Concept for metadata and time series data integration based on a material science application ontology

Paul Zierep, Dirk Helm

Fraunhofer IWM, Wöhlerstraße 11, 79108 Freiburg, Germany
paul.zierep@iwm.fraunhofer.de

1 Introduction

The matching between ontologies and classical datasets allows for an improved interpretability and understanding of the associated information and its interoperability. However, the integration of large tabular datasets poses various difficulties concerning the storage and accessibility, since the data cannot be accessed using only the ontology as an interface.

Whereas many categorical tabular datasets can be used to automatically populate ontologies, this approach is technically not feasible for large continuous datasets such as time series data. The conversion of such data into an ontology leads to limitations of the storage size and query speed of the used graph database backend. Although hybrid query systems such as R2RML and RML [1] allow for the mapping of relational databases and heterogeneous data formats to Resource Description Framework (RDF) graphs, these techniques are only suitable for trivial data schemas [2], which is not the case for our time series datasets.

In the presented poster, we propose a protocol to include time series data into a material science ontology using a two-step approach that combines ontology-based metadata querying with an additional functionality that allows for the performant retrieval of the associated time series data. The protocol demonstrates the first conceptual prototype of the envisioned system that will be used as a basis to implement a generic routine to complement large classical datasets with semantic meaning. We are looking forward to fruitful discussions with the community to further develop the proposed system.

2 Methods

2.1 Ontology development

To design the mapping concept, we developed a prototype ontology based on the tensile test experiment. The ontology concepts were derived from experimental datasets, the test standard ISO 6892-1 and interviews with domain experts. All concepts were classified into two superclasses (*MetaData* and *TimeSeriesData*), that provide the attributes required for the designed data population strategy. The structure of the superclasses are described in detail in the poster.

2.2 Data Integration Pipeline

To demonstrate data integration and data retrieval in our model system we implemented a basic pipeline. The pipeline was programmed using the ontology-based open-source Python framework SimPhoNy [3] and its major component OSP core (OSP: Open Simulation Platform). SimPhoNy allows for the manipulation of Abox individuals using CUDS (Common Universal Data Structure) objects.

Data parsing. To parse the raw datasets we implemented a mapping routine that reads the dataset and assigns the metadata to corresponding *MetaData* CUDS objects as well as each time series column to *TimeSeriesData* CUDS objects. The *TimeSeriesData* CUDS objects only store the information required to extract the specific column (e.g. column index, number of header rows and file path). The CUDS objects can be serialized to ontology individuals in a RDF graph.

Data retrieval. The data can be queried using the SimPhoNy application programming interface (API) as well as SPARQL queries. The *MetaData* individuals can be used to perform complex queries, such as filtering for tensile test experiments with specific properties (e.g., experiments that used as specific material X with a specimen width larger than Y). The corresponding *TimeSeriesData* individuals can be used to extract the column data using additional data analysis tools.

3 Results and Discussions

The implemented workflow enables the storing and retrieval of time series data in a semantically enriched ontology. The designed data parser allows for the parsing and mapping of semi-structured tabular data. The ontologically stored metadata can be used for complex queries of the time series data that would be difficult to perform using only the raw data files.

4 Funding

This research was funded by the Federal Ministry of Education and Research Germany (BMBF) within the project StahlDigital (funding code: 13XP5116C).

References

1. Zhao, Z., Han, S., Kim, J.: R2LD: Schema-based Graph Mapping of relational databases to Linked Open Data for multimedia resources data. Multimed Tools Appl. 78, 28835–28851 (2019). https://doi.org/10.1007/s11042-019-7281-5
2. Heyvaert, P., Dimou, A., Verborgh, R., Mannens, E.: Ontology-Based Data Access Mapping Generation Using Data, Schema, Query, and Mapping Knowledge. (2017)
3. OSP core. SimPhoNy. (2021). https://github.com/simphony/osp-core