A discipline-specific quality standard for atmospheric model data

Note: This document can be considered as the discipline-specific extension of the *generic quality indicator of DateCite DOIs*[1] proposed by the AtMoDat project.

Currently, we often face the issue that data sets are available in proprietary formats and lack discipline-specific standardization and metadata. This considerably hampers the reusability of these data sets and, thus, contradicts with the FAIR principles[2]. Within the AtMoDat project (Atmospheric Model Data), we aim to develop a **discipline-specific quality standard** for the publication of atmospheric model data which is closely aligned to the FAIR data principles. Besides making data more reusable, the work of data producers who provide high-quality data sets to the scientific community and beyond is recognized. As well, data repositories that perform high quality review and data curation and provide high quality data sets can be identified more easily. Finally, funding agencies have a strong interest that data whose production they have funded are reusable beyond the funding period in order to create added value in the long term.

This purely discipline-specific quality standard will be developed for atmospheric model data (e.g. named AtMoDatQ) and, later, can simply be transferred to other related disciplines such as ocean model data. **The standard consists of**

- an **extension of the list of mandatory DataCite metadata properties**,
- the requirement to show **all mandatory metadata** on the landing page (human- and machine-readable),
- **predefined keywords**, which can be supplemented with data-specific keywords,
- **specifications** for **open self-describing file formats** and file format conventions: files can be used with free software and are self-descriptive; e.g. netCDF, with a mandatory use of a standard (e.g. CF Conventions[3] with additional global attributes) and
- a **documentation of the quality control** of both data and metadata, executed by the data repository, which can be found via a PID (e.g. related identifier in metadata schema).

This discipline-specific quality standard can be **implemented without changing** the current **DataCite Metadata schema** 4.3[4] because the specifications of existing metadata properties are sufficiently flexible. However, this approach restricts the machine-readability of the metadata. Data users should be able to easily recognize that the data are standardized which could be realized e.g. by inclusion of a logo on the landing pages or a DOI branding.

Benefits: Firstly, high quality data will be clearly indicated. Secondly, data may be used and cited more often, if they are well documented and indexed. Such data will be valued higher than data without quality control. Data repositories stand out from other repositories by offering this ‘seal of quality’. Invested resources are utilized more efficiently by reusing existing data instead of generating new data.

[1] “Generic quality indicators in metadata of DataCite DOIs”, https://www.atmodat.de/p/quality_generic/
[2] Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi: https://doi.org/10.1038/sdata.2016.18
[3] Climate and Forecast (CF) Conventions, http://cfconventions.org/
[4] https://schema.datacite.org/