Towards a Common Description Vocabulary for Industrial Datasets

Christian Mader, Steffen Lohmann, Sören Auer
Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)

Abstract. In this position paper we motivate the need for establishing a common vocabulary for industrial datasets based on Linked Data technologies. We report on our progress in developing such a vocabulary within the Industrial Data Space project and providing insights on the general setting, the goals and our proposed methodology.

Introduction

The process of digitization is currently underway and is driving changes in society, businesses, and technology. An increasing amount of data becomes available throughout each step of the supply chain, from product development over shipment to customer delivery. However, data needs to be collected and processed throughout the whole product life-cycle to continuously create value. For example, information on a product's usage can be fed back to product planning and development. Exchanging this data on a bilateral basis in a secure, reliable way that conforms to the legal conditions is crucial, but making these data available for multiple interested stakeholders so that they can find, interpret and (re)use it becomes increasingly necessary. This would pave the way for treating data as a raw good to foster the development of new business models. Therefore, a consistent model for expressing dataset metadata information such as pricing, usage policies and clearance of datasets is a key requirement for successful future businesses.

Currently, information that is relevant for a company's business and needs to be shared, is still exchanged in proprietary formats (e.g., Excel) over inappropriate channels (e-mail, Skype, or uploaded to cloud services). This does not only violate data security principles, but also data sovereignty because data is copied to servers that are not within the premises of the rights holder, maybe even located in foreign countries under different legal framework. Furthermore, once sent out, the data cannot be pulled back again. Workflows become more complicated and tedious for both the publishing and the retrieving side, as additional communication effort is needed to explain data retrieval and usage.

In recent years, semantic technologies based on RDF have become popular for expressing metadata for digital libraries, open data portals and data catalogs in general. For these applications, they improve search and retrieval tasks as well as machine-supported interpretation and integration. We therefore argue to also use semantic technologies for describing and expressing industrial datasets in an agreed-upon way.

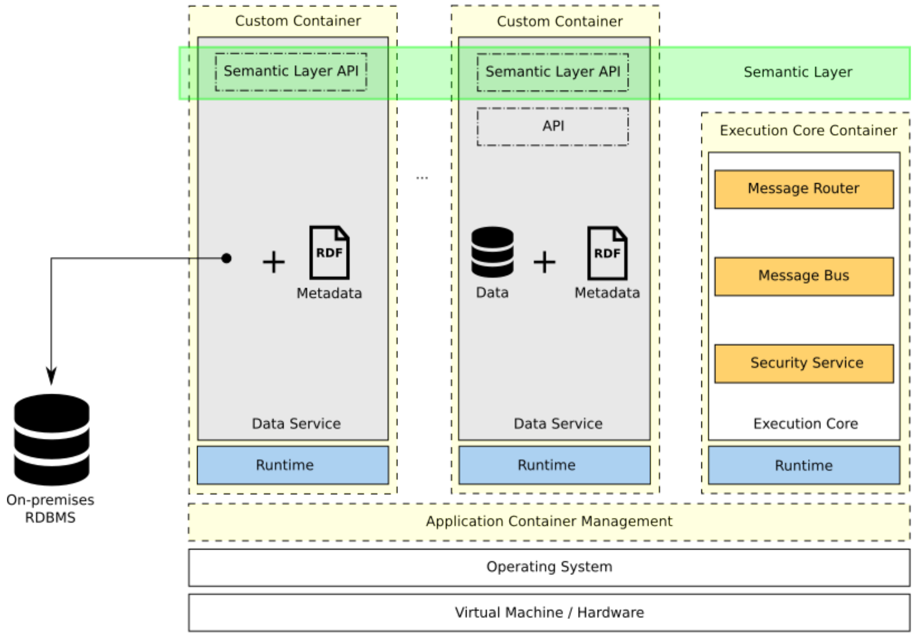
We believe that such a methodology will help to realize the vision of digital marketplaces for industrial datasets that allow to securely publish, find and combine datasets from various trusted providers. They can be key to the development of new business models that are based on the creation of additional value from existing data and thus act as a driver for the digital economy.

Motivating Setting

The Industrial Data Space (IDS) project[1] is a German nationally funded research project with the goal to create a specification and a demonstrator implementation of an infrastructure for industrial partners. It provides guidelines for the publication, retrieval and orchestration of datasets and services in a secure and interoperable way to address the challenges described above.

Architecture of the IDS Connector

To publish datasets and services for external access on the IDS, they must be deployed to the *Connector*, which is a runtime environment located inside a company network (or in a DMZ). Each dataset or service is encapsulated as a container that should implement an API of a *Semantic Layer* as outlined in Figure 1. This layer provides a REST API for retrieving (i) RDF metadata about the dataset and services, (ii) the data itself, as well as (iii) utility functions for, e.g., metadata querying.



Icon made by Freepik from www.flaticon.com

Figure 1. Schematic architecture of an IDS connector running two containers.

Metadata Dimensions for the IDS

When using the IDS connector for publication of datasets and services, their descriptive metadata must address different dimensions:

- General descriptive properties like provenance, recency, usage restrictions, or licenses.
- Properties related to the dataset’s content, such as topic, keywords, or quality metrics.
- Static metadata of a deployed container needed to evaluate if the encapsulated dataset or service can be used at a specific system. Examples are hardware requirements, security requirements, used technology and algorithms, or pricing.
- Runtime information of a deployed container for, e.g., tracking the usage of the dataset or service like deployment time or usage statistics.

[1] http:// www.industrialdataspace.org

For most of these dimensions, RDF-based vocabularies that provide properties for expressing this information already exist. Some of them are listed in Table 1.

Vocabulary	Domain	Namespace
DCMI Metadata Terms	General, Digital Library	http://purl.org/dc/terms/
vCard Ontology	People and Organizations	http://www.w3.org/2006/vcard/ns#
Friend of a Friend	People and Organizations, Social Networking	http://xmlns.com/foaf/0.1/
Creative Commons Rights Expression Language	Legal, Licensing	http://creativecommons.org/ns#
GoodRelations	eCommerce	http://purl.org/goodrelations/v1
Data Catalog Vocabulary	General, Dataset Metadata, Data Catalogs	http://www.w3.org/ns/dcat#
Data Quality Vocabulary	General, Dataset Quality	http://purl.org/eis/vocab/daq#
The PROV Ontology	Provenance	http://www.w3.org/ns/prov#

Table 1. Exemplary vocabularies for expressing IDS metadata.

When dataset content should be described, special ontologies which express topics or classifications can be adopted and reused. Examples of such are listed Table 2.

Vocabulary	Domain	Namespace
The Web Ontology for Products and Services	eCommerce	http://www.ebusiness-unibw.org/ontologies/eclass/5.1.4/#
Integrated Authority File (GND)	Library, Museum, Collections	http://d-nb.info/standards/elementset/gnd#
North American Industry Classification System (NAICS)	Economy	-
Library of Congress Subject Headings (LCSH)	Library, Museum, Collections	http://id.loc.gov/authorities/subjects/

Table 2. Exemplary vocabularies for dataset content annotation.

Furthermore, metadata descriptions should link to existing standards like those mentioned in the Industry 4.0 Standardization Landscape[2] in order to express the conformance level of the dataset. A metadata vocabulary should also be capable to state the algorithms and technologies used by service implementations. This way it is possible to describe the internal logic which is, e.g., necessary for deciding if a service must be taken offline or maintained because of newly discovered security issues.

[2] http://w3id.org/i40

Approach

Within the IDS project, our current efforts focus on the creation of the IDS metadata vocabulary whose classes and properties are aligned with already available vocabularies such as mentioned in the tables above. We follow a use-case driven methodology, i.e., based on the requirements of the industry partners within the IDS project, we incrementally build up the vocabulary using VoCol, a git-based collaborative vocabulary development platform [1]. The goal is to identify a concise set of classes and properties to express the requirements of the industry. We are currently in the process of reviewing how the dimensions pricing and usage restrictions can be expressed by this IDS metadata vocabulary. The immediate goal for exploiting the expressivity of the IDS metadata vocabulary will be to provide a search functionality that allows users to find and retrieve datasets that match specific requirements regarding the covered content, licence or usage restrictions.

Besides expressing information about the dataset itself, the IDS metadata vocabulary should also provide ways for describing the containers that encapsulate datasets and services. This makes it possible to manage and organize containers at a central registry (a IDS-specific *App-store*) so that they can be easily found (based on their functionality or popularity), downloaded and reused by participants of the IDS.

Discussion

The approach of the IDS for description and provision of datasets using metadata information is to some extent related to establishing a data catalog by using CKAN[3]. However, the IDS is built on the idea to publish all metadata information in RDF format and that the datasets themselves are not stored in a central place but at the premises of each IDS participant.

The Data Catalog Vocabulary (DCAT) [2] is an RDF-based vocabulary for describing datasets organized in data catalogs, addressing the problem of data catalog interoperability. Based on DCAT, DCAT-AP[4] has been developed as an application profile for data portals in Europe that hold public sector datasets. Similar to the DCAT-AP specification, we envision the IDS metadata vocabulary to also act as an agreed-upon basic vocabulary, but with the focus on describing industrial datasets.

Our immediate next steps are to continue with the creation and extension of the IDS metadata vocabulary for dataset descriptions, based on the use cases identified within the IDS project. We furthermore aim to ensure a broad uptake of the IDS Connector component by the industry. Therefore we are in the process of standardization as a DIN SPEC. It is our plan that we leverage the momentum of this standardization activity and extend it to also establish a national standards document covering the IDS metadata vocabulary terms.

References

[1] L. Halilaj, N. Petersen, I. Grangel-González, C. Lange, S. Auer, G. Coskun, and S. Lohmann, "Vocol: An integrated environment to support version-controlled vocabulary development," in Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016. Proceedings, 2016, to appear.

[2] Maali, F., Erickson, J. and Archer, P., 2014. Data catalog vocabulary (DCAT). *W3C Recommendation,* https://www.w3.org/TR/vocab-dcat/.

[3] http://ckan.org/

[4] https://joinup.ec.europa.eu/asset/dcat_application_profile/description