# Automatically generating ISO 19115-1:2014 metadata from relational databases

**Linda Baldewein,** Helmholtz-Zentrum Geesthacht (Germany), linda.baldewein@hzg.de
**Dietmar Sauer**, Helmholtz-Zentrum Geesthacht (Germany), dietmar.sauer@hzg.de
**Ulrike Kleeberg**, Helmholtz-Zentrum Geesthacht (Germany), ulrike.kleeberg@hzg.de
**Housam Dibeh**, Helmholtz-Zentrum Geesthacht (Germany), housam.dibeh@hzg.de
**Lars Möller**, Helmholtz-Zentrum Geesthacht (Germany), lars.moeller@hzg.de

Long-term storage solutions for scientific data are currently multiplying at an enormous rate as more and more journals require Digital Object Identifiers for the data used in their publications. The treatment of metadata, however, is somewhat lagging behind. Each data publisher requires different metadata fields and only few of them follow international standards. At the same time a variety of metadata catalogue systems, such as GeoNetwork, Geoportal, etc. exist. These catalogues provide standard interfaces to access, search and harvest the imported metadata, but simple tools to automatically generate the import files are missing.

At the Helmholtz Coastal Data Center (HCDC) we developed a method to extract all relevant information for the ISO 19115-1:2014 standard from different relational databases and to automatically generate XML metadata files for our datasets. Figure 1 gives an overview of the process. The relational databases contain measurement values, of for example biogeochemical samples. In addition, a lot of metadata information is stored across the various database tables, such as contact information, measurement parameters, units and methods and general information on the dataset, such as the campaign. Every table is checked for relevant metadata information. Each of the found attributes will be used to fill an ISO 19115-1:2014 element.

The metadata fields are loaded through SQL calls into the Feature Manipulation Engine (FME). FME is a software product to connect different data sources, manipulate their features and generate new results through automated workflows.

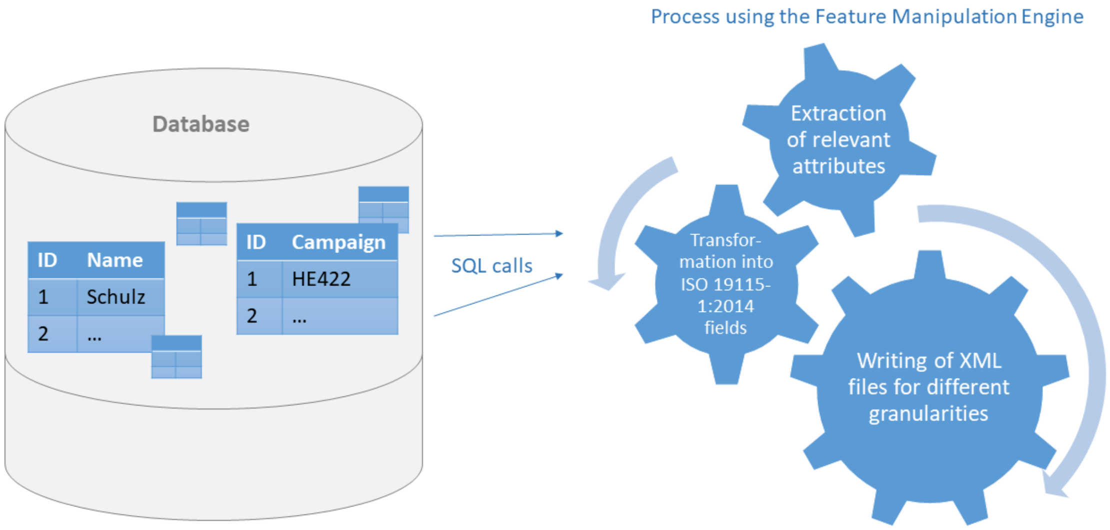
Process using the Feature Manipulation Engine
Database
ID Name
1 Schulz
2 ...
ID Campaign
1 HE422
2 ...
SQL calls
Extraction of relevant attributes
Transfor-mation into ISO 19115-1:2014 fields
Writing of XML files for different granularities

Figure 1: Generating XML files from database tables using the Feature Manipulation Engine

Using FME, the attributes gathered from the database are aggregated, transformed and mapped to the relevant metadata fields in the ISO 19115-1:2014 standard. For example, contact information is mapped to the field CI_ResponsibleParty and descriptions of parameters as well as standardized parameter names from controlled vocabularies, such as the ones provided by SeaDataNet, are mapped to MD_Keywords. The most evolved part of the process is the automated generation of the abstract of the dataset, which is part of the field CI_Citation. A generic template for all datasets is developed, that summarizes all known information of the campaigns, such as their purpose, and the involved projects. It also includes information on Digital Object Identifiers of the dataset.

As we have seen in the previous paragraphs, the metadata files are filled based on the information stored in the database. Thus, increasing the amount of such information stored is important. The sooner details about a measurement, like the coordinates of a station, are digitalized, the better the resulting metadata files will be. Since 2018 we have been generating metadata of campaigns already in the field, while the scientists are still on the research vessel. All campaigns have been equipped with high accuracy GNSS receivers and water proof tablets to electronically gather metadata. Using an app called Survey123 by Esri, the scientist fills in a questionnaire asking for example for the name of the station, the time of the sampling and the coordinates. The latter is directly received from the GNSS receiver, which is connected to the app to get the location of each sampling site with approximately one meter accuracy. The results of the questionnaire are then transferred via the mobile network from the tablet to a cloud storage. From there it is imported into the relational database, after carrying out quality checks (see Figure 2).

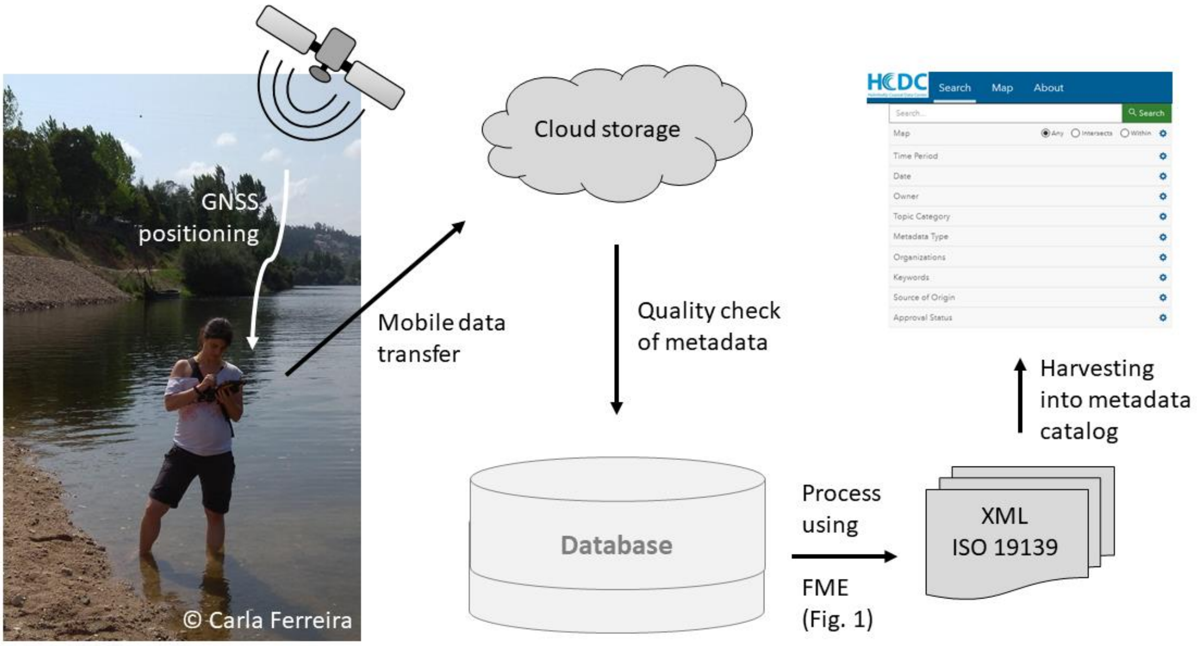

Figure 2: Metadata workflow from the field to the metadata catalogue

Once all metadata elements are mapped to the appropriate ISO 19115-1:2014 fields, an ISO 19139 XML file is generated by FME. This process is repeated for different granularities, such as each campaign and each project. The output XML files are stored in a folder, which is regularly harvested by the metadata catalogue system. A secondary batch process is set up to ensure that outdated metadata records are purged from the catalogue by directly deleting the old entries from the index. In our case of the metadata catalogue system Geoportal, old entries are removed from the open-source Elasticsearch instance and replaced by new information.

The automated process described above has been adapted to different relational databases, storing different data types, for example biogeochemical data and real time observational data. It has proven to be easily adjustable to new situations. After an initial setup phase, it automatically generates ISO 19115-1:2014 conform metadata files. In the future it might be possible to develop a user friendly generic tool from the described process that generates metadata files from any coastal research database.