Collaborative Metadata Definition using Controlled Vocabularies, and Ontologies

Ilia Bagov[‡], Christian Greiner[‡], Nikolay Garabedian[‡]

[‡] Karlsruhe Institute of Technology, Institute for Applied Materials, Karlsruhe, Germany

Corresponding author: Christian Greiner (christian.greiner@kit.edu)

Abstract

Data's role in a variety of technical and research areas is undeniably growing. This can be seen, for example, in the increased investments in the development of data-intensive analytical methods such as artificial intelligence (Zhang 2022), as well as in the rising rate of data generation which is expected to continue into the near future (Rydning and Shirer 2021). Academic research is one of the areas, where data is the lifeblood of generating hypotheses, creating new knowledge, and reporting results. Unlike proprietary industry data, academic research data is often subjected to stricter requirements regarding transparency, and accessibility. This is in part due to the public funding which many research institutions receive. One way to fulfil these requirements is by observing the FAIR (Findability, Accessibility, Interoperability, Reusability) principles for scientific data (Wilkinson et al. 2016). These introduce a variety of benefits, such as increased research reproducibility, a more transparent use of public funding, and environmental sustainability. A way of implementing the FAIR principles in practice is with the help of FAIR Digital Objects (FDOs) (European Commission: Directorate-General for Research and Innovation 2018). A FDO consists of data, an accompanying Persistent Identifier (PID), and rich metadata which describes the context of the data. Additionally, the data format contained in an FDO should be widely used, and ideally open. Our presentation is focused on the third of FDO's components mentioned previously – metadata. It outlines the concept for a framework which enables the collaborative definition of metadata fields which can be used to annotate FDO-encapsulated data for a given domain of research.

The first component of the presented framework is a controlled vocabulary of the domain related to the data which needs to be annotated. A controlled vocabulary is a collective that denotes a controlled list of *terms*, their *definitions*, and the *relations* between them. In the framework presented in this contribution, the *terms* correspond to the metadata fields used in the data annotation process. Formally, the type of controlled vocabularies used in the framework is a thesaurus (National Information Standards Organization 2010). Thesauri consist not only of the elements mentioned previously, but also allow for the inclusion of synonyms for every defined term. This eliminates the ambiguity which can occur when

using terms with similar definitions. Additionally, thesauri specify simple hierarchical relations between the terms in the vocabulary, which can provide an explicit structure to the set of defined metadata fields. The most important feature of our framework, however, is that the controlled vocabularies can be developed in a collaborative fashion by the domain experts of a given research field. Specifically, people are able to propose term definitions and edits, as well as cast votes on the appropriateness of terms which have already been proposed.

Despite their advantages, one limit of thesauri is their lacking capability of relating metadata fields to each other in a more semantically rich fashion. This motivated the use of the second component of the framework, namely ontologies. An ontology can be defined as “a specification of a conceptualization” (Gruber 1995). More precisely, it is a data structure which represents entities in a given domain, as well as various relations between them. After a set of metadata fields has been defined within a controlled vocabulary, that vocabulary can be transformed into an ontology which contains additional relations between the fields. These can extend beyond the hierarchical structure of a thesaurus and can contain domain-specific information about the metadata fields. For example, one such relation can denote the data type of the value which a given field must take. Furthermore, ontologies can be used to link not only metadata, but also data, as well as individual FDOs themselves. This can contribute to the Reusability aspect of FAIR Data Objects. For example, an FDO generated by a research group in a given domain can be linked to an existing domain ontology. Afterwards, the FDO can be reused more easily by researchers from the same scientific field, because the ontology will have already specified the FDO's relation to the subject area. Additionally, cross-domain ontologies can be combined with each other which can increase the reusability of FDOs beyond their domain boundaries.

The components described above are being implemented in the form of multiple software tools related to the framework. The first one, a controlled vocabulary editor written as a Python-based web application called VocPopuli, is the entry point for domain experts who want to develop a metadata vocabulary for their field of research or lab. The software, whose first version is already being tested internally, enables the collaborative definition, and editing of metadata terms. Additionally, it annotates each term, as well as the entire vocabulary, with the help of the PROV Data Model (PROV-DM) (Moreau and Missier 2013) - a schema used to describe the provenance of a given object. Finally, it assigns a PID to each term in the vocabulary, as well as the vocabulary itself. It is worth noting that the generated vocabularies themselves can be seen through the prism of FDOs: they contain data (the defined terms) which is annotated with metadata (e.g., the terms' authors) and provided with a PID.

The second software solution will facilitate the transformation of the vocabularies developed with the help of VocPopuli into ontologies. It will handle two distinct use cases – the from-scratch conversion of vocabularies into ontologies, and the augmentation of existing ontologies with the terms from a given thesaurus. As is the case with VocPopuli, the second tool is being developed in the Python programming language. The software solutions will be finally tested by two semi-overlapping groups of users from materials science. On the one hand, domain experts will input, edit, and discuss vocabulary terms in their area of interest, and thus create vocabularies. On the other hand, vocabulary and ontology administrators will oversee the vocabulary creation, and ontology transformation processes in a semi-automatic fashion.

After development is complete, the tools will be used in the creation of controlled vocabularies for various experimental procedures, as well as their transformation and/or integration into semantically richer ontologies. This will augment our already published work in the area (Garabedian et al. 2022) and will thereby test the integration of the new framework with already existing resources. The new vocabularies will describe processes in multiple domains, such as materials science, tribology, and metalworking. Afterwards, the developed thesauri will be used in the creation of metadata templates which can be used to annotate experimental data generated in the procedures mentioned above.

Keywords

research data, thesaurus, ontology engineering, Python

Presenting author

Ilia Bagov

Presented at

First International Conference on FAIR Digital Objects, poster

Funding program

Helmholtz Metadata Collaboration

European Research Council (ERC) Grant No. 771237 (TriboKey)

Alexander von Humboldt Foundation Postdoctoral Fellowship for Nikolay Garabedian

Grant title

MetaCook: The Metadata Cookbook

Hosting institution

- Institute for Applied Materials (IAM), Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, 76131, Karlsruhe, Germany

- KIT IAM-ZM MicroTribology Center (µTC), Strasse am Forum 5, 76131, Karlsruhe, Germany

Conflicts of interest

References

- European Commission: Directorate-General for Research and Innovation (2018) Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. https://op.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en. Accessed on: 2022-7-09.
- Garabedian N, Schreiber P, Brandt N, Zschumme P, Blatter I, Dollmann A, Haug C, Kümmel D, Li Y, Meyer F, Morstein C, Rau J, Weber M, Schneider J, Gumbsch P, Selzer M, Greiner C (2022) Generating FAIR research data in experimental tribology. Scientific Data 9 (1). https://doi.org/10.1038/s41597-022-01429-9
- Gruber T (1995) Toward principles for the design of ontologies used for knowledge sharing? International Journal of Human-Computer Studies 43: 907-928. https://doi.org/10.1006/ijhc.1995.1081
- Moreau L, Missier P (2013) PROV-DM: The PROV Data Model. W3C Recommendation. URL: https://www.w3.org/TR/2013/REC-prov-dm-20130430/
- National Information Standards Organization (2010) ANSI/NISO Z39.19-2005 (R2010) Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. URL: https://www.niso.org/publications/ansiniso-z3919-2005-r2010
- Rydning J, Shirer M (2021) Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts. https://www.idc.com/getdoc.jsp?containerId=prUS47560321. Accessed on: 2022-7-09.
- Wilkinson M, Dumontier M, Aalbersberg IJ, et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18
- Zhang D, et al. (2022) The AI Index 2022 Annual Report. AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University URL: https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf