Atmospheric **Mo**del **Dat**a
details at www.atmodat.de

v1.00

AtMoDat

AtMoDat: Improving the reusability of ATmospheric MOdel DATa with DataCite DOIs paving the path towards FAIR data

D. Neumann[1], A. Ganske[2], V. Voss[3], A. Kraft[2], H. Höck[1],
K. Peters[1], J. Quaas[4], H. Schlünzen[3] and H. Thiemann[1]

DKRZ

TIB

U·H
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

UNIVERSITÄT
LEIPZIG

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

[1]   [2]   [3]   [4]

Usage hints / Readme

This is an interactive presentation with several links to other slides within this document and to some external web pages.

- The **finger** indicates links **within** this **documents**

- This is an **example link**

click onto this box to go to a training page in the Supplement (there will a a link to go back)

- external links to internet resources are clearly indicated
  www.atmodat.de

click there to go back to first slide

top

Table of Contents

Usage hints

Motivation

Tasks in AtMoDat

Core Standard

Data Maturity Indicator

discipline-specific standardization

Summary

The End

Supplement

- data publication numbers (incl. DataCite DOI's) are increasing, however…

- data reusability often difficult
  - ○ wide variety of file formats
  - ○ incomplete metadata and/or no discipline-specific metadata
  - ○ insufficient quality information

Motivation for AtMoDat: improve the reusability of **at**mospheric **mo**del **dat**a

- affects entire scientific community
  - ○ data users: appropriate data difficult to find and to reuse
  - ○ data producers and repositories: less reuse of data, reduced publicity
  - ○ funding institutions: project outcomes not reusable
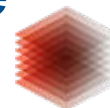
top

Tasks in AtMoDat

practice partners

infrastructure providers

Core Standard for improving reusability of atmospheric model data

Data Maturity Indicator for DataCite DOIs

feasible **discipline-specific standards** for **model output** of small sub-disciplines
- urban atmospheric models
- small MIPs (cloud modeling)

top

[1] https://www.atmodat.de/p/quality_discipline/
[2] https://www.atmodat.de/p/quality_generic/

Core Standard

AtMoDat

requirements for **DOI metadata** and for **metadata** on human- and machine-readable **landing pages** of datasets

file-level standardization based on the CMIP6 standard

Project report [German only]
https://www.atmodat.de/p/kernstandard/

Publication (Ganske et al., in prep.)

next slide

tasks

Core Standard: file-level standardization

back

file-level standardization based on **CMIP6 standard**

netCDF files

compliance with **CF Conventions**

controlled vocabularies (CVs) for **global attributes**

CVs at GitHub:
https://github.com/AtMoDat/AtMoDat_CVs

checker plugin at GitHub:
https://github.com/AtMoDat/cc-plugin-cmip6-cv (Test for CMIP)

plugin for a **netCDF checker** for checking these **CVs**

tasks
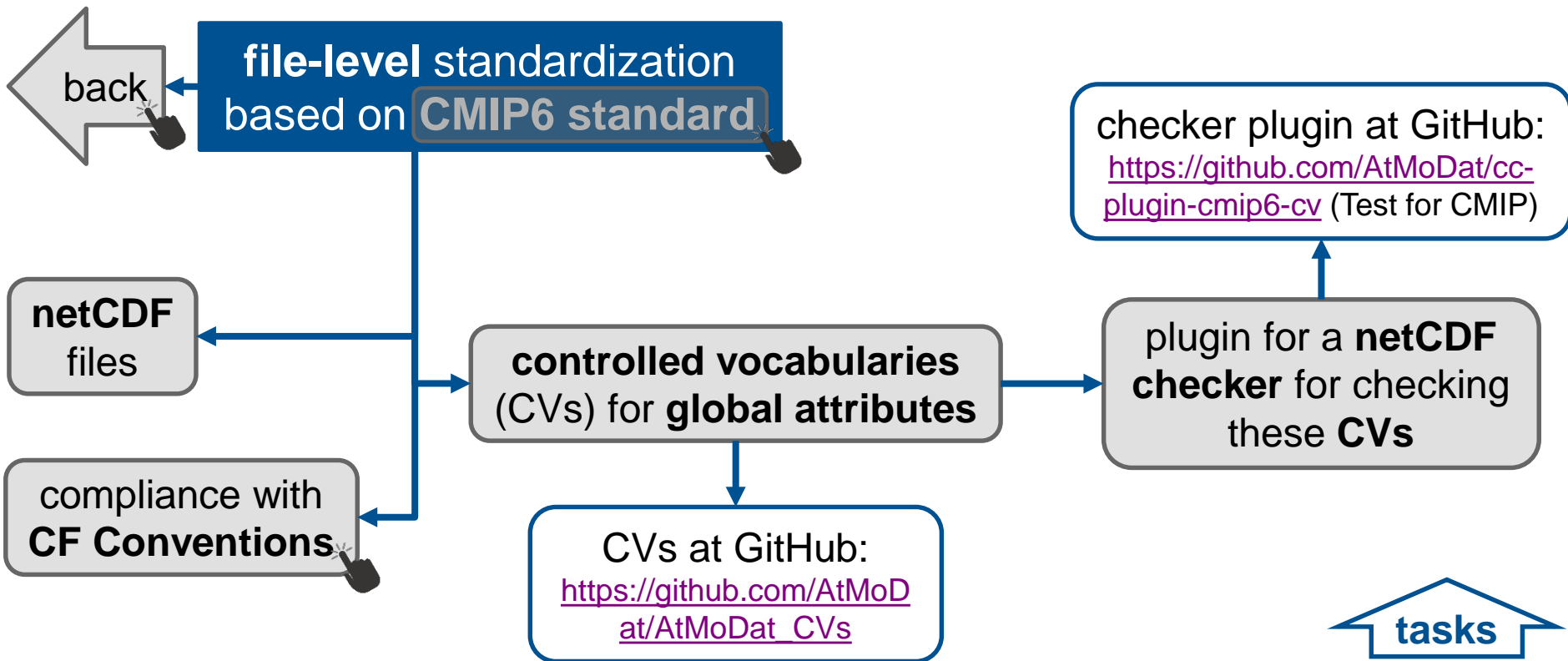
Data Maturity Indicator – **DRAFT**

work in progress

AtMoDat

General idea

add a property to the **DataCite Metadata Schema** to host data maturity information:

DataCite Metadata:
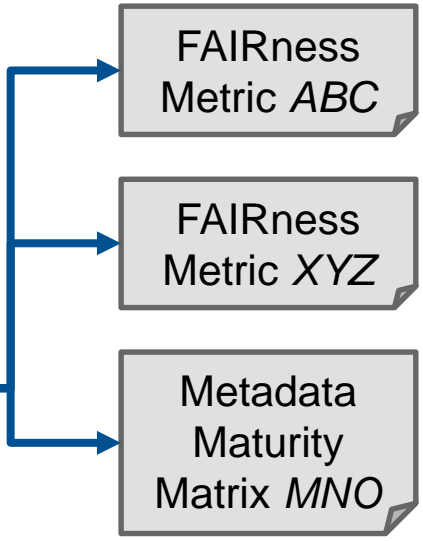
title: …
description: …
author: …
…
dataMaturityIndicator: ▪

…

<dataMaturityIndicator>
- name, version, …
- **external maturity metric** ▪ (linked via PID)
- **result report** linked via PID
- **machine readable results** of maturity check

FAIRness Metric *ABC*

FAIRness Metric *XYZ*

Metadata Maturity Matrix *MNO*

Data Maturity Indicator **draft**:
https://github.com/AtMoDat/data-maturity-indicator

tasks

discipline-specific standardization

work in progress
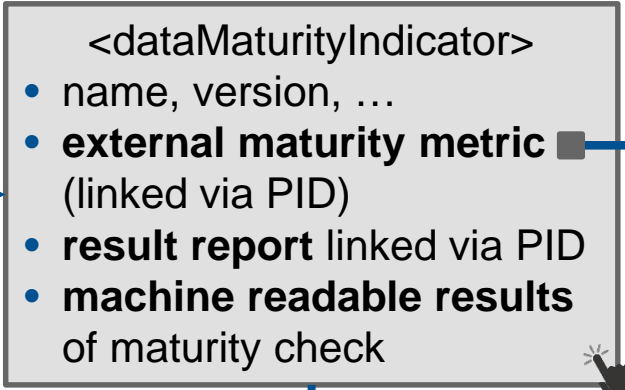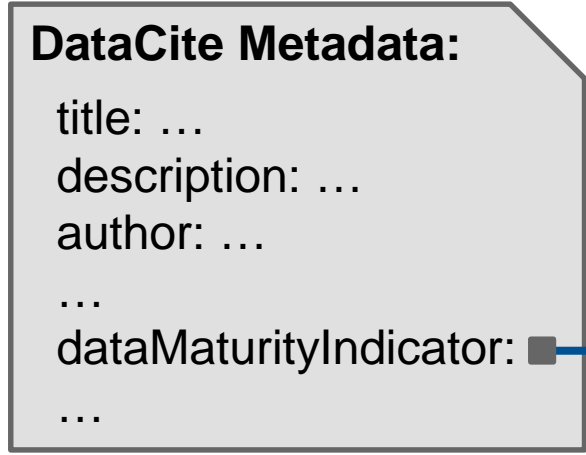
AtMoDat

microscale atmospheric modeling
(urban scale / obstacle resolving)

issues / questions

specific input/output data not present in global and regional atmospheric models:

- buildings, trees and similar ▪

- variables at building surface: temperature, radiation ▪

- highly spatially resolved terrain data and model grids ▪

Should we and how to implement these objects in netCDF files?

CF Standard Names not existing => many new name proposals

Proper definition of projections and spatial reference system important

survey on best practices and standardization in microscale modeling

https://uhh.de/orm-survey

tasks

see also EGU2020-2822, Voss et al. (2020), doi: 10.5194/egusphere-egu2020-2822

9

v1.00

current situation: issues with respect to reusability of scientific data

development of **concepts to raise reusability**

- draft proposal to **store data maturity information** in DataCite metadata
- **netCDF** file header **standardization** between **CF Conventions** and **CMIP6 standard**
- **discipline-specific extensions to CF Conventions**
- recommendations for **metadata** on **human- and machine-readable** landing pages

aim: improve reusability ($\rightarrow$ **FAIRness**) of archived atmospheric model data of small atmospheric science sub-disciplines

top

The end

Acknowledgements

We thank our **colleagues** for answering questions and providing support on individual topics of the AtMoDat project.

AtMoDat is funded by the **German Federal Ministry for Education and Research (BMBF)** within the framework of "*Forschungsvorhaben zur Entwicklung und Erprobung von Kurationskriterien und Qualitätsstandards von Forschungsdaten*" (FKZ: 16QK02A).

Questions?

Contact us:
info@atmodat.de

Visit us:
https://www.atmodat.de

Visit us on GitHub:
https://github.com/AtMoDat

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

The following slides contain supplementary information to the "main" slides. The supplementary slides might not be fully understandable when viewed individually without context of the corresponding "main" slide.

SUPPLEMENT

top

Authors

Daniel Neumann
German Climate Computing Center (DKRZ)
https://orcid.org/0000-0001-8574-9093
contact: daniel.neumann@dkrz.de

Anette Ganske
Technische Informationsbibliothek (TIB)
https://orcid.org/0000-0003-1043-4964

Vivien Voss
University of Hamburg, Meteorological Institute
https://orcid.org/0000-0001-7329-8192

Angelina Kraft
Technische Informationsbibliothek (TIB)
https://orcid.org/0000-0002-6454-335X

Heinke Höck
German Climate Computing Center (DKRZ)
https://orcid.org/0000-0002-0131-1404

Karsten Peters
German Climate Computing Center (DKRZ)
https://orcid.org/0000-0003-0158-2957

Johannes Quaas
University of Leipzig, Leipzig Institute for Meteorology
https://orcid.org/0000-0001-7057-194X

K. Heinke Schlünzen
University of Hamburg, Meteorological Institute
https://orcid.org/0000-0002-5711-547X

Hannes Thiemann
German Climate Computing Center (DKRZ)
https://orcid.org/0000-0002-2329-8511

top

Supplementary usage hints – Training Page

- This is a training page to learn/train navigating in this presentation. Apart from this purpose, this slide contains no valuable information.

- This is some dummy text in order to have some content available for the reader.

- Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut …

click there to go back to first slide

click there to go back the slide „Usage hints / Readme"

usage hints

top
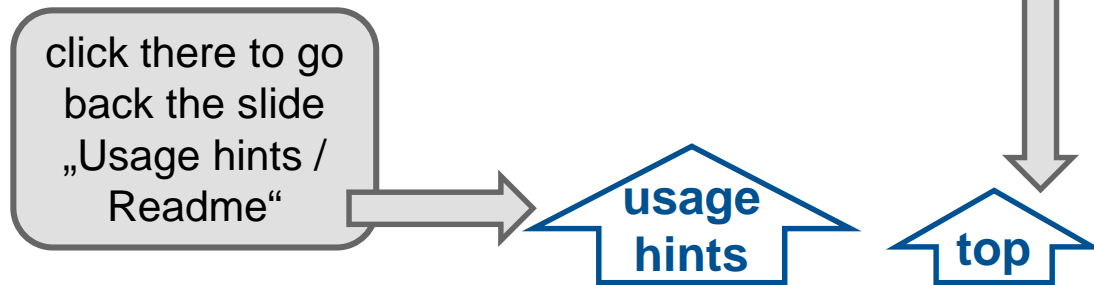
AtMoDat

The project

The project acronym AtMoDat abbreviates "*Atmosphären-Modelldaten: Datenqualität, Kurationskriterien und DOI-Branding*" ("*Atmospheric Model Data: data quality, curation criteria and DOI branding*").

AtMoDat started in June 2019 and is funded for three years by the German Federal Ministry for Education and Research (BMBF) within the framework of "*Forschungsvorhaben zur Entwicklung und Erprobung von Kurationskriterien und Qualitätsstandards von Forschungsdaten*".

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Aim

Archived research data should be Findable, Accessible, Interoperable, and Reusable (FAIR). Quality assured data and transparent curation criteria are important to ensure effective partial and subsequent reusability of research data by other research communities and public authorities. See slide "**Tasks**" for details on the actual work done in AtMoDat.

Web links

AtMoDat Homepage: https://www.atmodat.de/

Details on AtMoDat: https://www.atmodat.de/about-atmodat

AtMoDat on GitHub: https://github.com/AtMoDat

Details on the BMBF funding framework on Curation Criteria and Quality Standards, German only: https://www.bmbf.de/foerderungen/bekanntmachung-1791.html

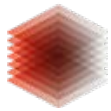tasks   top

German Climate Computing Center (DKRZ)

Hosting the World Data Center for Climate (WDCC), the DKRZ provides a certified long-term archive (LTA) for scientific data. One key aspect is the support of the "FAIR Data" principles for scientific data management and stewardship (FAIR: findable, accessible, interoperable, re-useable). Currently, the DKRZ is focused on archiving global climate modeling data, such as CMIP data.

In AtMoDat, the DKRZ plans to extend its services towards further modeling communities, i.e. small model inter-comparison projects (MIPs) and urban climate modeling. For this purpose, existing metadata standards are to be adapted in order to cover the needs of these communities and to enable data curation. Additionally, a field specific "atmospheric model data DOI" (AMD-DOI) concept including data quality information should be established and applied to example cases. The AMD-DOI concept allows the data user to quickly identify high quality data sets. Quality controlled data characterized with an AMD-DOI leads to an increased re-use potential of the archived data. The curation processes developed and introduced in the course of the project will also lead to an improvement of processes already established at DKRZ.

DKRZ's homepage: https://www.dkrz.de/

WDCC at DKRZ: https://cera-www.dkrz.de/

tasks   top

TIB – Leibniz Information Centre for Science and Technology

As a pioneer in the field of research data management, the DKRZ and the TIB coordinated the registration process for the Digital Object Identifier (DOI) as early as 2005 in a worldwide cooperation. In 2009 the international consortium DataCite e.V. was founded, whose office is located at the TIB. Today, the TIB currently supports more than 160 data centres that register DataCite DOIs for digital objects.

The AtMoDat work plan for the TIB is divided into two scopes. Firstly, the DataCite DOI metadata scheme will be extended and adapted for a quality assured 'Atmospheric Model Data DOI (AMD DOI) Branding' for the data sets of urban climate research and for small meteorological model comparison studies. Suitable metadata parameters are selected and described and the technical development is coordinated with DataCite. With the establishment of the AMD DOI, AtMoDat will introduce a discipline-shaped persistent identifier service to the scientific modelling communities. This service will provide all functionalities of the well-known DataCite DOIs, and at the same time provide a high level of quality control and therefore trustworthiness for the so published data.

Furthermore, the development of a technical vocabulary based on DCAT is planned in order to optimize the machine reusability of the data sets according to the FAIR Data Principles. For the (meta)data quality standards developed, the curation processes will then be adapted for operationalization so that the desired AMD-DOI branding can be incorporated into the workflows and business processes of the TIB's DOI Services.

TIB's homepage: https://www.tib.eu/

tasks   top

University of Hamburg, Meteorological Institute

The Meteorological Institute of the University of Hamburg (MI UHH) develops local and global scale meteorological models, operates one of largest and most modern boundary layer wind tunnels in Europe, and offers one of the few meteorology degree programs in Germany. The working groups "Mesoscale and Microscale Modelling" and "Environmental Wind Tunnel Laboratory" of the MI participate in AtMoDat. They develop urban climate models and do practical research in the wind tunnel, respectively. Both groups are very interested in advancing the standardization of (meta-)data in their fields of research and contribute their input as practice partners to AtMoDat.

Homepage: https://www.mi.uni-hamburg.de/arbeitsgruppen/memi.html

The research group Clouds and Global Climate of Institute for Meteorology of the University of Leipzig investigates clouds, aerosols, radiation and precipitation via modeling on regional and global scale.

Scientific progress in the field of atmospheric modeling is often achieved through model intercomparison projects (MIPs). In MIPs, the same situations are simulated with different atmospheric or climate models. Initial and boundary conditions and model setups are precisely prescribed. The diversity that nevertheless occurs allows systematic and statistical errors to be identified and possibly traced back to their causes. In general, these MIPs also aim at specific comparisons with observation data. For large MIPs, the model output data is also well defined in terms of format, provision and archiving.

The research group Clouds and Global Climate is involved in several smaller "grassroots" MIPs, including two projects in which scientists from Leipzig take leading roles: AEROCOM and ACPC. Within the "AEROCOM" project, Leipzig leads is a comparative study on warm rain, i.e. rain formed from liquid water clouds. New and not yet standardized model diagnostics have to be implemented and evaluated. On the one hand, standardization is necessary to achieve comparability. On the other hand, it cannot be too rigid, since changes may occur in the course of evaluation and of further scientific development. In the initiative Aerosols, Clouds, Precipitation and Climate (ACPC) the team of the University of Leipzig participates in the simulation of aerosol-cloud interactions with cloud system resolution models over limited model areas. Here, standardization is far less advanced than in global modelling, and it results in corresponding problems in standardizing and making data available.

Homepage: http://research.uni-leipzig.de/climate/

tasks   top

CMIP6 Standard

AtMoDat

Requirements of the CMIP6 Standard for model output

- file format: **netCDF**
- **one variable per file**
- compliance with **CF Conventions** **next slide**
- **30 global attributes**
- **controlled vocabularies** (CVs) for **attribute** values based on DRS
- data variables mapped on specific vertical and horizontal model grids
- Files stored in unambiguous folder hierarchy based on DRS

Summary of requirements: "CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV's": https://goo.gl/v1drZl

CMIP6 global attribute CVs:
https://github.com/WCRP-CMIP/CMIP6_CVs

GMD Special Issue on CMIP6 Experimental Design and Organization:
https://www.geosci-model-dev.net/special_issue590.html

Overview over CMIP6 by the World Climate Research Programme:
https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

Coupled Model Intercomparison Project Phase 6 (CMIP6):
CMIP6 provides an organizational structure for a specific set of climate simulations experiments. The simulations within CMIP6 form the basis for peer-reviewed publications on which the upcoming IPCC assessment report will be based on. Besides CMIP there are 23 endorsed MIPs under the umbrella of CMIP6, such as Ocean MIP (OMIP).

Data Reference Syntax (DRS):
The DRS consists of 17 "components" which allow to unambiguously identify each variable of each individual simulation performed within an endorsed MIP. Some components:
- variable_id,
- institution_id,
- realm (atmosphere, ocean, …),
- variant_label (identifies ensemble member),
- acitivity_id (CMIP, OMIP, …).

core standard
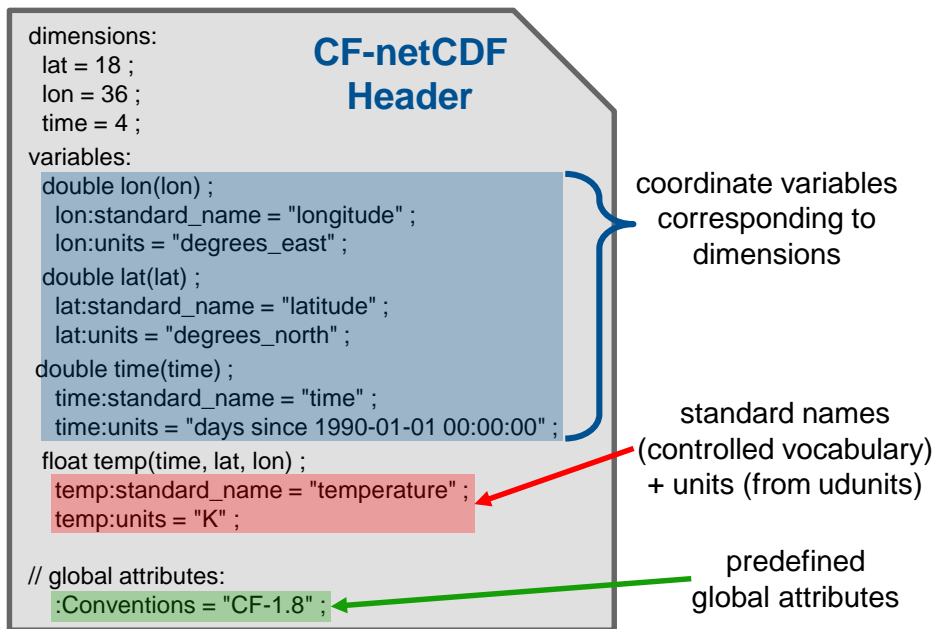
top

CF Metadata Conventions

AtMoDat

"*The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created with the NetCDF API. The CF conventions are increasingly gaining acceptance and have been adopted by a number of projects and groups as a primary standard. The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.*" (CF Conventions Homepage)

Homepage: http://cfconventions.org/

CF Standard Names:
http://cfconventions.org/standard-names.html

GitHub: https://github.com/cf-convention/cf-conventions

UDUNITS (unidata, not CF):
https://www.unidata.ucar.edu/software/udunits/

CF-netCDF Header

```
dimensions:
  lat = 18 ;
  lon = 36 ;
  time = 4 ;
variables:
  double lon(lon) ;
    lon:standard_name = "longitude" ;
    lon:units = "degrees_east" ;
  double lat(lat) ;
    lat:standard_name = "latitude" ;
    lat:units = "degrees_north" ;
  double time(time) ;
    time:standard_name = "time" ;
    time:units = "days since 1990-01-01 00:00:00" ;
  float temp(time, lat, lon) ;
    temp:standard_name = "temperature" ;
    temp:units = "K" ;

// global attributes:
  :Conventions = "CF-1.8" ;
```

coordinate variables corresponding to dimensions

standard names (controlled vocabulary) + units (from udunits)

predefined global attributes

core standard

top

Draft Data Maturity Indicator – **Draft** – Details

work in progress

AtMoDat

ID	DataCite-Property	Occ	Definition
99	DataMaturityCheck	0-n	
99.1	maturityCheckMetricName	1	name of the metric
99.2	maturityCheckMetricType	0-1	type of the metric document
99.3	maturityCheckMetricIsMachineReadable	0-1	is the metric machine-readable / can it be automatically evaluated
99.4	maturityCheckMetricIdentifier	0-1	PID to the metric definition
99.5	maturityCheckMetricVersion	1	version of the metric
99.6	maturityCheckPerformedBy	1-n	information on who performed the quality assurance
99.6.1	maturityCheckPerformedByName	1	name of person/organization
99.6.2	maturityCheckPerformedByIdentifier	1-n	ID of person/organization
99.7	maturityCheckResult	1-n	result of the evaluation
99.7.1	maturityCheckResultDate	1	date when the evaluation was performed
99.7.2	maturityCheckResultIdentifier	0-1	URI to a evaluation report or a similar document
99.7.3	maturityCheckResultIndicator	0-n	evaluation results of one indicator of the metric

The Data Maturity Indicator (DMI) is a draft extension to the DataCite Metadata Schema. The DMI would allow to place results of data maturity checks in the DataCite Metadata in a standardized format. Currently (DataCite Metadata Schema v4.3), such information can only be provided via relatedIdentifiers.

This concept does only provide metadata fields. No specific maturity matrix/metric/measure is proposed.

Details: https://github.com/AtMoDat/data-maturity-indicator

Data Maturity Indicator

top