The Application of Semantic Resources and Technologies for the Discovery and Integration of Geo- and Biosciences Data

Michael Diepenbroek [1]

[1] *MARUM, University of Bremen, Bremen, Germany*

Abstract

Large-scale and complex questions in science, such as global warming, invasive species spread, and resource depletion, increasingly require the collection of disparate data sets from various data sources building on different knowledge domains in science and society. Structured data and metadata with consistent semantics are prerequisites for data usability, in particular for findability of data and efficient data integration. Ontologies, thesauri, and vocabularies for various science domains have been evolving tremendously during the last decade and play a key role for the harmonization of data. Nevertheless, the application of terminologies in the context of data production, archiving, and publication is still at its beginning. In addition, features and usability of terminology services vary greatly. The situation is aggravated by the complexity and dynamic growth of measurement and observation types (parameters) including used methods which are essential for integrating data from distributed sources.

The ISC World Data Center PANGAEA (www.pangaea.de) with ~200.000 parameters and methods linked to more than 400.000 data sets covers a large part of scientific fields in the earth and environmental sciences. For the harmonization of parameters and methods PANGAEA has (1) embedded a term catalogue (TC) comprising various relevant terminologies including taxonomies into its editorial system, (2) has conceptualized parameters and methods by setting up a basic syntax and rule set, and (3) has implemented routines based on full text search for matching parameter and method names with terms from the TC. Despite these measures being quite successful it must be noted that the approach is limited to PANGAEA as a single data provider - the needed effort is high.

More recently, Germany launched the National Research Data Infrastructure (NFDI - https://www.nfdi.de/) initiative with a number of consortia covering various science domains. The NFDI4BioDiversity (www.nfdi4biodiversity.org) consortium, having started in 2020, leads the development of a multi-cloud-based infrastructure supported by almost all existing consortia. The so-called NFDI Research Data Commons (RDC) will enable uniform access to data, software, and compute resources as well as sovereign data exchange and collaborative work. Harmonization of the semantics of data during the ETL process will be supported by terminology services enhanced by AI technologies. With the RDC, a paradigm shift from data to function shipping is initiated.

The approach aligns well with initiatives like the European Open Science Cloud, EOSC (https://eosc-portal.eu/), the NCI RDC (https://datascience.cancer.gov/data-commons) or the Australian RDC (https://ardc.edu.au/). Nevertheless, integrating and harmonizing data into the conceived cloud based systems remains a major challenge: (1) Terminology services need to improve in quality and functionality; (2) AI technologies are not yet part of the integration process; (3) more convergence towards cross-domain usable metadata standards like schema.org - allowing community specific extensions like BioSchemas (https://bioschemas.org/) - is needed to keep the ETL process manageable; and finally (4) a comprehensible, generally applicable model for the definition of parameters and methods would make the task considerably easier. The latter should build in part on the UCUM system (https://ucum.org/) for scientific units.

S4BioDiv 2021: 3rd International Workshop on Semantics for Biodiversity, held at JOWO 2021: Episode VII The Bolzano Summer of Knowledge, September 11-18, 2021, Bolzano, Italy

Bibliography

Dr. Michael Diepenbroek, Geologist and IT Specialist. 1992 PhD in Geology at the Free University of Berlin; 1992-94 computer center of the AWI, Bremerhaven; 1994-97 conception and implementation the scientific information system PANGAEA®; 1998-2021 at MARUM, University Bremen, where he was responsible for the management of PANGAEA®. During the last 10 years he took a leading role establishing PANGAEA as a global service provider for scientific data, in particular through mandates from the ISC (ISC World Data System - Vice-Chair of the Scientific Committee 2009-2016), the WMO (WMO Information System), collaborations with major science publishers, and as Chair/Co-chair in various RDA groups. Coordinator of the German Federation for Biological Data (GFBio) (2013-2020). Since 2017 engaged in the National Research Data Infrastructure Initiative (NFDI), in particular in the conception and preparation of NFDI4BioDiversity (https://www.nfdi4biodiversity.org). Since 2021 working for GFBio e.V. as part of NFDI4BioDiversity.