A Semantic Catalogue for the Data Market Austria

Bernd-Peter Ivanschitz[1], Thomas J. Lampoltshammer[2], Victor Mireles[3], Artem Revenko[3], Sven Schlarb[4], and Lőrinc Thurnay[2]

[1] Research Studios Austria, Thurngasse 8/16, Vienna, Austria
bernd.ivanschitz@researchstudio.at
[2] Danube University Krems, Dr.-Karl-Dorrek-Str. 30, Krems an der Donau, Austria,
[3] Semantic Web Company, Neubaugasse 1, Vienna, Austria
[4] AIT Austrian Institute of Technology, Giefinggasse 4, Vienna, Austria

Abstract. The Data Market Austria (DMA) is an ecosystem of federated data and service infrastructures. It aims at making data from various data providers accessible and interoperable by allowing the submission, storage, management and dissemination of static datasets or streaming data services. By creating a metadata vocabulary, standardizing the ingest of data and ensuring the quality and completeness of metadata, it lays the ground to enable participants to share or consume datasets residing in different infrastructures. This demo focuses on the mapping services used in the DMA to standardize data from different sources using a modified version of the DCAT metadata schema. We present tools that enable inter organizational integration of datasets, in a manner that is both user-friendly and powerful enough to handle vast amounts of data.

Keywords: Metadata mapping, semantic enrichment, RDF,distributed systems, RML, Metadata catalogue

1 Introduction

The amount of data produced every day is growing at breathtaking speed – data has become an important asset that is of high importance in nearly every industry sector worldwide [6]. Therefore, a healthy data economy and a successfully functioning data-services ecosystem enable and ensure sustainable employment and growth and thereby societal stability and well-being [4,5]. Several issues have been identified as hindering the data economy in the Austrian case [2], among them the lack of interconnection between different infrastructures hosting data and data related services.

The Data Market Austria (DMA)[5] project addresses these problems by developing the technological, infrastructural, regulatory, and economic foundations for a comprehensive, innovation-supporting, sustainable Austrian data-services ecosystem. The technological foundation includes Blockchain technology

[5] https://datamarket.at/

for provenance, smart contracts and security, interconnected clouds, data access, constraint-preserving processing and analysis algorithms, semi-automated data quality improvement, and recommender-based brokerage technology. Additionally, two pilots in the areas of ICT for Mobility and ICT for Earth Observation are being developed to demonstrate the first usage scenarios of DMA.

The DMA is a network of participating (or member) organizations that contribute to the data market by offering their products in form of datasets or services to customers of the DMA. Each participating node must implement a defined set of services and mandatory standard interfaces. These are, for example instances of a *Data Crawler* a *Metadata Mapper*, a *Blockchain peer*, and *Data Management* and *Storage* components. Together with a common conceptual model, these standard interfaces represent the basis of interoperability for the use of datasets in the DMA.

The gateway to this network of nodes containing data and providing services is the *DMA portal* which, while not hosting any data or providing major services, collects information from all nodes to keep an up to date catalogue of available datasets. The focus of this demo is the design and implementation of this unified catalogue.

2 A Semantic Catalogue for a Data Market

Since the data in the DMA lies in a set of distributed repositories, it is necessary to build a unified catalogue to enable end users to search all available data sets and services. Furthermore, a single catalogue can be exploited for recommendation, deduplication, and various metadata quality measures. In the DMA, the creation of this unified catalogue is approached by creating i) a single metadata standard for unified representation of data sets, including standardized vocabularies for describing resources, ii) tools for facilitating the compliance of existing metadata with the previous points and iii) the technological foundation for the building and maintenance of the catalogue itself.

Metadata standard

The DMA metadata catalogue is based on DCAT-AP, the DCAT application profile for data portals in Europe[6] and extends the schema for DMA use cases. This standardization enables future cooperation with international data portals and ensures that the DMA is easily accessible for cooperating companies with a certain data quality standard. The DMA extension of the DCAT-AP, the Data Maket Core Vocabulary (DMAV), provides more classes and properties for describing datasets and services that are accessible on the DMA. The extension focuses on the business use case of the DMA and adds predicates covering topics like price modeling and dataset exchange, not present in the original DCAT-AP catalogue. The *dmav:priceModel* predicate, for example, allows us to handle the transaction fees for commercial datasets that are being made available in

[6] https://joinup.ec.europa.eu/release/dcat-ap-v11

the DMA. The *dmav:SLA* (Service Level Agreement) class allows to model the condition of a service contract in more details.

In the DMA metadata catalogue, every dataset constitutes an RDF[7] resource. There is a set of predicates that link every resource to different literals, which constitute the values of the metadata fields. These values can be of two types: i) literals, as in the case of *dcat:description* or *owl:versionInfo*, or ii) elements of a controlled vocabulary, as in the case of Language or License. These controlled vocabularies, which are managed by PoolParty Semantic Suite[8], enable accurate search, filtering and linking of different datasets. Additionally, the DMA includes a series of semantic enrichment services which automatically annotate free-text fields (such as *dcat:description* or *dcat:title*) with elements of controlled vocabularies.

Tools for adoption of the metadata standards

Since the DMA aims at making available data which was not originally produced for commercialization, we must assume that the metadata describing it does not comply to any particular standard. This is specially true because the data in each node is managed by a different organization. Therefore, the conversion to the unified metadata standard described above must be treated in a case by case basis.

The DMA provides two tools to facilitate this. The first is a UI component in which a node's administrator can upload a sample (in XML or JSON) of the metadata they wish to make abailable in the DMA. They are then prompted to select, for each of the metadata fields required by the DMA, which fields of their metadata schema should be used. This UI tool, called the *Metadata Mapping Builder* is, in essence, a user-friendly way to generate XPath and JSONPath expressions. Once these expressions have been generated, they are arranged into an RML[1] file, which is then used to produce RDF from similarly structured XML or JSON files.

Catalogue compilation and maintenance

Each node in the DMA that wishes to make a series of datasets available, must implement the following workflow. First, the *Data Harvesting Component*, which must be configured by the node's administrator to find the different datasets within the node, sends the corresponding metadata files to the *Metadata Mapping Service*, which uses the mapping file created as described above to generate, for each dataset, a set of RDF triples (serialized in Turtle format).

Afterwards, the dataset, its original metadata, and the corresponding RDF are ingested into the *Data Management* component which takes care of the packaging, versioning and assignment of unique identifiers to all datasets, whose

[7] https://www.w3.org/RDF/
[8] https://www.poolparty.biz/

hashes are furthermore registered in the Blockchain. Next The node's Data Management component publishes, through a ResourceSync[9] interface, links to metadata files in RDF format of recently added or updated datasets. This way, the node's metadata management is decoupled from the process of incorporating metadata into the DMA catalogue.

In the DMA's central node, the *Metadata Ingestion* component constantly polls the ResourceSync interfaces of all registered nodes, and when new datasets are reported, harvests their RDF metadata which, let us recall, already complies with the DMA metadata vocabulary. This metadata is then enriched semantically. The enrichment is based on EuroVoc[10], which is used in DMA as the main thesaurus. The NLP interchange format [3] is used for annotations, which are done in stand-off mode. The mapped and enriched metadata is then ingested into the *Search and Recommendation Services*. The high quality of the metadata and its compliance to the chosen scheme guarantees that the datasets and service are discoverable by the users of DMA.

With small variations, the processes described above are also used for ingesting publicly available data from goverment portals as well as ingesting small amounts of data that an individual would like to make available in the DMA.

Acknowledgements The Data Market Austria project is funded by the "ICT of the Future" program of the Austrian Research Promotion Agency (FFG) and the Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant no. 855404

References

1. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: Rml: A generic language for integrated rdf mappings of heterogeneous data. In: LDOW (2014)
2. Fernandez Garcia, J.D., Kiesling, E., Kirrane, S., Neuschmid, J., Mizerski, N., Polleres, A., Sabou, M., Thurner, T., Wetz, P.: Propelling the potential of enterprise linked data in austria. roadmap and report (2016)
3. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating nlp using linked data. In: International semantic web conference. pp. 98–113. Springer (2013)
4. Höchtl, J., Lampoltshammer, T.J.: Social Implications of a Data Market. In: CeDEM17 - Conference for E-Democracy and Open Government. pp. 171–175. Edition Donau-Universität Krems (2017)
5. Lampoltshammer, T.J., Scholz, J.: Open Data as Social Capital in a Digital Society. In: Kapferer, E., Gstach, I., Koch, A., Sedmak, C. (eds.) Rethinking Social Capital: Global Contributions from Theory and Practice, pp. 137–150. Cambridge Scholars Publishing, Newcastle upon Tyne (2017)
6. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity (2011)

[9] http://www.openarchives.org/rs/1.1/resourcesync
[10] http://eurovoc.europa.eu/