	FundingReference (with name, identifier, and award related sub-properties)	R	Deutsche Forschungsgemeinschaft, http://dx.doi.org/10.13039/501100001659	It is strongly recommended to name the funder and to add the respective DOI of the Funder registry (https://www.crossref.org/services/funder-registry/)

Table 3: For the DataCite DOI and the landing page needed metadata, which is provided by the data producers. Only those metadata are listed here , that describe the dataset/dataset collection, to which the DOI is assigned to. For the metadata of the individual files, see additional Table 10. Status is either mandatory (M) or recommended (R). **N**: Examples are enriched metadata of the dataset collection of Neumann et al. (2017)

Name	**ATMODAT Status**	**Example**	**Description**	**More Information**
Creator(s) (with optional family name, given name, name identifier and affiliation)	M	**N**: Neumann, Daniel,. . . , `https://orcid.org/0000-0001-8574-9093`	It is strongly recommended to provide ORCIDs for persons and RORs for affiliations, see Appendix G.	`https://orcid.org`, `https://ror.org`
Title	M	**N**: Concentrations of gaseous pollutants and particulate compounds over Northwestern Europe and nitrogen deposition into the North and Baltic Sea in 2008	Dataset title	
Subject	M	**N**: EASYDAB, ATMODAT, meteorology and atmospheric sciences, atmosphere	Always use several keywords, which must at least include: EASYDAB, ATMODAT, the field of science and the realm of the model , which must be taken from controlled vocabularies (CVs) . More than one realm is possible.	List of CVs in Appendix F
	R	atmospheric chemistry, climate,...	It is strongly recommended to add further keywords, which also should be taken from CVs, if applicable.	
Contributor(s) (with optional given name, family name, name identifier and affiliation)	M	**N**: Matthias, Volker	All Contributors with names and ORCIDS, if applicable.	`https://orcid.org`, `https://ror.org`
Date Created or Updated	M	**N**: 2017-06-08	Creation date or date of update (if applicable) of the dataset/dataset collection.	Table 4
Date: End of Embargo	R	2017-06-08	Date of the end of an embargo.	Table 4.
temporal information of the simulation, e.g. simulation period	R	**N**: 20080101/20081231	It is strongly recommended to mention all temporal information of the simulation, if data is time dependent.	Figure 2.
Language	M	en	The language of the dataset must always be specified according to ISO 639-1	ISO639 (2002)
Grid Specifications	R	Data was calculated on a grid	It is strongly recommended to mention the grid.	Table 5
Description	M	This dataset contains the results of a simulation . . .	Abstract/Description of the dataset/dataset collection, to which the DOI was assigned.	Table 7
. . .	. . .	. . .	. . .	. . .

Name	ATMODAT Status	Example	Description	More Information
...	...	...	...	...
Alternate Identifier	R		If there is an alternative Identifier, like a second DOI, it should be specified.	
Documentation of the model, that was used for the simulations	R	http://doi.org/10.1029/2001JD001409,	It is strongly recommended to refer to the documentation of the model used, preferably with a PID, see Appendix G.	
Boundary Conditions of the model, that was used for the simulation	R	http://doi.org/10.5194/gmd-4-47-2011	It is strongly recommended to indicate the boundary conditions, that were used for the simulations, preferably with a PID, see G.	Table 8.
Validation	R		It is strongly recommended to mention any validation, that were performed and described, preferably with a PID.	
maturity checks	R	https://cera-www.dkrz.de/WDCC/ui/cerasearch/entry?acronym=quality_checking_report	It is strongly recommended to mention any maturity checks, that were performed, preferably with a PID.	
Dataset was cited in	R		It is strongly recommended to mention any publication, that cites the dataset, preferably with a DOI.	
Simulation is part of	R		If applicable, PID or URL of the description of the Model Intercomparison Project, in which the simulation was made.	
Comparable Datasets	R		It is strongly recommended to mention any datasets, that are comparable, e.g. because they were calculated with the same model or with the same boundary conditions. Use a PID, if applicable.	
Size	R	12459482227 Bytes	Size of the dataset or dataset collection, to which the DOI is assigned to.	
File Format	M	application/x-netcdf	For ATMODAT standard always netcdf	
Version	R	**N:**1	It is strongly recommended to versionize the datasets.	
Rights	M	**N:**Creative Commons Attribution 4.0 International (CC-BY-4.0)	The ATMODAT standard requires an open licence that is at least available in English language. Use standard name, if applicable.	chapter 4.4 and https://spdx.dev
GeoLocation(either point, box or polygon)	R	**N:** geoLocationBox : eastBoundLongitude=36.9, northBoundLatitude=67, southBoundLatitude=42.6, westBoundLongitude=18.9, geoLocationPlace = Northwestern Europe	For specifying all metadata connected to the location of the data (e.g. area)	Figure 3
Funding Reference (with name, identifier, and award related sub-properties)	R	Deutsche Forschungsgemeinschaft, http://dx.doi.org/10.13039/501100001659	It is strongly recommended to name the funder and to add the respective DOI of the Funder registry.	https://www.crossref.org/services/funder-registry/

Table 4: Cases, where it is strongly recommended to state dates of publications of datasets or changes of published datasets via the DataCite property *Date* and its subproperties. At least one of the sub-properties *created, updated* or *issued* must be filled.

Action	**Date of the action**	**DataCite Subproperty of *Date***	**Information is given by whom**
Model results were calculated for the first time	Date, at which the computation of the dataset/dataset collection was finished	DateType = Created	Data producer
Published dynamical dataset,which has been prolongated	Date of the prolongation of the dataset	DateType=Updated	Data producer and curator
These model results were published for the first time	Date, at which the dataset/dataset collection was published	DateType = Issued	Curator
Embargo: the reuse of a dataset is only allowed after a specific date	Date of the end of the embargo	DateType = Available	Data producer

Table 5: Description of the model's horizontal discretisation, written in the DataCite properties *Description* and *ResourceType* and on the landing page. Table should be used by data producers and curators.

Horizontal Discretisation	**Specify grid sizes**	**DataCite *ResourceType***	In **DataCite *Description* and on Landing Page**
Fixed grid with constant grid sizes	yes	ResourceType=grid, resourceTypeGeneral = Dataset	Specify grid and corresponding grid sizes
Grid with variable grid sizes	no	ResourceType=grid, resourceTypeGeneral = Dataset	Specify grid and ranges of corresponding grid sizes
Spectral	no	ResourceType = Digital, resourceTypeGeneral=Dataset	"Spectral horizontal discretisation"
Other unstructured grid	no	ResourceType = Digital, resourceTypeGeneral = Dataset	Name of horizontal discretisation

Table 6: DataCite property *RelatedIdentifier* for all information about the production of the data and connections to other datasets.

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.
Quality Check	*RelatedIdentifier: relationType = "IsReviewedBy"*	Link to the documentation of the performed quality checks.

Table 7: DataCite property *Description*: Recommended elements should be mentioned, if applicable. Table is made for data producers. Use controlled vocabulary from the lists in Appendix F, as much as possible.

Topic	**ATMODAT Status**	**Example**	**Explanation**
Abstract	M	Description of the simulation and the project	Describe either the dataset or the dataset-collection, to which the DOI is assigned to.
simulation time information	R	1.1.2010 - 31.21.2010	State all temporal information of a simulation according to Figure 2) in the Appendix.
Calender used	R	Non-Gregorian calender	Unless otherwise noted, it is assumed that a Gregorian calender was used, see Figure 2 in the Appendix.
Grid	R	Data was calculated on a rectangular grid	Specify the grid, on which data was calculated, see Table 5. Use controlled vocabulary of ES-DOC, see Appendix F
Model	M	CCLM	Name of the model, which was used for the calculation of the data
Model version	R	1.3	Version of the model, which was used for the calculation of the data
Horizontal Resolution	R	10 km x 10 km	Horizontal resolution of the data, see Table 5.
Geographic Reference System	R	UTM	If not stated, WGS84 is assumed.
Vertical Coordinate	R	pressure [hPa]	Vertical coordinate of the model, e.g. height, sigma, pressure, . . .
Spatial Coverage	R	North Sea	Lon/lat Box and/or name of a place or a region, taken from a CV in Appendix F. Further information can be found in Figure 3 in the Appendix.
Basic Approximations	R	hydrostatic	Basic approximations used, e.g. hydrostatic, non-hydrostatic,. . .
Boundary Conditions	R	ERA-Interim	Specify boundary conditions used, see Table 8. Cite DOI or PID of the boundary data, if possible.
Possible usage of the data	R	Climate model calculations: only for investigations of mean values, where means are taken over at least 4 adjacent grid points and 10 years.	What was the motivation to calculate the data and for what can the data be used? Also mention degree of accurracy and any known problems with the data.

Table 8: Cases, where it is strongly recommended to mention boundary conditions in the DataCite property *Description* **and** on the landing page

Simulation	**Lateral Boundary Conditions**	**Vertical Boundary Conditions**	**Initial Conditions**
global model/data	no	yes	yes
limited area model (regional, urban, . . .)	yes	yes	yes
point data / time series	no	no	yes
vertical profile	no	yes	no
trajectory	no	no	yes

Table 9: Information about the dataset, which is necessary but with no obvious DataCite properties. It is strongly recommended or even mandatory to write this information on the landing page.

Landing Page	**AtMoDat Status**	**Example**	**Comment**
Model	M	CCLM	Name of the model, which was used for the calculation of the data
Model version	M	1.3	Version of the model, which was used for the calculation of the data
grid	R	regular grid, grid sizes 10 km x 10 km	see Table 5
Projection	R	–	Used geographic projection
Vertical Coordinate	R	pressure [hPa]	Vertical coordinate of the model, e.g. height, sigma, pressure, . . .
Temporal Coverage	R	Date: dateType = "Valid"	Temporal coverage of a time series
Spatial Coverage	R	GeoLocation	Lon/lat box and/or name of a place or a region
Basic Approximations	R	hydrostatic	Basic approximations used, e.g. hydrostatic, non-hydrostatic,. . .

Table 10: Information about each single file or variable: if only one DOI is assigned to a dataset collection including many datasets, it is strongly recommended to give this information 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
Size	Size	File size

I. Flow Charts

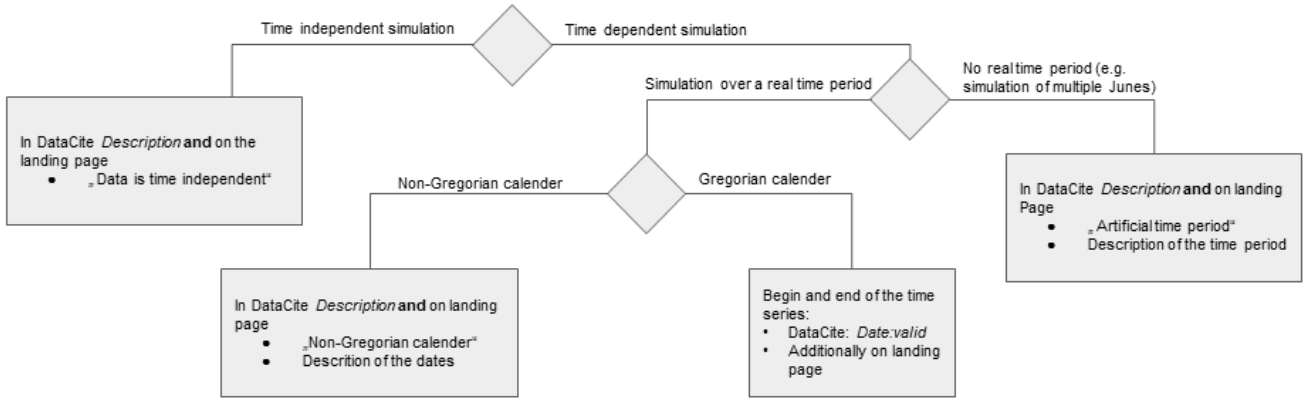

Figure 2: Flow charts for the metadata, describing the temporal properties of a dataset. Names of DataCite properties are written in *italic*.

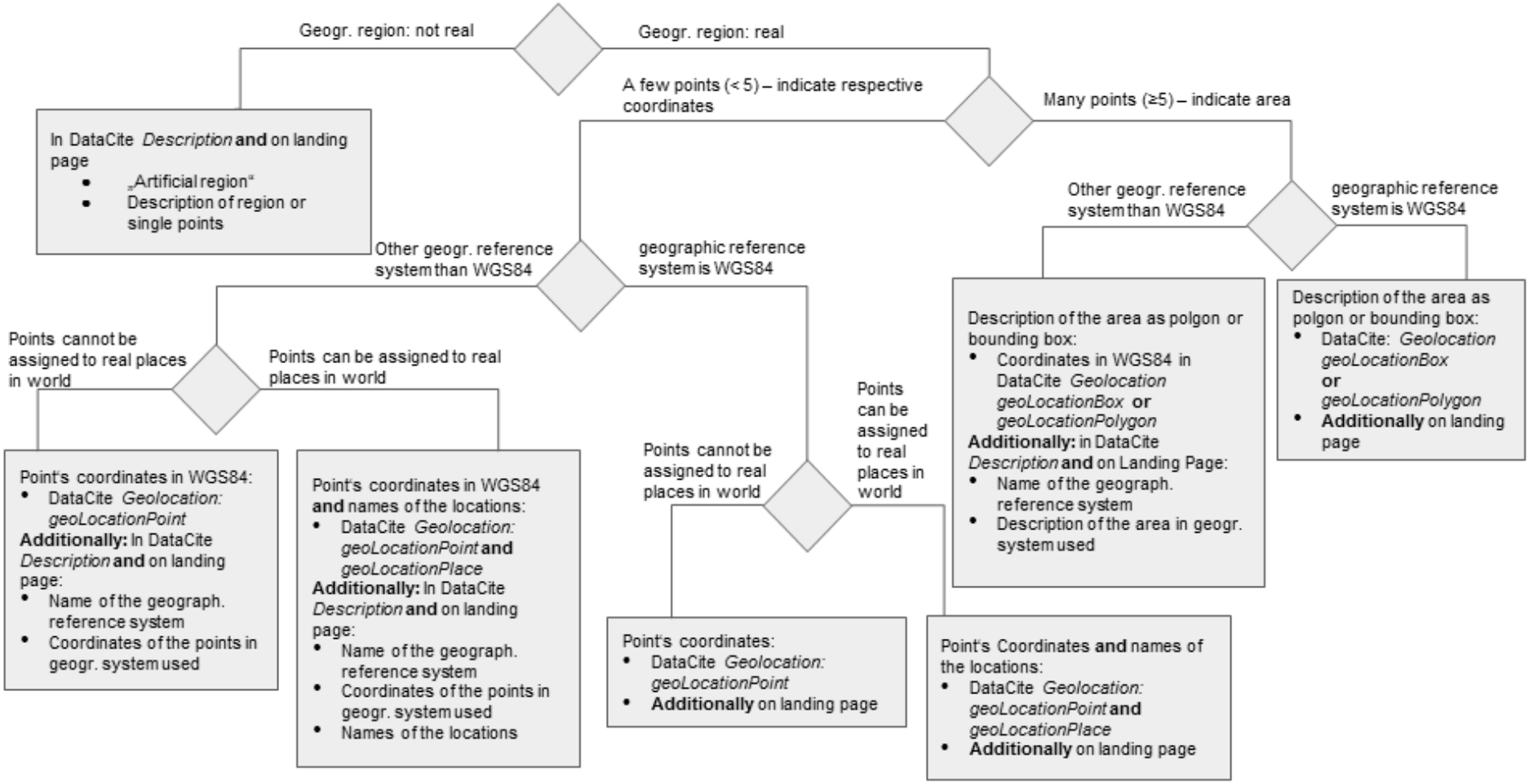

Figure 3: Flow charts for the metadata, describing the spatial properties of a dataset. Names of DataCite properties are written in *italic*.

J. File Format Tables

Table 11: Detailed table: mandatory (*status:* M), recommended (R), optional (O) und special (S, see footnote) global attributes for netCDF files complying with the ATMODAT standard. Table **??** is a short version of this table.

Attribute	**type**	**content**	**status**	**vocab.**	**example**	**Description**
comment	string	free form	O	CF-1.8		Miscellaneous information about the data, not captured elsewhere. This attribute is defined in the CF Conventions.
contact	string	free form	R	CMIP5	"John Smith, ABC Institute, john.smith@abc-institute.com"; "Scientific Data Contact Point at ABC Institute, data@abc-institute.com"	A contact person and contact details. Does not need to be the creator. Reasonable to provide contact details that are valid in long term.
Conventions	string	space separated list, containing "ATMODAT-Ver" (R) and "CF-Ver" (M).	M	CF-1.8	"CF-1.8 ATMODAT-2.4"	A space-separated list (or comma-separated if and only if a used conventions contains an empty space) of the conventions that are followed by the dataset (see https://www.unidata.ucar.edu/software/netcdf/conventions.html). Files following the ATMODAT Standard have to include "AtMoDat-VERSION" in the "Conventions" attribute. "VERSION" is the version of the ATMODAT Standard that was applied (e.g. "AtMoDat-2.4"). Files that follow the CF Conventions (what they must if they follow the AtMoDat Core Standard), have to include "CF-VERSION". Where "VERSION" is the applied version of CF Conventions (e.g. "CF-1.8"). For files that follow the ACDD (optional), the string 'ACDD-VERSION' has to be included.
creation_date	string	ISO 8601 time stamp	R	CMIP6	"2020-06-19T10:00:00", "2020-05-01"	Creation date of this netCDF file. Should be a time stamp conform with ISO 8601 (YYYY-MM-DD[THH:MM:SS])
creator	string	free form	R		"Max Smith, ABC Institute, orcid: ..., url: www.abc-institute.com/max_smith"	Person, who created the dataset. Does not need to be the long term contact person. The full name and ORCID of the creator should be included.
crs (coordinate reference system)	string	free form	R		"WGS84", "spherical Earth with radius of 6370 m"	coordinate reference system
featureType	string	CV (CF Conventions)	O[a]	CF-1.8	"point", "timeSeriesProfile"	Specifies the type of discrete sampling geometry to which the data in the file belongs, and implies that all data variables in the file contain collections of features of that type. CV: http://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.html#discrete-sampling-geometries
frequency	string	CV (extended from CMIP6)	R	CMIP6	"1hr", "1hrPt", "120s", "120sPt"	sampling frequency; note: "value + unit" denotes a mean over the provided time period, appending "C" denotes a climatology over the time period, appending "Pt" means a point value within the period; see for details: https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_frequency.json
further_info_url	string	free form	O	CMIP6	0	location of documentation
geospatial_lat_resolution	string	number + unit (CV for unit) (exception for e.g. 51° 14 '4,2 ")	R	ACDD-1.3	"100 m", "7 km", "10.3 degree"	information about the targeted spacing of points in latitude
geospatial_lon_resolution	string	number + unit (CV for unit) exception for e.g. 51° 14 '4,2 ")	R	ACDD-1.3	"100 m", "7 km", "10.3 degree"	information about the targeted spacing of points in longitude
geospatial_vertical_resolution	string	number + unit (CV for unit)	R	ACDD-1.3	"25 m"	information about the targeted vertical spacing of points
...	...	...	...	...	...	

Attribute	type	content	status	vocab.	example	Description
...	...	...	...	...	...	...
history	string	free form	R	CF-1.	0	Provides an audit trail for modifications to the original data. This attribute is also in the NetCDF Users Guide: 'This is a character array with a line for each invocation of a program that has modified the dataset. Well-behaved generic netCDF applications should append a line containing: date, time of day, user name, program name and command arguments.' To include a more complete description you can append a reference to an ISO Lineage entity; see NOAA EDM ISO Lineage guidance.
institution	string	free form	M	CF-1.	"XYZ University, ABC Institute"	The name of the institution principally responsible for originating this data. This attribute is recommended by the CF convention.
institution_id	string	free form	R	CMIP6	"XYZ, ABC" ,"MPI-M", "HZG", "UHH, MeMi", "Ulei, GlobCli"	institution identifier
keywords	string	free form	R	ACDD-1.3	"EIONET: Nitrogen dioxide", "EIONET: Precipitation", "GCMD: AIR QUALITY"	A comma-separated list of key words and/or phrases. Keywords may be common words or phrases, terms from a controlled vocabulary (GCMD is often used), or URIs for terms from a controlled vocabulary (see also "keywords_vocabulary" attribute).
keywords_vocabulary	string	free form	O	ACDD-1.3	"EIONET: Eionet Data Dictionary, GCMD: GCMD Keywords"	If you are using a controlled vocabulary for the words/phrases in your "keywords" attribute, this is the unique name or identifier of the vocabulary from which keywords are taken. If more than one keyword vocabulary is used, each may be presented with a prefix and a following comma, so that keywords may optionally be prefixed with the controlled vocabulary key.
license	string	free form	R	CMIP6	"CC-BY 4.0"	Provide the name of a commonly known license, provide an URL to a standard or specific license, or describe any restrictions to data access and distribution in free text.
metadata_link	string	free form	O	ACDD-1.3	0	A URL that gives the location of more complete metadata. A persistent URL is recommended for this attribute.
nominal_resolution	string	CV (CMIP6, extended)	R	CMIP6	"1 km", "0.4x1 km2", "0.1 degree", "1x0.5 degree"	approximate horizontal resolution. CV: `https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_nominal_resolution.json`
processing_level	string	free form	O	ACDD-1.3	"raw data", "level 2", "bias corrected"	A textual description of the processing (or quality control) level of the data.
product_version	any	free form	R	ACDD-1.3	"1.3"	Version identifier of the data file or product as assigned by the data creator. For example, a new algorithm or methodology could result in a new product_version.
program	string	free form	O	ACDD-1.3	"GHRSST", "NOAA CDR", "NASA EOS", "JPSS", "GOES-R".	The overarching program(s) of which the dataset is a part. A program consists of a set (or portfolio) of related and possibly interdependent projects that meet an overarching objective.
project	string	free form	O	ACDD-1.3	"Projekt ABC", "PATMOS-X", "Extended Continental Shelf Project", "PALMOD"	The name of the project(s) principally responsible for originating this data. Multiple projects can be separated by commas, as described under Attribute Content Guidelines.
realm	string	CV	R	CMIP6	"atmos", "aerosol", "ocean"	The Earth's realm (compartment) that is represented by the model. CV, see: `https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_realm.json`
...	...	...	...	...	...	

Attribute	type	content	status	vocab.	example	Description
...	...	...	...	...	...	...
references	string	free form	O	CF-1.8	"Juckes et al. (2020), doi: https://doi.org/10.5194/gmd-13-201-2020; Eyring et al. (2016), doi: https://doi.org/10.5194/gmd-9-1937-2016"	Published or web-based references that describe the data or methods used to produce it. Recommend URIs (such as a URL or DOI) for papers or other references. This attribute is defined in the CF conventions.
source	string	registered content	M	CF-1.8	"temperature from CTD #1234", "world model v0.1", "wind tunnel Uni Hamburg"	The method of production of the original data. If it was model-generated, source should name the model and its version. If it is observational, source should characterise it. This attribute is defined in the CF Conventions.
source_type	string	CV (CMIP6, extended)	R	CMIP6	"AGCM", "CHEM"	model configuration if applicable: see: https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_source_type.json
standard_name_vocabulary	string	free form	R	ACDD-1.3	"CF Standard Name Table v27"	The name and version of the controlled vocabulary from which variable standard names are taken. (Values for any standard_name attribute must come from the CF Standard Names vocabulary for the data file or product to comply with CF.).
summary	string	free form	R	ACDD-1.3	"This is a detailed summary of this file ... ."	A paragraph describing the dataset, analogous to an abstract for a paper.
title	string	free form	R	CF-1.8	"This is a great dataset"	A short phrase or sentence describing the dataset. In many discovery systems, the title will be displayed in the results list from a search, and therefore should be human readable and reasonable to display in a list of such names. This attribute is also recommended by the NetCDF Users Guide and the CF conventions.

[a] The global attribute *featureType* is mandatory when the data are of type *point*, *time series* (at one spatial location), *trajectory*, *profile*, *time series profile* or *trajectory profile*. The global attribute *featureType* must not be set if the data are spatially gridded. Chapter *Discrete Sampling Geometries* of the CF Conventions describes this in more detail (Appendix D).

K. Summary Metadata Tables

Table 12: Requirements for the DataCite Metadata

Comment	Requirements	Status
	All information about persons and institutions are complemented with a PID (see Appendix G).	R
	All links to documents, homepages etc. are provided as PIDs (see Appendix G).	R
	The languages of the dataset is specified according to ISO 639-1, e.g. "en" for English.	M
	All temporal information in the property date are given in a standardised way, according to ISO 8601 and ISO 19108.	M
	The subject contains EASYDAB, AtMoDat, the field of science and the realm of the model. The terms are chosen from controlled vocabularies (CVs, see Appendix F). Additional subjects are taken from CVs (see Appendix F).	M R
	Usage rights/licence must be open and always be described in English language.	M
The following attributes are submitted to DataCite.	Identifier (with mandatory type sub-property)	M
	Creator(s) Creator(s) sub-properties (family name, given name, name identifier and affiliation sub-properties)	M R
	Title (with optional type sub-properties)	M
	Publisher	M
	PublicationYear	M
	Subject Subject sub-properties	M R
	Contributor(s) Contributor(s) sub-properties (given name, family name, name identifier and affiliation sub-properties)	M R
	Date with at least one sub-property from created or updated Date with other date-types	M R
	Language	M
	ResourceType (with mandatory general type description sub-property)	M
	AlternateIdentifier (with type sub-property)	R
	RelatedIdentifier (with appropriate sub-properties)	R
	Size	R
	File Format	M
	Version	R
	Rights	M
	Description	M
	GeoLocation (with point, box and polygon sub-properties)	R
	Funding Reference (with name, identifier, and award related sub-properties)	R
	temporal information of the simulation, e.g. simulation period	R
Elements are included in Description	Abstract	M
	simulation time information	R
	Calender used	R
	Grid	R
	Model	M
	Model version	R
	Horizontal Resolution	R
	Geographic Reference System	R
	Vertical Coordinate	R
	Spatial Coverage	R
	Basic Approximations	R
	Possible usage of the data	R

Table 13: Requirements for the Landing Page

Comment	Requirements	Status
	The landing page contains human- and machine-readable metadata.	M
	The landing page contains a complete citation of the dataset in human-readable format and the DOI itself.	M
	The DOI is stored in the machine-readable part of the landing page.	R
	The landing page contains information on how to access the data, as long as the dataset is available.	M
	All metadata fields that are in the metadata record of the DataCite DOI are listed.	M
	The machine-readable part of the landing page is compliant to schema.org or an equivalent structure (i.e. based on W3C DCAT).	M
	If the DOI was assigned to a dataset collection with several datasets or to datasets with several variables: The top level of the landing page exposes the metadata of the low level granularity and it contains a list of all available files/variables.	R
Additional metadata fields which are not included in the DataCite ones.	Model	M
	Model version	R
	Grid	R
	Projection	R
	Vertical Coordinate	R
	Temporal Coverage	R
	Spatial Coverage	R
	Basic Approximations	R
If DOI is assigned to a Dataset collection these metadata fields are dispalyed on the landing page.	Variable/ Dataset Name	R
	Temporal Aggregation	R
	Spatial Aggregation	R
	Dimension	R
	Size	R

Table 14: Requirements for the Data Files

Comment	Requirements	Status
	The file format is netCDF.	M
	The value of the Conventions global attribute includes the version number of the used CF convention in the form "CF-Ver".	M
	The value of the Conventions global attribute includes the version number of the used ATMODAT Standard in the form "ATMODAT-Ver".	R
The following global attriubutes are included in the netCDF header and are strings.	comment	O
	contact	R
	Conventions	M
	creation_date	R
	creator	R
	crs (coordinate reference system)	R
	featureType	S
	frequency	R
	further_info_url	O
	geospatial_lat_resolution	R
	geospatial_lon_resolution	R
	geospatial_vertical_resolution	R
	history	R
	institution	M
	institution_id	R
	keywords	R
	keywords_vocabulary	O
	license	R
	metadata_link	O
	nominal_resolution	R
	processing_level	O
	program	O
	project	O
	realm	R
	references	O
	source	M
	source_type	R
	standard_name_vocabulary	R
	summary	R
	title	R
taken from a CV	featureType	S
	frequency	R
	nominal_resolution	R
	realm	R
	source_type	R
	The global attribute geospatial_lat_resolution is in the form: number + unit.	R
	The global geospatial_lon_resolution is in the form: number + unit.	R
	The global geospatial_vertical_resolution is in the form: number + unit.	R
	The global attribute product_version is included in the netCDF header.	R
	CF version 1.4 or higher are applied.	M
	netCDF file headers includes a time axis if the data is not static in time.	M
	netCDF file headers include a vertical axis if the data has reasonable vertical information.	M
	netCDF file headers include coordinate axes if the data is horizontally resolved.	M
	The global attribute Conventions is a space separated list.	M
	The creation_date is a ISO 8601 time stamp.	R

Table 15: Information, that has to be provided by the data producer. Explanations and examples can be found in Table 3.

Comment	Requirements	Status
	All information about persons and institutions are complemented with a PID (see Appendix G).	R
	All links to documents, homepages etc. are provided as PIDs (see Appendix G)	R
	The subject contains at least EASYDAB, AtMoDat, the field of science and the realm of the model. The terms are chosen from controlled vocabularies (CVs, see Appendix F). Additional subjects are taken from CVs (see Appendix F).	M R
	Usage rights/licence must be open and always be described in English language.	M
The following informations are needed for DataCite DOI metadata.	Creator(s) Creator(s) family name, given name, name identifier and affiliation)	M R
	Contributor(s) Contributor(s) given name, family name, name identifier and affiliation)	M R
	Title	M
	Subject/Keywords	M
	Description of subject vocabulary	R
	Description	M
	Alternate Identifier (if applicable) (with type description)	R
	Metadata Language	M
	Either date of creation or update of the dataset/dataset collection	M
	Additional dates related to publication of dataset/dataset collection	R
	End of Embargo (date)	R
	Temporal information of the simulation, e.g. simulation period	R
	Grid Specifications	R
	Documentation of the model, that was used for the simulations	R
	Boundary Conditions, that were used for the simulation	R
	Location of the model area (either point, box or polygon)	R
	Validation	R
	Maturity Checks	R
	Dataset was cited in	R
	Simulation is part of	R
	Comparable Datasets	R
	Size of the dataset/dataset collection	R
	File Format	M
	Version of the dataset/dataset collection	R
	Rights	M
	Funding Reference (with name, identifier, and award related sub-properties)	R
Elements are included in Description	Abstract	M
	Simulation time information	R
	Calender used	R
	Grid	R
	Model	M
	Model Version	R
	Horizontal Resolution	R
	Geographic Reference System	R
	Vertical Coordinate	R
	Spatial Coverage	R
	Basic Approximations	R
	Boundary Conditions	R
	Possible usage of the data	R

L. Example for DOI Metadata as JSON


```
{
   "data":{
      "id":"10.1594/wdcc/cmaq_cclm_hzg_2008",
      "type":"dois",
      "attributes":{
         "doi":"10.1594/wdcc/cmaq_cclm_hzg_2008",
         "prefix":"10.1594",
         "suffix":"wdcc/cmaq_cclm_hzg_2008",
         "identifiers":[
            {
               "identifier":"https://doi.org/10.1594/wdcc/
                  cmaq_cclm_hzg_2008",
               "identifierType":"DOI"
            }
         ],
         "alternateIdentifiers":[

         ],
         "creators":[
            {
               "name":"Neumann, Daniel",
               "nameType":"Personal",
               "givenName":"Daniel",
               "familyName":"Neumann",
               "nameIdentifiers":{
                  "nameIdentifier":"https://orcid.org/0000-0001-8574-
                     9093",
                  "nameIdentifierScheme":"ORCID",
                  "schemeUri":"https://orcid.org"
               },
               "affiliation":{
                  "name":"Leibniz-Institut fuer Ostseeforschung
                     Warnemuende (IOW)",
                  "affiliationIdentifier":"https://ror.org/03xh9nq73"
                     ,
                  "affiliationIdentifierScheme":"ROR",
                  "SchemeURI":"https://ROR.org"
               }
            },
            {
               "name":"Matthias, Volker",
               "...":"..."
            },
            {
               "name":"Bieser, Johannes",
               "...":"..."
            },
            {
               "name":"Aulinger, Armin",
               "...":"..."
```

```
            }
        ],
        "titles":[
            {
                "title":"Concentrations of gaseous pollutants and
                    particulate compounds over Northwestern Europe and
                    nitrogen deposition into the North and Baltic Sea
                    in 2008"
            }
        ],
        "publisher":"World Data Center for Climate (WDCC) at DKRZ",
        "container":{

        },
        "publicationYear":2017,
        "subjects":[
            {
                "subject":"Climate"
            },
            {
                "subject":"aerosol"
            },
            {
                "subject":"air pollution"
            },
            {
                "subject":"air quality"
            },
            {
                "subject":"atmospheric chemistry"
            },
            {
                "subject":"chemistry transport modelling"
            },
            {
                "subject":"coastDat-2"
            },
            {
                "subject":"coastdat"
            },
            {
                "subject":"sea salt"
            }
        ],
        "contributors":[
            {
                "name":"Matthias, Volker",
                "nameType":"Personal",
                "givenName":"Volker",
                "familyName":"Matthias",
                "nameIdentifiers":{
```

```
                "nameIdentifier":"https://orcid.org/0000-0003-0519-8805",
                "nameIdentifierScheme":"ORCID",
                "schemeUri":"https://orcid.org"
            },
            "affiliation":{
                "name":"Helmholtz-Zentrum Geesthacht, Zentrum fuer Material- und Kuestenforschung GmbH (HZG)",
                "affiliationIdentifier":"https://ror.org/03qjp1d79",
                "affiliationIdentifierScheme":"ROR",
                "SchemeURI":"https://ROR.org"
            },
            "contributorType":"ContactPerson"
        }
    ],
    "dates":[
        {
            "date":"2017-06-08",
            "dateType":"Created"
        },
        {
            "date":"2017",
            "dateType":"Issued"
        },
        {
            "date":"20080101/20081231",
             "dateType":"valid"
        }
    ],
    "language":"en",
    "types":{
        "ris":"DATA",
        "bibtex":"misc",
        "citeproc":"dataset",
        "schemaOrg":"Dataset",
        "resourceType":"Digital",
        "resourceTypeGeneral":"Dataset"
    },
    "relatedIdentifiers":[
        {
            "relationType":"IsDocumentedBy",
            "relatedIdentifierType":"DOI",
            "DOI":"http://doi.org/10.1029/2001JD001409"
        },
        {
            "relationType":"isDerivedFrom",
            "relatedIdentifierType":"DOI",
            "DOI":"http://doi.org/10.5194/gmd-4-47-2011"
        },
        {
            "relationType":"isReviewedBy",
```

```
                "relatedIdentifierType":"URL",
                "URL":"https://cera-www.dkrz.de/WDCC/ui/cerasearch/
                   entry?acronym=quality_checking_report"
             },
             {
                "relationType":"References",
                "relatedIdentifierType":"DOI",
                "DOI":"http://doi.org/10.5194/acp-16-739-2016"
             }
          ],
          "sizes":[
             "12459482227 Bytes"
          ],
          "formats":[
             "application/x-netcdf"
          ],
          "version":"1",
          "rightsList":[
             {
                "rights":"Creative Commons Attribution 4.0
                   International (CC BY 4.0)"
             }
          ],
          "descriptions":[
             {
                "description":"Project: coastDat-2 - CoastDat-2 (http:
                   //www.coastdat.de/about_us/index.php) comprises a
                   compilation of coastal analyses (that is hindcasts
                   and reconstructions) and scenarios for the future
                   obtained from numerical models. The atmospheric
                   part of the dataset is simulated using the COSMO-
                   CLM (http://www.clm-community.eu). The objective is
                    to provide a consistent meteo-marine dataset.
                   Summary: The concentrations and deposition of
                   atmospheric constituents such as air pollutants
                   were modelled with the Community Multiscale Air
                   Quality (CMAQ) Model system for the year 2008.....
                   ",
                "descriptionType":"Abstract"
             }
          ],
          "geoLocations":[
             {
                "geoLocationBox":{
                   "eastBoundLongitude":"36.9",
                   "northBoundLatitude":"67",
                   "southBoundLatitude":"42.6",
                   "westBoundLongitude":"18.9"
                },
                "geoLocationPlace":"Northwestern Europe"
             }
          ],
```

```
        "fundingReferences":[

        ],
        "url":"http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=
           CMAQ_CCLM_HZG_2008",
        "contentUrl":null,
        "metadataVersion":1,
        "schemaVersion":null,
        "source":"mds",
        "isActive":true,
        "state":"findable",
        "reason":null,
        "viewCount":0,
        "viewsOverTime":[

        ],
        "downloadCount":0,
        "downloadsOverTime":[

        ],
        "referenceCount":0,
        "citationCount":0,
        "citationsOverTime":[

        ],
        "partCount":0,
        "partOfCount":0,
        "versionCount":0,
        "versionOfCount":0,
        "created":"2017-12-08T11:03:54.000Z",
        "registered":"2017-12-08T11:04:15.000Z",
        "published":"2017",
        "updated":"2020-06-08T10:41:49.000Z"
     },
     "relationships":{
        "client":{
           "data":{
              "id":"dkrz.wdcc",
              "type":"clients"
           }
        },
        "media":{
           "data":{
              "id":"10.1594/wdcc/cmaq_cclm_hzg_2008",
              "type":"media"
           }
        },
        "references":{
           "data":[

           ]
        },
```

```
            "citations":{
               "data":[

               ]
            },
            "parts":{
               "data":[

               ]
            },
            "partOf":{
               "data":[

               ]
            },
            "versions":{
               "data":[

               ]
            },
            "versionOf":{
               "data":[

               ]
            }
         }
      }
}
```