Generic quality indicators in metadata of DataCite DOIs

The generation of research data is both work extensive and costly. Data users are searching for standardized high-quality data. Funding agencies wish to encourage the reuse of data, whose production they have funded, beyond the original usage. To foster the reuse of research data, the FAIR[1] data principles were established: findable, accessible, interoperable and reusable. Assigning DataCite DOIs and metadata to datasets are the first steps to meet the FAIR principles. However, DOIs and metadata do not guarantee that data are actually reusable if quality information on the data and metadata are not provided and data are saved in proprietary or undocumented file formats. The AtMoDat project (Atmospheric Model Data) aims on establishing a generic quality indicator in the DataCite DOI metadata schema which

- supports data users to identify high quality research data in and beyond their field of expertise,
- incentivizes data producers to publish high quality data, and
- honors data repositories which perform extensive data curation.

Instead of a purely discipline-specific implementation of a quality indicator, a generic implementation within the DataCite schema makes sense, as we want to facilitate the reuse of research data by an interdisciplinary audience.

The core of this approach would be that published data records would receive a quality indicator - a new property either named '*Quality Assessment*' (Response: '*yes/no/n.a.*') or '*Quality level*' (increasing level with increasing number of fulfilled quality aspects). The *Quality levels* following the FAIR principles would have to be discussed, developed, communicated, and implemented by the different scientific communities. We are aware that this is a long-term process and must be driven by active community discussions and that the outcome will differ amongst different disciplines in their 'level of practicality'. However, given the high volumes of digital scientific data, such a quality indicator will help data producers, consumers, data centers, and even funders in the long run. The quality indicator should refer to the dataset, rather than to the repository or data center itself. That means that such a *Quality* indicator will differ from repository certificates such as Core Trust Seal[2] or DINI-Certificate[3]. These certificates usually include a time-period during which the certificate is valid. In contrast, the quality status proposed here is permanently assigned to the published dataset and has not to be assigned to each dataset of a particular repository.

Aspects described by this quality indicator should be

- the use of all applicable recommended and optional DataCite metadata,
- documentation by means of discipline-specific metadata,
- the use of open self-describing file formats which follows discipline-specific standards, and
- checks of metadata, file formats and discipline-specific standards for correctness and completeness by the publishing data repository

The approach described here would improve the labelling of quality-checked data and, thus, significantly contribute to their re-use leading further towards FAIR data.

We invite you to get in touch with us via info@atmodat.de.

[1] Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi: https://doi.org/10.1038/sdata.2016.18
[2] https://www.coretrustseal.org/
[3] https://dini.de/dienste-projekte/dini-zertifikat/