Discussion Topic

Automated Data Integration –How to enable a not (yet) Redeemed Promise?

Sirko Schindler, Friederike Klan (DLR Institute of Data Science, German Aerospace Center, Germany)

One of the Semantic Web's promises is easy ad-hoc integration of data from multiple, heterogeneous sources. Ontologies describe both data and metadata in a machine readable way. With these descriptions, semantically enabled applications can automatically pull data from different sources, superseding the time-consuming process of manually created ETL-tasks.

Ontologies by themselves do not dictate how to arrange a domain's concepts and relations. While this enables a plethora of possible use cases, this very flexibility provides a challenge in its own right: With the variety in options, there comes a variety in solutions. Over time different models for the same domain evolve, each valid and justified in its own right. However, when integrating data described using different models, this fragmentation of solutions yields similar problems to data integration it tried to solve in the first place.

One area, where this is particularly evident, is the formal description of observational data and measurements as present, e.g., in the life and geo sciences, or in citizen observations. Over time several standards with overlapping semantics have emerged: The OGC/W3C Semantic Sensor Network Ontology (SSN, https://www.w3.org/TR/vocab-ssn/), the OGC/ISO Observation and Measurement (O&M) conceptual model (https://www.iso.org/standard/32574.html), or the W3C RDF Data Cube Vocabulary (https://www.w3.org/TR/vocab-data-cube/) to name just a few.

However, practitioners (e.g., data managers of research projects or large research data infrastructures) have developed their own vocabularies and models, rarely adopting existing standards: Examples include EnvThes (http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn, Center for Ecology and Hydrology), or Anaeethes (https://fairsharing.org/FAIRsharing.49bmk). These domain-driven models often adopt less rigorous ways of describing and interlinking vocabularies. They oftentimes use lightweight vocabularies like SKOS (http://www.w3.org/TR/skos-reference, W3C) or provide mere term-lists to describe their observational data.

This variety of data models hinders the automated integration of observational data, leaving it a tedious manual task. This hampers large-scale data analyses needed, e.g., for assessing global trends.

One appraoch to this problem is to find a canonical model at least for a particular domain that subsumes all the different aspects of models created. This approach has often been attempted, but failed as least as much. If not for opinions and personal reservations, then mutually exclusive modeling decisions make it next to impossible to find unified data models that are adopted by communities as a whole. With the rise of scientific data portals, several community-driven efforts attempted to harmonize the description of observational data in order to facilitate data integration. In parallel to domain-agnostic standards, those efforts lead to data models such as OBOE: The Extensible Observation Ontology (https://github.com/NCEAS/oboe) or the Biological Collections Ontology (http://www.obofoundry.org/ontology/bco.html). However, a consensus on one canonical model has not been reached yet.

This is not surprising. Although two data models might refer to the same part of the world, they are created with different use cases in mind, imposing different requirements and thus leading to structurally and semantically different models.

A viable alternative to a unified model is to allow for a well-designed set of co-existing models linked by mappings to mediate between the different perspectives taken by the models. First attempts to interlink existing models have been made, e.g., by SSN offering alignments to O&M and OBOE.

To date however, there are serious barriers to the vision of easy, ad-hoc data integration driven by interlinked data models.

High entry barrier for rich semantic data models

While standardization efforts focus on semantically rich data models expressed in formal knowledge representation languages, practitioners seem to prefer semantically lightweight modeling approaches like SKOS. A reason might be that practitioners are typically experts in their respective application domain, but not familiar with formal logic. By lowering the barriers for describing data models using Semantic Web standards, SKOS fosters sharing of standardized data models. This comes at the cost of semantic richness. SKOS offers properties such as `skos:broader`, `skos:narrower`, `skos:related` and `skos:closeMatch`, `skos:exactMatch`, `skos:broadMatch`, `skos:narrowMatch` and `skos:relatedMatch` to indicate relationships between concepts within and among concept schemes. However, it does not provide means to (and was never meant for) precisely describing the semantics of concepts and the relationships between these. Data models described using SKOS thus do not offer sufficient information for automated data integration. We argue, that to fill this information gap means are required that enrich SKOS-based thesauri and taxonomies making their hidden semantics explicit, e.g., by aligning them with more formal and rich data models.

Missing standards for domain-agnostic properties

An important prerequisite for automated data integration is explicit knowledge about the precise relationship between concepts defined within and between data models. RDFS and OWL 2 offer great flexibility by leaving the definition of properties to ontology designers. This results in each data model coming with its own set of properties, which hampers data integration across models. To harmonize the definition of properties across domains and models, we argue in favor of a set of universal, i.e. domain-agnostic properties with a well-defined meaning . One attempt towards this goal is the Relation Ontology (RO) (http://www.obofoundry.org/ontology/ro.html) developed within the OBO Foundry (http://www.obofoundry.org/), Beside domain-specific relationships, RO defines 10 generic properties intended for cross-domain use (https://github.com/oborel/obo-relations/wiki/ROCore)including, e.g., `part of` or `located in`. Currently omitted is the fact that relationships might just hold for a certain period of time.

Regardless of their semantically overlapping aspects, data models can largely differ in their structure, i.e. the relationships defined between concepts. When aligning different data models , this can lead to complex mappings,Here, available language elements such as `owl:equivalentClass`, `owl:equivalentProperty`, or `owl:propertyChainAxiom` may not suffice, e.g., for mappings involving not just chains of properties but more complex structures.

We want to use the workshop to discuss those and other barriers to easy, ad-hoc data integration and how these can be addressed by efforts under the umbrella of W3C. We would like to emphasize easy to use but yet expressive approaches that can find support with domain experts and ontology engineers alike.

Background of the authors

The authors have several years of experience with data management in scientific projects in the life and geosciences as well as in citizen science projects and management of socio-economic databases. This includes work and research on semantic data models, mainly for observational data and measurements, as well as research on data search, data integration (http://ceur-ws.org/Vol-1933/paper-9.pdf), ontology quality (http://semantic-web-journal.net/system/files/swj1825.pdf), and techniques supporting the reuse of ontologies (http://fusion.cs.uni-jena.de/fusion/wp-content/uploads/2017/09/ekaw2016.pdf). Previous work also includes comparative studies of ontologies in the field of units of measurement and the modeling of observations and phenomena (https://www.w3.org/community/owled/files/2016/11/OWLED-ORE-2016_paper_5.pdf).

They are involved in standardization efforts for observational data models in the context of citizen science and earth observation.

- This includes activities in an emerging working group dedicated to enabling interoperability between observed parameters. The group comprises of practitioners involved in different national and international projects and research infrastructures Working Group (https://icei2018.uni-jena.de/wp-content/uploads/2018/08/paper_151.docx) and is supposed to be hosted by the Research Data Alliance (https://www.rd-alliance.org/).
- The authors are also involved in similar initiatives in the context of Citizen Science and Erath Observation, e.g. In the OGC Interoperability Experiment (https://www.opengeospatial.org/projects/initiatives/citsci-ie) and a community of practice for citizen observatories recently established within the EU project WeObserve (http://www.iiasa.ac.at/web/home/research/researchPrograms/EcosystemsServicesandManagement/event/180603-WeObserveCOPsLaunch.html) as well as in data standardization activities,e.g. within the Citizen Science COST Action CA15212 (https://www.cs-eu.net/wgs/wg5) or the OGC (description of Earth Observation Products).