

Automatic assessment of metadata quality in ISO 191xx

Susanne Feistel, Leibniz Institute for Baltic Sea Research Warnemünde (Germany),
susanne.feistel@io-warnemuende.de

Ulrike Kleeberg, Helmholtz-Zentrum Geesthacht (Germany), ulrike.kleeberg@hzg.de

Jörn Kohlus, Nationalpark Wattenmeer, (Germany), joern.kohlus@lkn.landsh.de

Rainer Lehfeldt, Federal Waterways Engineering and Research Institute (Germany),
rainer.lehfeldt@baw.de

Carsten Schirnick, GEOMAR Helmholtz Centre for Ocean Research Kiel (Germany),
cschirnick@geomar.de

Susanne Tamm, Federal Maritime and Hydrographic Agency (Germany), susanne.tamm@bsh.de

Sharing geoinformation and data across communities is becoming more important, which is enabled by web services. The Open Data initiative (e.g. data.gov) promoted by public agencies and research institutions enhances this development. However, poor data quality with insufficient quality information may be hindering the acceptance and usage of the data by the scientific community in the future. The mandatory metadata within ISO 191xx is not sufficient for a comprehensive quality assessment or long-term usability of any dataset. Quality information is mostly optional and not easily found or filtered. It might be stored within the metadata as information in LI_Lineage, DQ_DataQuality or even within the dataset itself, possibly coded. Ultimately, in most cases, quality information is not directly accessible for interested users.

On the other side, for the data creator the documentation of quality information can get time consuming and overwhelming, especially if required all at once in the reporting phase of a project. Quality Flag Schemes have been proposed by multiple organisations. In UNESCO 2013, Manuals and Guides 54, the most commonly used are presented and compared. Mutual conversion rules between them are provided at http://odv.awi.de/fileadmin/user_upload/odv/misc/ODV4_QualityFlagSets.pdf. Furthermore, quality information gets significantly more complex if more than one dataset is considered as in data products such as maps or modelling results.

We present a more practical approach to get a sufficient and standardized quality assessment (quality flag), dynamically generated from the entire quality information of the metadata XML file. While the ISO standard offers a complex array of optional fields to provide quality information, we recommend a manageable number of fields to be filled. As prepared in (Feistel et al. 2018, in prep.) there are specific information a user needs first and foremost to evaluate the usefulness of a dataset.

Our task group provides a set of tools to assess quality information recorded in the recommended fields of the ISO standard. The first is a web-based form to manually assess a single dataset, the second is a program for syntactically analysing multiple dynamically generated XML in ISO 19139. To make the quality assessment compatible and comparable between platforms and schemes, as well as machine-friendly, we propose a coded summary string in the pattern of the scheme plus a flag, e.g. "SDN::1" to be put in the ISO field "DQ_StandaloneQualityInformation".

For data creators, metadata templates can be prepared in advance. This is especially useful for repeating data collection or creation procedures within projects, as well as long-term data series. Thus, the compilation of metadata can be simplified for individual scientists. The advantage of using an automatic assessment of metadata quality (ISO XML checker) is primarily in getting a standardized quality flag based on a common quality flag scheme. A time consuming individual control is not necessary. The quality control mechanism supports users in finding data relevant for their own work more efficiently.