Ontology-Enabled Metadata Schema Generator: The Design Approach

Jian Qin
School of Information Studies
Syracuse University
jqin@syr.edu

Xiaozhong Liu
School of Informatics and Computing
Indiana University Bloomington
liu237@indiana.edu

Miao Chen
Data to Insight Center
Indiana University Pervasive Technology Institute
Mchen14@syr.edu

Metadata standards are important for normalizing descriptions of publications and research data and for information discovery and use. Large, complex metadata standards, however, can complicate the creation, sharing, and maintenance of metadata and incur high costs for metadata operations, especially in the domain of scientific data (Qin et al., 2010; Qin Ball, & Greenberg, 2012; Qin & Li, 2013). One strategy to solve the problems of large, complex metadata standards is to break them into independent modules to allow for reuse of elements and maximal possibility of automation. To implement this strategy, we need a metadata infrastructure that contains elements, vocabularies, and other metadata artifacts and that is easy to use. This short paper describes the design approach to an ontology-enabled metadata schema generator as part of the metadata infrastructure.

Elements in metadata standards in the scientific data domain tend to follow a pattern that a small number of (super-) general elements co-occur in a large number of standards and those co-occurred in 2-4 standards tend to be field-general. Even though semantically same elements co-occurred across different standards, they often varied in singular-plural forms, capitalization, or complete different words (Qin & Li, 2013). These inconsistencies and varying naming conventions can be mitigated by ontologies. These ontologies as the semantic underpinning for scientific metadata will have different types, e.g., entity ontologies for person, organization, project, study, dataset, and so on, or temporal ontologies for date and time. They will be built by following the portability principle (Qin, Ball, & Greenberg, 2012).

Many semantic resources have been made available in linked data format and can be utilized to avoid reinvent the wheels in the process of building a metadata infrastructure. Using the open identity metadata ORCID and ResearcherID as an example, FIG. 1 shows the design approach that uses the open identity metadata and portable metadata schemes in the form of ontologies as the input for the metadata scheme generator. The generator will then output the elements and relations selected by the user in the form of an RDF schema or other format.

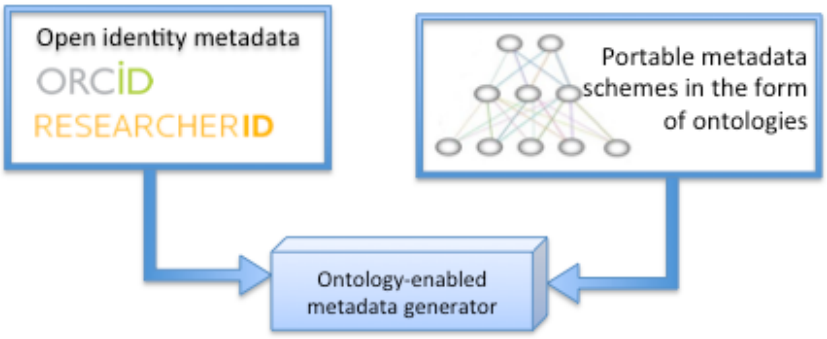

FIG. 1. A conceptual structure of an ontology-enabled metadata generator

Envisioning how such a metadata infrastructure might affect the metadata creation process, we developed a case scenario for the ontology-enabled metadata schema generator. The Ecological Metadata Language (EML) and Content Standard for Digital Geospatial Metadata (CSDGM) are two standards used to describe ecological data and geospatial data respectively. Both of them contain sections/modules of metadata elements and share some general descriptive elements. The EML eml-resource module contains elements such as title, creator, pubDate, abstract, keyword, and keywordThesaurus, which overlap with similar elements in CSDGM, namely title, originator, pubdate, abstract, keywords, and themekt. The two standards also share similar elements for describing time and space. For example, EML's eml-coverage module contains granular modules of "time", "beginDate", and "endDate", which correspond to similar temporal elements in CSDGM. The element "boundingCoordinates" is identical in both standards. These observations demonstrate that different metadata standards contain similar elements and the shared elements may be generalized as ontologies and reused in building new metadata schemas.

When a collection of ontologies is created, differences in naming semantically same elements can be smoothed out and structures established between elements based on their relations. A metadata generator can take the advantages of the ontology collection as well as external semantic resources for building customized metadata schemas. The ontology-enabled metadata generator will have the functions and capabilities to:

- Facilitate interactive metadata selection and structure definition in developing customized metadata schemas;
- Preload instances as the values for frequently used elements , e.g., project team members' names and affiliations, or project/study description;
- Automatically acquire identity information for entities and elements; and
- Build portable metadata to enable faster dissemination, access, and reuse.

To fulfill this design, there are more technical issues beyond metadata to be solved. For example, what kind of architecture for the back-end will be needed to enable a visualized, interactive, and dynamic front-end for the metadata schema generator? What are the boundaries between metadata modules and how will they affect the size and structure of ontologies? How will the ontologies be defined and maintained? For a novel thinking of metadata schema creation, we recognize that there may be more questions than answers at this early stage of development. Work is already underway for building a prototype as a proof of concept. We are hoping to have an example for the demo at the time of this workshop.

References

EML. http://knb.ecoinformatics.org/software/eml/

FGDC. http://www.fgdc.gov/metadata/geospatial-metadata-standards

Qin, J., A. Ball, & J. Greenberg. (2012). Functional and architectural requirements for metadata: Supporting discovery and management of scientific data. *Dublin Core International Conference DC-2012*, *Kuching, Malaysia, September 3-7, 2012*.

Qin, J. & K. Li. (2013). How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure. *Dublin Core International Conference DC-2013*, *Lisbon, Portugal, September 2-6, 2013*.

Qin, J., M. Chen, X. Liu, & A. Wiggins. (2010). Linking entities in scientific metadata. In: *Proceedings of the Dublin Core International Conference DC-2010*, Pittsburg, PA, October 20-22, 2010.