Federated Vocabulary Hubs as a Building Block for Semantic Layers in Data Spaces

Robert David[1,2], Vladimir Alexiev[3], Petar Ivanov[3], Wouter van den Berg[4], Jan Pieter Wijbenga[4] and Michiel Stornebrink[4]

[1]*Semantic Web Company GmbH, Austria*

[2]*Vienna University of Economics and Business, Austria*

[3]*Ontotext AD, Bulgaria*

[4]*TNO, The Netherlands*

Abstract

This paper proposes a layered architecture for federating vocabulary hubs, the component within a single data space responsible for shared agreement on semantics. Until now, each data space has their own vocabulary hub as prescribed by the IDS-RAM. The proposed design addresses the vocabulary hub federation problem. Moreover, it enables a semantic layer that provides semantic services over these federated data spaces. The European Commission's recent standardization request for a European Trusted Data Framework underlines the relevance of this work. This paper discusses relevance, requirements, like scalability and data sovereignty, and a layered architecture that integrates IDS and Semantic Web standards. Existing tool implementations are discussed that implement parts of the architecture and that can form a basis for future implementation of the full architecture for this semantic layer over data spaces.

Keywords

federation, dataspaces, vocabulary hub, vocabularies, semantic layer, semantic interoperability

This paper integrates with the W3C Dataspaces Community Group by mainly addressing the generic dataspaces issue https://github.com/w3c-cg/dataspaces/issues/10. It also addresses the issues https://github.com/w3c-cg/dataspaces/issues/2 for data discovery, https://github.com/w3c-cg/dataspaces/issues/6 to enable semantic metdata for data (payloads) and https://github.com/w3c-cg/dataspaces/issues/8 for lifecycle management and effective reuse of knowledge.

1. Introduction

Data space (DS) [1] architectures, like defined by IDS in the IDS-RAM[1] are designed for distributed decentralized data sharing. To make best use of data in a DS, data assets require interoperability. While this is very well solved on the syntactic side by established standards, like JSON[2], it still remains a challenge on the semantic side. To cope with this challenge, IDS defines vocabulary hubs[3] as central components for DSs, which provide standardized semantic descriptions to describe metadata and data of shared assets. IDS defines these vocabularies to be machine-readable and to use RDF as the data model for representation. Vocabulary hubs host, maintain, publish, and document vocabularies and make them available to all DS participants.

However, having one central component comes with challenges to govern and operate vocabularies in a DS. In this paper, we propose an architecture for federated vocabulary hubs, which address these

[1]https://docs.internationaldataspaces.org/knowledge-base/ids-ram-4.0

[2]https://www.json.org/

[3]https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_6_vocabulary_hub

challenges, and provide a flexible and extendable approach to scale vocabulary and interoperability services, while preserving sovereignty for governing parties. Our work builds on earlier research on federated vocabulary hubs[2] and extending their role for semantic layers[3], and combines industry expertise from TNO and Graphwise (the merger of Semantic Web Company and Ontotext) to expand the architectural vision with concrete implementation experiences from both open-source and commercial vocabulary hub platforms.

This work is especially timely because it aligns with the European Commission's recent standardization request for a European Trusted Data Framework. Our proposed architecture for federated vocabulary hubs directly addresses the need for "an implementation framework for trusted ontologies and data models" as outlined in the standardization request to CEN/CENELEC. This framework is currently being developed under the newly established CEN/CLC JTC 25 'Data management, Dataspaces, Cloud and Edge' technical committee[4] .

1.1. Use Cases and Requirements from Projects

We investigate use cases of EU-funded DS projects to identify specific requirements for vocabulary hubs.

DataBri-X - Legal Data Space   The Horizon Europe project DataBri-X[5] looks into three use cases where DSs provide value and enable to operate data management and processing tools in a distributed and sovereign environment. One of these use cases is the Legal Data Space, which aims to develop a preliminary version of a European DS, which fulfills legal requirements related to AI and which generally benefits European industry. The following three scenarios were identified in the project:

- Creation of a European Legal Data Space Nucleus.
- Enrichment and analysis of Legal Data.
- Addressing legal requirements on AI and data.

All these challenges require governed legal vocabularies, which are commonly shared among the users, but where authors retain sovereignty, to enable clear and unambiguous semantic descriptions to address legal requirements.

UNDERPIN - Time Series Schemas   The Digital Europe project UNDERPIN[6] develops a DS for manufacturing to drive mainly predictive maintenance on refineries and wind farms. The data sources of UNDERPIN provide time series data for training machine learning algorithms. on consolidated data. For consolidation of this data, different schema mapping and transformation approaches can be used [4]. Schemas created for this purpose are created to give semantic meaning to time series data and enable integration. Often the meaning of data is difficult to determine from the data itself, so schemas are provided by experts, which are often the providers of data, to be shared and used by participants in the DS. However, the sovereignty and governance of the schemas should remain with these experts.

Digital Product Passports (DPP)   The European DPP initiative is one of the most urgent use cases for federated vocabulary hubs. Products often need to conform to multiple passport templates based on their product type(s), which requires the integration of various semantic models and vocabularies. DPPs of physical products can bring to light the life cycle of the product and its compliance to relevant regulations, whereas DPPs of software products, as employed in the UNDERPIN project, can inform about licensing and the conformity to standards. A federated vocabulary search capability would help manufacturers and regulators identify and implement the right combination of passport templates for specific products.

---

[4]https://standards.cencenelec.eu/dyn/www/f?p=205:7:0::::FSP_ORG_ID:3485479&cs=1EF27AE97B5DBDA9B990D3DAF8BD63366
[5]https://databri-x.eu/
[6]https://underpinproject.eu/

All these use cases show the need for vocabularies in a broader scope which we aim to cover with our proposed architecture for federated vocabulary hubs. The following section describes the specific challenges identified for vocabulary hubs.

1.2. Challenges for Vocabulary Hubs

Vocabulary sovereignty A key principle for DSs is the sovereignty of data (owners) to decide how and when their data is shared and used. The same principle applies to vocabularies as well. Usually, there is a governing party responsible for development and maintenance of a vocabulary. They need to decide how and when they expose vocabularies to the DS and change requests from outside need to be governed by them as well. Exposure still stays as defined by the IDS-RAM and every DS participant can access published vocabularies and retrieve semantic descriptions.

Vocabulary governance Another important aspect is the governance of vocabularies. Besides having an internal working process and only exposing vocabularies once they are ready to be shared and used, we face the challenge of versioning of published vocabularies. DS participants should be able to use specific versions of a vocabulary and they should be supported with services when switching versions to help them manage the changes that come with these different versions.

Vocabulary findability A significant challenge for vocabulary hubs is ensuring that valuable vocabularies can be discovered across different data spaces. If a data space participant is not aware of the existence of an important vocabulary that happens to be maintained in a different data space, they might miss opportunities for semantic interoperability or end up recreating similar vocabularies. Federated vocabulary hubs can help address this challenge by enabling participants to discover and access vocabularies from other hubs as if they were present in their own hub.

Scalability & High Availability Scalability and high availability are two important technical attributes for vocabulary hubs. As central components, they need to provide a high uptime in order for the DS to use their services. Scalability might also be an issue because of large data sets or load of access of DS participants, depending on the use case.

Extending the Vocabulary Ecosystem Another challenge is to make vocabulary hubs easily extendable. Partially this is covered by scalability on the technical side. However, on the data governance side, this is a different challenge. We want to easily and transparently extend the vocabulary ecosystem and introduce new governing parties and their vocabularies without changes to the existing ones.

Services Finally, exposing vocabularies to a DS is an important part to achieve semantic interoperability, but often this requires additional services, like data integration, mapping, or reasoning services, to fully leverage their potential. Such services should also be able to build on a scalable and sovereign vocabulary infrastructure and provide transparent access to DS participants.

2. Preliminaries

Preliminaries include the IDS architecture for DSs, which defines vocabulary hubs, the DSSC blueprint, the concept of linked data to create federated data structures, and technologies like SPARQL to implement federated use of these structures.

2.1. IDS Data Spaces

The International Data Spaces Association (IDSA)[7] is an organisation which aims to standardize DS architectures for sovereign data sharing to drive the digital economy in Europe and beyond. IDSA publishes the IDS Reference Architecture Model (IDS-RAM)[8], which defines the architecture of a DS and the components it uses. One of these components is the vocabulary hub[9], which is a central service for providing standardized vocabularies to all DS participants to enable semantic interoperability. The IDS vocabulary hub is defined on a high level and does not go into details of governance for vocabularies or even federation. Federation is currently not supported for any services in IDS, which limits the vocabulary hub, being a central component, regarding flexibility, scalability, and sovereignty of vocabularies.

2.2. DSSC blueprint

The Data Space Support Centre (DSSC)[10] blueprint provides guidelines for data interoperability in data spaces via vocabulary services, which ensure consistent use of common data models based on semantic standards within a data space. The blueprint also suggests linking data sets to their corresponding data models through the use of the Data Catalog Vocabulary (DCAT)[11], which is the W3C Recommendation for describing data sets and services using RDF.

2.3. RDF and (inter)linking

The Semantic Web [5] is a set of standards for knowledge representation based on the World Wide Web. The basic data model, defined by the Resource Description Framework (RDF), is a knowledge graph, where nodes and edges are represented as resources on the Web. Information about resources can be retrieved via the URI, and one can follow edges, which are hyperlinks, between nodes to navigate this web of data. Furthermore, by creating links to published resources, it can be expanded in a decentralized and open way.

2.4. SPARQL federation

SPARQL[12] is the query language for RDF data. It can be used to query distributed data sources exposed via SPARQL endpoints. These endpoints need to conform to the defined W3C recommendation to be interoperable, but otherwise they can be driven by databases, which store RDF natively or by services, which map non-RDF to RDF and expose it in the context of queries. SPARQL queries for graph patterns, supports mandatory and optional patterns, conjunction, disjunction, aggregation, negation, and many more capabilities. SPARQL endpoints can support federation[13] (since SPARQL 1.1), which enables them to delegate parts of a query to other SPARQL endpoints and thus query the web of data as a distributed knowledge base.

3. An Architecture for Federated Vocabulary Hubs

Vocabulary hubs, as defined by IDS, represent central DS components, which host and expose vocabularies. These vocabularies are intended to be used by participants for common semantic descriptions with the aim of interoperability. However, this architecture limits the DS to a centralized approach and does not address the challenges identified above. Therefore, we propose an extended architecture,

---

which is built on top of the IDS-RAM, and which implements a decentralized approach to enable federated vocabulary hub services. DS participants can use any hub in the federation and hubs can also be used across multiple DSs. The architecture aims to address the challenges described above and provide concrete technical solutions. On the technical level, we reuse existing standards, like SPARQL federation. We call this decentralized architecture consisting of multiple loosely coupled vocabulary hubs a *vocabulary hub ecosystem*.

In the following, we introduce our proposed architecture for federated vocabulary hubs. We build this architecture incrementally by starting with a i) basic architecture built on standards, then expanding it towards a ii) service-oriented architecture, and then finally concluding with a iii) Semantic layer built on a federated vocabulary hub ecosystem.

3.1. Connecting the Nodes - A basic Architecture

As described above in section 2.3, RDF can be used in a flexible manner to link together graphs on the Web. This open approach works decentralized, and there is no central or governing authority which needs to establish or confirm such links. Using SPARQL, query endpoints expose the vocabularies and by enabling federation on the query engines we can implement a basic vocabulary federation architecture purely built on Semantic Web standards.
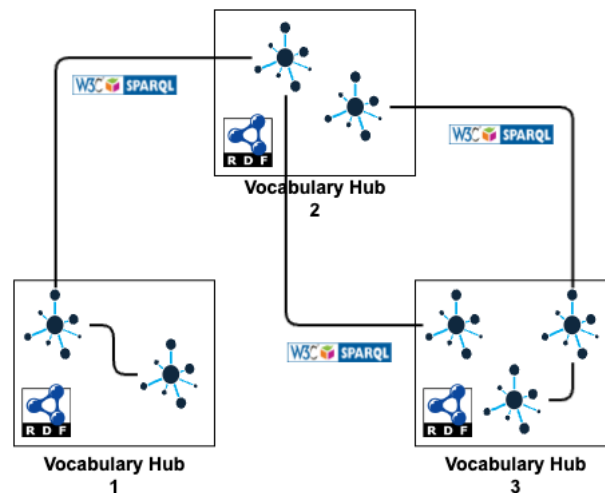
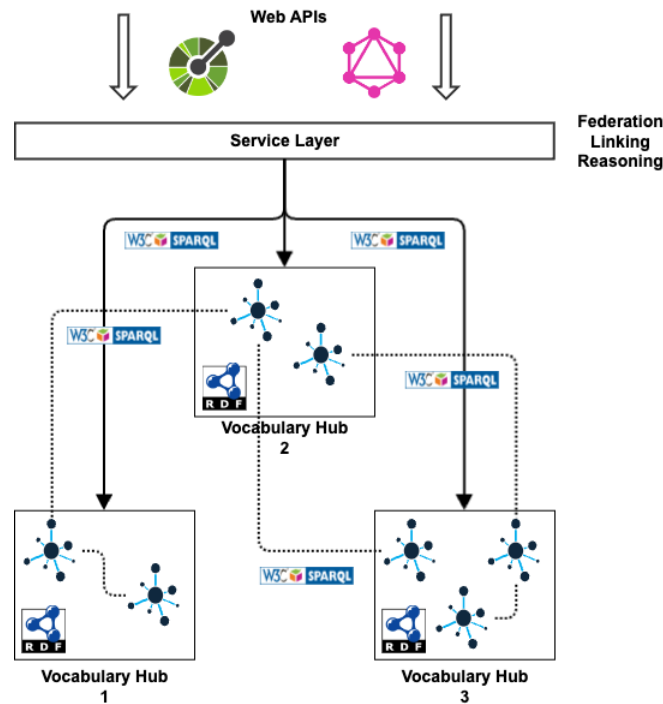**Figure 1:** Federated Vocabulary Hub - Basic Architecture.

Figure 1 shows the basic architecture. Data is exposed as RDF via SPARQL query endpoints, which support federation. Links between different vocabularies can be established by referring to resources in vocabularies, which can be hosted locally or remotely, and can be retrieved by query federation. With this basic architecture, we implement vocabulary federation based on existing standards, i.e. W3C recommendations.

The architecture naturally addresses findability via linked resources between vocabularies. It addresses the challenge of vocabulary sovereignty, because vocabulary hub operators manage their hosted vocabularies and decide for exposure, while still being able to interconnect to vocabularies hosted elsewhere via links. It also partially addresses the challenge of vocabulary governance, because the exposed vocabularies, being RDF data, can use common practices for web service versioning, e.g. providing the version number as part of the IRI. Scalability is also addressed by this decentralized architecture, which allows for high flexibility in building the hosting ecosystem. Finally, extendability is also provided because of the flexibility of RDF to interconnect vocabularies independent of the hosting location.

Our basic architecture addresses many challenges well. However, there are still gaps regarding governance and extension. We also aim to add services to support easy integration and use of vocabulary data on a higher abstraction level than SPARQL endpoints provide.

3.2. Added Value via Federation Services

To expand the basic architecture, we introduce a service layer on top of the technical components, like the SPARQL query endpoint. These services abstract the technical details. They expose web services via commonly used standards for integration, like OpenAPI[14], JSON[15] and GraphQL[16] and they provide a facade for transparent access to the vocabulary ecosystem. Specific details of the federation are hidden, and accessing one vocabulary hub node provides users with access to find and use the full set of vocabularies of the ecosystem. Also, services can manage challenges of having different versions of vocabularies in in different hubs. Such challenges arise when using service endpoints for exposure instead of linked data. Furthermore, services, like reasoning over combined data sets, are provided.

Figure 2: Federated Vocabulary Hub - Added Services.

This architecture improves vocabulary governance, because services can decide for the functionality they expose. They can collect, pre-process, and transform RDF data sources and present them in a way that is convenient for integrators to consume. Extendability is better addressed by services, because the service layer makes accessing the vocabulary hub ecosystem transparent. Finally, we can add arbitrary services on top of RDF vocabularies to provide functionality which cannot be covered by SPARQL endpoints and RDF directly. These services raise the role of an IDS vocabulary hub [3] to provide value-added services, which can solve challenges of semantic interoperability within data spaces.

## 3.3. The Vocabulary Hub Ecosystem as a Building Block for Semantic Layers

As the final step in our architectural evolution, we discuss how the vocabulary hub ecosystem relates to the principle of semantic layers. Vocabulary hubs drive semantic interoperability. In the context of data spaces, this primarily means semantic interoperability of metadata of assets. However, in many scenarios of data integration or consolidation, we require semantic interoperability at the data level to make best use of distributed, and possibly heterogeneous, data sources. With federated vocabulary hubs, we provide an essential building block for a unified semantic view on data via vocabularies.

---

[14]https://github.com/OAI/OpenAPI-Specification
[15]https://www.json.org/
[16]https://graphql.org/
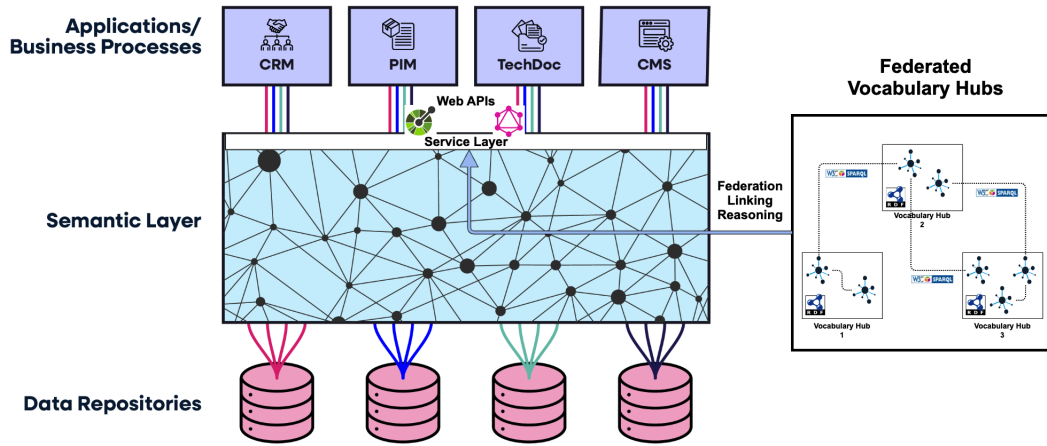
Figure 3: Federated Vocabulary Hubs - A Building block for Semantic Layers.

The connection between vocabulary hubs and data space catalogs is crucial for enabling semantic interoperability at the data level. If every data set or data service catalog entry contains a link to its semantic specification in a vocabulary hub (e.g. through the use of dcterms:conformsTo in DCAT), it becomes possible to query across the catalog and discover relevant data assets based on their semantic descriptions. Data consumers can then search for data sets that use specific concepts from known vocabularies. Additionally, semantic relationships between different vocabularies can be used to expand searches across related concepts (e.g. through relations like owl:equivalentProperty).

In this way, semantic layers drive business processes and associated applications and thereby enable enterprises to make best use of their data sources. Figure 3 shows how a semantic layer semantically connects isolated data repositories and provides a unified semantic view on data as input to applications and business processes.

4. Implementing a Vocabulary Hub Ecosystem

In the following, we present software products that are already used as vocabulary hubs in data spaces. We describe how they can be used to implement the previously described federated vocabulary hub architecture.

4.1. Semantic Treehouse

Semantic Treehouse[17] is an open-source[18] vocabulary hub implementation that provides comprehensive management of shared data models while supporting federation through DCAT integration. The platform implements support for multiple abstraction layers of data models, ranging from basic vocabularies to complex ontologies, application profiles, and technical schemas. Through its adoption of DCAT, it enables standardized vocabulary exchange, making data models discoverable and accessible across different data spaces. Collaborative governance is supported with built-in version control, issue tracking, and review processes that facilitate community-driven vocabulary development and maintenance, and fine-grained access control enables sovereignty in a federated ecosystem.

The platform generates multiple technical artifacts from semantic models. These include JSON Schema, XML Schema, OpenAPI, and RDF/SHACL shapes. This allows data space participants to implement the vocabularies using familiar technologies while preserving semantic consistency.

Semantic Treehouse demonstrates how vocabulary hubs can evolve from simple repositories to active participants in a federated ecosystem. It has been in development and use since 2016. Data space

---

[17]https://www.semantic-treehouse.nl
[18]https://gitlab.com/semantic-treehouse

projects that have applied Semantic Treehouse include Enershare[19], ZeroW[20] and CIRPASS 2[21], among others. The implementation of DCAT has been the first step towards the vision of a decentralized vocabulary ecosystem that enables semantic interoperability across data spaces.

Semantic Treehouse facilitates federation through the uniform exports in DCAT of each vocabulary hub instance through an API endpoint. Merges of content can be achieved manually through unions of the triple data coming from different Semantic Treehouse environments, e.g. by loading multiple DCAT exports into a triple store. Future work includes exposing this knowledge graph by means of an SPARQL endpoint, thereby offering a query and viewing facility that allows searching for vocabularies based on keywords, descriptions, or content.

4.2. PoolParty Thesaurus Server

PoolParty Thesaurus Server[22] – part of the novel Graphwise Platform[23] – provides management of taxonomies and ontologies for AI applications. Taxonomies and ontologies are developed using SKOS [24] and OWL_{strict} [6], an OWL[25] subset for unambiguous semantic descriptions, and can be exposed via Web APIs, as linked data on the Web and via SPARQL endpoints to create a federated graph data architecture. Regarding the software architecure of Graphwise platform, PoolParty is backed by the Ontotext GraphDB[26] graph database, which is highly scalable, supports a cluster architecture and provides ACID compliance, thereby fulfilling scalability & high availability requirements. It also features SPARQL 1.1 support and reasoning capabilities.

With these features, PoolParty provides several services to implement a vocabulary hub with federation support. The basic architecture is covered by exposing a SPARQL endpoint with enabled federation and by publishing vocabularies and ontologies via linked data on the Web. By supporting these means of publication, PoolParty fulfills the role of an IDS vocabulary hub. With added services, we develop it towards a federated vocabulary hub node supporting semantic layer strategies. PoolParty is designed as a middleware and provides a RESTful Web API for vocabulary and ontology management functionalities, which makes it easy to integrate with other services. Finally, PoolParty offers various capabilities around data consolidation and integration, like ETL and schema mapping features, which makes it fully support semantic layer strategies.

4.3. Complementary Approaches to Vocabulary Hub Implementation

While Semantic Treehouse and PoolParty Thesaurus Server have slightly different technical approaches to vocabulary management, they share architectural principles that align with the federated vocabulary hub vision. Both platforms adopt and adhere to open standards, like XML, JSON, SKOS and RDFS/OWL, so they are interoperable by design. They are both developed by organizations that bridge research and practice, which means the use of state-of-the-art semantic technologies is combined with practical applicability in real-world data space implementations. Indeed, both tools have successfully been deployed in data space projects, as described in section 1.1.

5. Conclusions

In this paper, we presented an architecture for federated vocabulary hubs as an evolution of IDS vocabulary hubs. We identified major challenges for vocabulary hubs in data spaces and discussed how we can address them with the proposed architecture. We concluded with a presentation of existing

---

[19]https://enershare.eu/
[20]https://www.zerow-project.eu/
[21]https://cirpass2.eu/
[22]https://www.poolparty.biz/
[23]https://graphwise.ai/
[24]https://www.w3.org/2009/08/skos-reference/skos.html
[25]https://www.w3.org/TR/owl-features/
[26]https://www.ontotext.com/products/graphdb/

software products, which form building blocks to implement the proposed architecture. The federated vocabulary hub architecture is the first step in the evolution of IDS data spaces towards implementing decentralized vocabulary hub ecosystems.

The key to vocabulary hub federation lies in a) the adoption of DCAT as means to catalog vocabularies, and b) exposing the catalogue of vocabularies as a knowledge graph by means of a SPARQL endpoint. This loosely couples vocabulary hubs, links vocabularies and allows performing federated SPARQL queries. Thereby enabling the functionalities of the service layer as described in the architecture of the vocabulary hub ecosystem.

The timing of this work is particularly relevant given the European Commission's recent standardization request for a European Trusted Data Framework. Our proposed architecture for federated vocabulary hubs provides a foundation for implementing the "trusted ontologies and data models" framework required by this request; it specifically calls for technical specifications to "specify criteria for the selection of semantic assets" and "specify methods for the semantic annotation of shared data". Our federated vocabulary hub architecture directly supports these requirements and positions our work as a potential building block for the standards being developed under CEN/CLC JTC 25.

Additionally, when we look forward, we see that federated vocabulary hubs will play a crucial role in supporting AI-powered systems, including impacting semantic interoperability via GraphRAG and agentic AI. These systems will leverage the federation of vocabulary hubs to automatically discover and adopt the most appropriate semantic models for specific use cases.

Our next steps include to incrementally implement the proposed architecture based on the presented software products, to apply it to (further) use cases, and to contribute to standardization for data spaces. Following our architectural proposal, we will expand the services of Semantic Treehouse and PoolParty Thesaurus Server accordingly and enable federation support for building vocabulary hub ecosystems to be tested in real-world projects. Furthermore, we will contribute with standardization proposals for the federated architecture to extend IDS.

Acknowledgements

References

[1] B. Otto, M. Hompel, S. Wrobel, Designing Data Spaces: The Ecosystem Approach to Competitive Advantage, Springer International Publishing, 2022.

[2] J. Bootsma, J. P. Wijbenga, L. Oosterheert, M. Stornebrink, W. van den Berg, Establishing semantic interoperability across data spaces: a solution for sharing vocabularies, Technical Report, TNO, 2024. URL: https://coe-dsc.nl/knowledge-base/original-content/deliverables/.

[3] R. David, P. Ivanov, V. Alexiev, Raising the Role of Vocabulary Hubs for Semantic Data Interoperability in Dataspaces, in: Third workshop on Semantic Interoperability in Data Spaces, Budapest, Hungary, 2024. URL: https://raw.githubusercontent.com/underpin-project/papers/refs/heads/main/EBDVF-2024/EBDVF-2024-presentation/presentation.pdf.

[4] M. Andresel, V. Siska, R. David, S. Schlarb, A. Weißenfeld, Adapting ontology-based data access for data spaces, in: The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference, May 26–27, 2024, Hersonissos, Greece: SDS 2024 co-located with ESWC 2024, 2024.

[5] O. Lassila, J. Hendler, T. Berners-Lee, The Semantic Web, Scientific American 284 (2001) 34–43.

[6] R. David, A. Ahmeti, S. Ahmetaj, A. Polleres, OWL_{strict}: A Constrained OWL Fragment to avoid Ambiguities for Knowledge Graph Practitioners, in: The Semantic Web: 22th International Conference, ESWC 2025, Portorož, Slovenia, June 1 – June 5, 2025, Proceedings, 2025.