EGU2020-7058
https://doi.org/10.5194/egusphere-egu2020-7058
EGU General Assembly 2020

Enabling Data Reuse Through Semantic Enrichment of Instrumentation

Robert Huber[1], Anusuriya Devaraju[1], Michael Diepenbroek[1], Uwe Schindler[1], Roland Koppe[2], Tina Dohna[1], Egor Gordeev[1], and Marianne Rehage[1]

[1]Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany
[2]Alfred Wegener Institut Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany

Pressing environmental and societal challenges demand the reuse of data on a much larger scale. Central to improvements on this front are approaches that support structured and detailed data descriptions of published data. In general, the reusability of scientific datasets such as measurements generated by instruments, observations collected in the field, and model simulation outputs, require information about the contexts through which they were produced. These contexts include the instrumentation, methods, and analysis software used. In current data curation practice, data providers often put a significant effort in capturing descriptive metadata about datasets. Nonetheless, metadata about instruments and methods provided by data authors are limited, and in most cases are unstructured.

The 'Interoperability' principle of FAIR emphasizes the importance of using formal vocabularies to enable machine-understandability of data and metadata, and establishing links between data and related research entities to provide their contextual information (e.g., devices and methods). To support FAIR data, PANGAEA is currently elaborating workflows to enrich instrument information of scientific datasets utilizing internal as well as third party services and ontologies and their identifiers. This abstract presents our ongoing development within the projects FREYA and FAIRsFAIR as follows:

- Integrating the AWI O2A (Observations to Archives) framework and associated suite of tools within PANGAEA's curatorial workflow as well as semi-automatized ingestion of observatory data.
- Linking data with their observation sources (devices) by recording the persistent identifiers (PID) from the O2A sensor registry system (sensor.awi.de) as part of the PANGAEA instrumentation database.
- Enriching device and method descriptions of scientific data by annotating them with appropriate vocabularies such as the NERC device type and device vocabularies or scientific methodology classifications.

In our contribution we will also outline the challenges to be addressed in enabling FAIR vocabularies of instruments and methods. This includes questions regarding reliability and

trustworthiness of third party ontologies and services. Further, challenges in content synchronisation across linked resources and implications on FAIRness levels of data sets such as dependencies on interlinked data sources and vocabularies.

We will show in how far adapting, harmonizing and controlling the used vocabularies, as well as identifier systems between data provider and data publisher, improves the findability and re-usability of datasets , while keeping the curational overhead a slow as possible. This use case is a valuable example of how improving interoperability through harmonization efforts, though initially problematic and labor intensive, can benefits to a multitude of stakeholders in the long run: data users, publishers, research institutes, and funders.