EGU2020-9412
https://doi.org/10.5194/egusphere-egu2020-9412
EGU General Assembly 2020

Semantic harmonization of geoscientific data sets using Linked Data and project specific vocabularies

Martin Schiegl[1], Gerold W. Diepolder[2], Abdelfettah Feliachi[7], José Román Hernández Manchado[4], Christine Hörfarter[1], Olov Johansson[3], Andreas-Alexander Maul[5], Marco Pantaloni[6], László Sőrés[8], and Rob van Ede[9]

[1]Geologische Bundesanstalt (GBA), Vienna, Austria
[2]Bayerisches Landesamt für Umwelt – Geologischer Dienst (LfU), Augsburg, Germany
[3]Geological Survey of Sweden (SGU), Uppsala, Sweden
[4]Instituto geológico y minero de España (IGME). Madrid. Spain
[5]Federal Institute for Geosciences and Natural Resources (BGR), Hannover, Germany
[6]Geological Survey of Italy (ISPRA) Rome, Italy
[7]Bureau de Recherches Geologiques et Minieres (BRGM), Orléans, France
[8]Magyar Bányászati És Földtani Szolgálat (MBFSZ), Hungary
[9]Nederlandse Organistie voor Toegepast Natuurwetenschappelijk Onderzoek (TNO), The Netherlands

In geosciences, where nomenclature naturally has grown from regional approaches with limited cross-border harmonization, descriptive texts are often used for coding data whose meanings in the international context are not conclusively clarified. This leads to difficulties when cross border datasets are compiled. On one hand, this is caused by the national-language, regional and historical descriptions in geological map legends. On the other hand, it is related to the interdisciplinary orientation of the geosciences e.g. when concepts adopted from different areas have a different meaning. A consistent use and interpretation of data to international standards creates the potential for semantic interoperability. Datasets then fit into international data infrastructures. But what if the interpretation to international standards is not possible, because there is none, or existing standards are not applicable? Then efforts can be made to create machine-readable data using knowledge representations based on Semantic Web and Linked Data principles.

With making concepts reference able via uniform identifiers (HTTP URIs) and crosslinking them to other resources published in the web, Linked Data offers the necessary context for clarification of the meaning of concepts. This modern technology and approach ideally complements the mainstream GIS (Geographic Information System) and relational database technologies in making data findable and semantic interoperable.

GeoERA project (Establishing the European Geological Surveys Research Area to deliver a Geological Service for Europe, https://geoera.eu/) therefore provides the opportunity to clarify expert knowledge and terminology in the form of project specific vocabulary concepts on a scientific level and to use them in datasets to code data. At the same time, parts of this vocabulary

might be later included in international standards (e.g. INSPIRE or GeoSciML), if desired. So called “GeoERA Project Vocabularies” are open collections of knowledge that, for example, may also contain deprecated, historical or only regionally relevant terms. In an ideal overall view, the sum of all vocabularies results in a knowledge database of bibliographically referenced terms that have been developed through scientific projects. Due to the consistent application of the data standards of Semantic Web and Linked Data nothing stands in the way of further use by modern technologies such as AI.

Project Vocabularies also could build an initial part of a future EGDI (European Geological Data Infrastructure, http://www.europe-geology.eu/) knowledge graph. They are restricted to linguistic labeled concepts, described in SKOS (Simple Knowledge Organization System) plus metadata properties with focus on scientific reusability. In order to extend this knowledge graph, additionally they also could be supplemented by RDF data files to support project related applications and functionality.