Raising the Role of Vocabulary Hubs for Semantic Data Interoperability in Dataspaces

ROBERT DAVID, The Semantic Web Company, Austria
VLADIMIR ALEXIEV and PETAR IVANOV, Ontotext, Bulgaria

Dataspaces are an important enabler for industrial sharing data (either commercially licensed or private). Europe is investing heavily into sectoral dataspaces, federation and orchestration platforms like SIMPL, Eclipse DSC, GXFS, etc. Still, dataspaces enable shared data access, but do not solve the data interoperability problem. For that, the consumer would like to see the data from different providers in a harmonized and semantically integrated form. The Vocabulary Hub service (part of the IDSA RAM) provides a repository for ontologies and vocabularies. We describe an approach of raising the role of the vocabulary hub to also allow richer metadata description (e.g. the meaning of every column in a tabular dataset), and binding semantic descriptions to ingested datasets, thus providing on-the-fly data semantization and easing data querying. This is achieved through the integration of two commercial semantic products (PoolParty and GraphDB), leveraging the partnership between the Semantic Web Company and Ontotext, and is being developed within the frame of the Digital Europe project UNDERPIN, with applications to refinery and wind farm data.

Additional Key Words and Phrases: dataspaces, semantic interoperability, semantic technologies, ontologies, vocabulary hub, oil and gas, renewable energy, refineries, windfarms

ACM Reference Format:
Robert David, Vladimir Alexiev, and Petar Ivanov. . Raising the Role of Vocabulary Hubs for Semantic Data Interoperability in Dataspaces. In *Proceedings of* . ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

The European data economy heavily depends on the availability of data and the technological foundation to make use of it. Different industries and communities can benefit mutually by sharing data with each other, thereby supporting their digitization business goals. Machine-learning systems heavily rely on high volumes of high-quality training data. Examples are the well-known mobility dataspace and use cases like energy communities for running simulations. To cope with the challenge of data sharing, Dataspace approaches like the IDS RAM were introduced to not only solve the technical aspects of data sharing, but also to establish methods for data sovereignty so that dataspace participants can determine themselves how and when others can make use of their data.

For making use of shared data, interoperability is a crucial factor. While syntactic data exchange is well covered by clearly defined data formats, there is still the open challenge of semantic interoperability. IDS RAM is firmly based on semantic metadata for all aspects and actors of a dataspace. Furthermore, it can leverage semantic interoperability approaches through the Vocabulary Hub that stores ontologies and semantic thesauri for data descriptions and can dereference URIs to provide semantic details of data assets.

Authors' Contact Information: Robert David, The Semantic Web Company, Austria; Vladimir Alexiev; Petar Ivanov, Ontotext, Bulgaria.

In this paper, we present our proposal for a technological solution of an IDS Vocabulary Hub built on the integration of the GraphDB semantic graph database and PoolParty Semantic Suite. We not only cover the core Vocabulary Hub functionalities and responsibilities in a dataspace, but also extend the idea towards a Semantic Layer for dataspaces, which provides even more services to easily and semi-automatically manage semantic interoperability. Thus we raise the role of the Vocabulary Hub to not only store semantic assets, but also to bind semantic descriptions dynamically to incoming data, therefore enabling richer discovery and semantic data integration between datasets.

2 PROJECTS AND USE CASES

In the expanding data economy landscape supported by dataspaces, we are working on several projects which address practical data sharing needs and where we experience the importance of semantic interoperability to fulfill various use cases. Our solution is developed with main focus on the UNDERPIN project with use cases of refineries and wind farms, but other projects, such as DataBri-X provide use cases which were considered.

2.1 Manufacturing/Maintenance Dataspace

Operating refineries and wind farms involves high maintenance costs. Predicting potential failures and optimizing the maintenance operations to minimize the costs can be business-critical. For establishing predictive maintenance, we need large amounts of high quality training data that can be obtained by data consolidation across different refinery and wind farm measurement sources.

The UNDERPIN project addresses this problem by creating a dataspace for European manufacturers in the refinery and renewable energy domains and various SMEs across their value-chain ecosystem, such as equipment suppliers. Extensive time series data is exchanged, but their format and meaning is not standardized. Describing the meaning of these time series facilitates semantic interoperability between datasets from different providers. This description includes measured variable, quantity kind and unit of measure, which sensor took the measurements, what it is attached to, the detailed model and connectivity of that industrial equipment.

2.1.1 The Refinery Dataspace. Refineries can be optimized regarding costs and efficiency by considering not only individual components, but the production chain as a connected system. In this use case, we aim to improve the maintenance process as well as the decision making for preventive maintenance so as to minimize the downtime and the impact on the production capabilities.

2.1.2 The Wind Farm Dataspace. Operating wind farms involves high costs for maintaining the wind turbines. Predicting potential failures and optimizing the maintenance operations to minimize the costs can be business critical. For establishing predictive maintenance, we need large amounts of high qualitative training data, which can be achieved by consolidating different wind farm measurement sources. This use case provides different kinds of tabular data, where consolidation needs semantic harmonization of data.

2.2 Energy and Legal Dataspace

Beyond the UNDERPIN project as the main use case provider, we also consider use cases from other projects. We present 2 examples from the DataBri-X project, which work on different content types of data and have different usage scenarios. We considered these as well when designing our solution to make it applicable to a broader range of use cases.

2.2.1 The Energy Dataspace. In this use case, we run simulations for energy communities, which predict the behavior of the energy grid. To provide precise predictions, there is the need of high amounts of example data. This data is of numeric type and different tabular data sources need to be semantically consolidated.

2.2.2 The Legal Dataspace. Analyzing documents in the legal domain is a highly challenging task because of the complexity of legal information. We provide recommender services for insights into corpora of legal documents that identify the meaning via a legal knowledge graph, NLP analysis and semantic annotations.

2.3 The Problem

"Standard" dataspace solutions provide access to data and metadata descriptions, but they do not typically address the data integration problem. If a consumer wants to consume datasets from different providers in an integrated way, they need to harmonize the data to some common model, convert it, store it in a database, and potentially implement entity linking (correlation of different records that are about the same real-world thing) and data fusion.

Above we gave just 4 examples to illustrate the heterogeneity of data in the context of different use cases. We aim to develop solutions that not only fulfill the dataspace needs, but also improve the situation regarding semantic interoperability by providing services that can be easily leveraged to implement such use cases.

3 APPROACH AND IT-SOLUTION

To tackle this problem, we introduce Semantic Web standards and technologies to model and process data. Vocabularies and ontologies (expressed using RDFS, OWL and SKOS) can represent data with a clear semantics based on standards. We leverage GraphDB and PoolParty as two software components that implement these standards to process the data.

3.1 IDS Vocabulary Hub

In this paper, we focus on the IDSA approach to dataspaces. IDSA defines the architecture of dataspaces in the reference architecture model IDS-RAM (current version 4). One of the defined architectural components, the Vocabulary Hub, has the role of providing vocabularies to annotate and describe data assets and services. These vocabularies form a common language to enable semantic interoperability among the dataspace participants. The Vocabulary Hub is responsible for resolving annotations of data assets and for providing details in the form of standardized semantic descriptions. It supports the use of Semantic Web standards for defining and representing RDF vocabularies and ontologies and thereby provides semantic interoperability in a standardized and machine-readable way.

3.2 GraphDB Graph Database

TODO

3.3 PoolParty Semantic Suite

PoolParty Semantic Suite is a semantic middleware platform based on W3C standards, specifically the Semantic Web, which provides a wide variety of functionality in the area of knowledge graph management, graph-based NLP, semantic search and recommender systems.

At the core of PoolParty is the Thesaurus Manager, which can be used to create and maintain RDF-based vocabularies, including ontologies, schemas and different kinds of knowledge models. These vocabularies can be applied for various use cases, from bridging data silos in enterprises by implementing data consolidation via the vocabularies, to annotating

documents with semantic annotations to implement semantic search and recommendations which provide insights into data.

PoolParty is well suited to form the basis of a vocabulary management system within a dataspace, effectively implementing not only the Vocabulary Hub role, but also providing additional services that improve the expressiveness of common vocabularies and automate semantic interoperability.

3.4 The GraphDB Poolparty Integration

The integration of GraphDB and PoolParty leverages synergies of a high-performance graph database with a platform for knowledge graph management.

Both GraphDB and PoolParty are based on W3C recommendations like RDF, OWL and SPARQL. They therefore seamlessly integrate with each other, but also with other services and components in dataspaces which support these machine-readable standards for semantic interoperability.

4 A SEMANTIC SOLUTION AND ITS BENEFITS

When harmonizing data, we need to provide a method to automatically map and process it, so this can be done efficiently on large datasets and/or regular updates in dataspaces. The benefit provided should support the ML training in such a way that the consolidated dataset is exposed as a uniform training dataset.

4.1 Semantic Layer Approach

The Vocabulary Hub, as IDS defines it in the IDS-RAM, is a basic building block to achieve semantic interoperability in dataspaces by provisioning common vocabularies. We aim to go one step further and envision a Semantic Layer for dataspaces, which provides vocabulary-based services for advanced and automated semantic descriptions of metadata and data. The Semantic Layer is implemented as services based on the GraphDB and PoolParty and provided as a one seamlessly integrated component.

In the following, we discuss i) the services provided for the Vocabulary Hub component and ii) additional services as Vocabulary Hub extensions that implement a Semantic Layer for dataspaces.

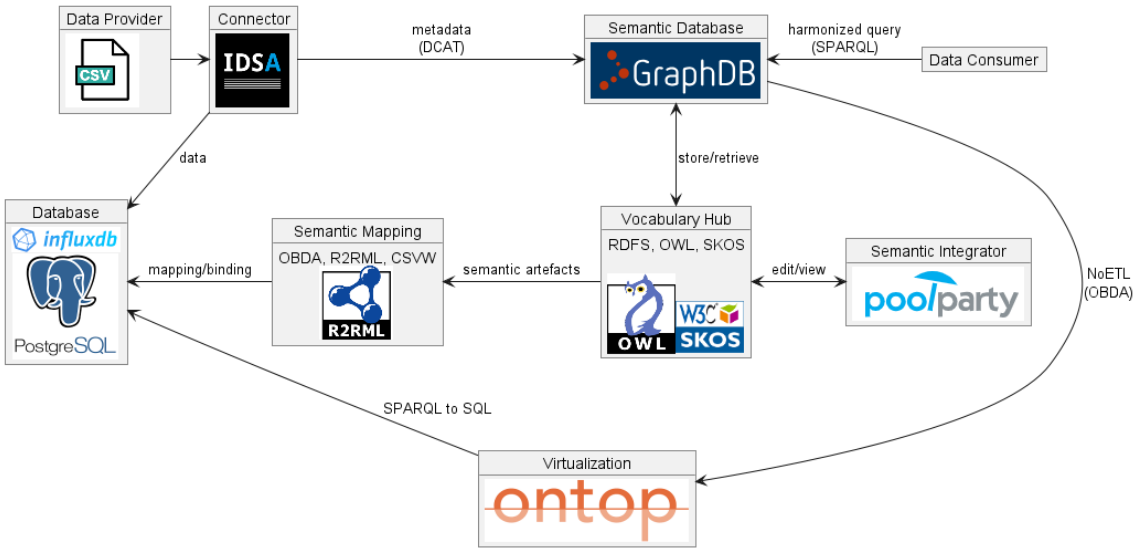
Vocabulary Hub Services: - RDF based: both GraphDB and PoolParty store and manage data based on the RDF data model. RDF is the basic building stone and underlying vocabulary definitions. - Scalable vocabulary storage: cluster implementation for higher throughput and a fault tolerant architecture. - Knowledge graph management: easy management of vocabularies using a rich web user interface and integration via Web APIs. Vocabularies can be created and edited based on standards like RDFS, OWL and SKOS, and existing vocabularies can be reused and combined to fit specific dataspace needs. - SPARQL service for semantic lookups: dereference URIs describing data assets to retrieve a semantic description of the entity from the graph database.

Semantic Layer Services support both unstructured and structured data assets: - Unstructured data: semantic annotation services for documents based on concept tagging. PoolParty's KG-based NLP primarily uses SKOS taxonomies to identify concepts in documents, but can be extended with OWL ontology elements (classes, properties) for increased expressiveness. - Structured data: we support structured data harmonization methods, including semantic mapping and linking SKOS concepts - Ontology-Based Data Access (OBDA) is used for virtualization of relational data, i.e. accessing relational data as semantic data. It is well suited for semantic querying of large amounts of data, while keeping the original relational data in place (NoETL approach). Ontologies are leveraged to construct the OBDA mapping; they are governed via the Semantic Layer and can harmonize data assets of various formats using a precise mapping.

Furthermore, it can be used to establish a fine-grained access control, thereby deciding for individual data elements if they should be exposed ot not. - RDB to RDF Mapping Language (R2RML) [3] and RDF Mapping Language (RML) [1] are intended for data transformation processes within a dataspace, where the original source is used as a basis to create an RDF representation (ETL approach). - GraphDB supports both OBDA and R2RML through its integration of ONTOP [2]. The same mappings can be used for NoETL virtual data access, or for ETL (to materialize the semantic data). - CSV on the Web (CSVW) [4] is a set of W3C standards that allow detailed semantic description of tabular data using a JSON "manifest" file. - SKOS concepts are used for structured data in several ways. Concepts can be directly added to RDF data via semantic links (properties from an ontology), they can be used to annotate specific literals of an RDF graph, or they can be used to associate documents with (specific parts of) an RDF graph. - Inference services: we provide metadata expansion services based on interlinked vocabularies in combination with ontologies. This inference approach can be used with semantic tagging for unstructured data, as well as the structured data approaches. - Inference tagging is provided by a PoolParty tagging service in combination with an expansion query for implementing the inference. The query can be configured to work for various scenarios. - Vocabulary crosswalks are a data modeling approach using interlinks to relate: - Different vocabularies from similar or the same domain that cover different areas or aspects of the domain. By relating concepts with each other using 1:1, 1:n and m:n relations, we can automatically expand queries and annotations by traversing the interlinked graph. - Different versions of the same vocabulary to automatically translate between those versions. - Metadata inference based on interlinked vocabularies is similar to crosswalks, but intended to represent different levels of abstraction, so that we are able to infer from a more general vocabulary to a more specific vocabulary and vice versa. For example, we can start with DCAT themes vocabularies and link to further vocabularies, which are more specific for certain areas or which go into more detail.

4.2 Software Architecture

The following figure shows the architecture of our approach:



- A Data Provider sends data through the Connector, and at the same time provides basic metadata (DCAT)
- Selected data can be stored in Database(s) managed by the dataspace, for easier access and querying.

  - For large-volume time series data, we recommend to use InfluxDB, PostgreSQL, Google BigQuery, etc.
  - Of course, access control is very important here. We'll leverage GraphDB's Fine-grained Access Control, but access control in the time series database is also important.
  - Depending on Provider preferences and agreements, other data may stay at the source, and be delivered to the Consumer only dynamically through the Connector.
- The Vocabulary Hub stores relevant semantic assets: ontologies and thesauri. These are added based on relevant use cases and datasets, following a dataspace governance process.
- The Vocabulary Hub also stores mapping assets, such as OBDA or R2RML mappings and CSVW manifests.
- PoolParty Semantic Integrator is used to edit, view and manage the semantic assets, and to bind mapping assets to incoming data.
- For virtual access to relational data, ONTOP is used to translate SPARQL to SQL dynamically and to convert returned data to SPARQL result format.
- The Data Consumer benefits from:
  - More powerful dataset discovery by using richer semantic queries, e.g.
    * "Give me all time series related to compressor C123" (equipment instance)
    * "Give me all time series related to temperature" (measured quantity)
    * "Give me all time series of bearings" (equipment kind or machine part)
  - Harmonized querying of datasets from multiple providers.

5 CONCLUSION AND FUTURE WORK

We presented an approach for dynamically binding datasets to semantic descriptions in a dataspace's Vocabulary Hub, therefore facilitating data harmonization and easier data consumption. We have implemented a first prototype of our approach, which builds on the integration of GraphDB and PoolParty. As next steps, we will bring this solution into use cases and into broader discussion to gain insight and develop it further.

The next steps include: - Bringing it into actual dataspaces in practice, representing and integrating datasets from different dataspace participants. - Implementing use cases for richer discoverability, harmonized querying and support for different content types of structured and unstructured data. - Explore how ML can benefit regarding quality in practice when providing consolidated and cleaned data via dataspaces. - Discuss how we can extend the IDS-RAM with services to improve the support for semantic interoperability provided by our solution.

6 ACKNOWLEDGEMENTS

This work is partially supported by the Digital Europe programme project UNDERPIN (grant agreement 101123179)

REFERENCES

[1] 2024. RML Introduction |. https://rml.io/docs/rml/introduction/.

[2] Timea Bagosi, Diego Calvanese, Josef Hardi, Sarah Komla-Ebri, Davide Lanti, Martin Rezk, Mariano Rodríguez-Muro, Mindaugas Slusnys, and Guohui Xiao. 2014. The Ontop Framework for Ontology Based Data Access. In *The Semantic Web and Web Science*, Dongyan Zhao, Jianfeng Du, Haofen Wang, Peng Wang, Donghong Ji, and Jeff Z. Pan (Eds.). Springer, Berlin, Heidelberg, 67–77. https://doi.org/10.1007/978-3-662-45495-4_6

[3] Souripriya Das, Seema Sundara, and Richard Cyganiak. 2012. *R2RML: RDB to RDF Mapping Language.* W3C Recommendation.

[4] Swirrl. 2024. CSVW - Standards. https://csvw.org/standards.html.