The Vocabulary Hub as a Catalog for Semantic Artifacts for Discovery and Alignment of Datasets

Ruben Dedecker*, Julián Rojas Meléndez and Pieter Colpaert

IDLab, Ghent University – imec, Technologiepark-Zwijnaarde 126, 9052 Gent, Belgium

Abstract

The EU Common Data Spaces initiative aims to enable secure, sovereign and interoperable data sharing across organizational and national boundaries. However, the high heterogeneity of underlying data models and formats, prevents semantic interoperability from being realized. Publishers can address this challenge by exposing their internal knowledge by adopting continuous publishing models that reduce operational overhead for both publishers and consumers. Yet for data consumers, costly alignments still remain a necessity when the semantics of published datasets differ from their expected internal data models and schemas. Data spaces require mechanisms to define, discover, and govern such alignments throughout their entire lifecycle, enabling eventual interoperability. In this paper, we show that considering additional semantic artifacts as part of the vocabulary hub, namely *dataset profiles* defining structural and semantic constraints, and *profile alignments* (e.g., in the form of SPARQL construct queries), could provide consumers with a semantic entry point for dataset discovery and integration. We focus on the interaction patterns afforded by the additions of these semantic artifacts and provide a demonstrator implementation of a user interface that integrates this functionality. We validate our approach through a use case from the DeployEMDS project, focused the automatic discovery and alignment of traffic measurements. The extended vocabulary hub enables clients to discover datasets based on profile characteristics such as shapes, ontologies, and publishing data models, while also identifying available alignment pathways toward target consumer data models. It shows how the technical barriers for creating and relying on semantic alignments are lowered, enabling the consumption of data using the desired vocabularies and schemas. Future work will focus on integrating this component with existing data space connector implementations to further automate semantic interoperability by enabling semantic and profile-based content negotiation for data exchanges.

Keywords

Data Spaces, Vocabulary Hub, DCAT Catalog, Dataset Profile, Semantic Alignment,

1. Introduction

The EU Common Data Spaces initiative[1] initiative represents a shift towards a unified digital market where data can flow securely across borders and sectors in the European Union. This initiative aims to establish a trusted environment for data sharing that guarantees sovereignty over data resources. Central to this vision is the achievement of high-level data interoperability, which allows diverse stakeholders to share and access data across various strategic domains, including health, energy, and mobility [1].

Semantic interoperability is achieved when interacting systems attribute the same meaning to an exchanged data entity, ensuring consistent handling across systems regardless of individual formats [2]. In order to support interoperable data exchanges, current data spaces reference architectures, most notably, the IDS-RAM[2] from the International Data Space Association (IDSA), have devised components such as the Vocabulary Hub and the Metadata Broker. The Vocabulary Hub aims to facilitate the management and harmonize the shared understanding of domain-specific terms and ontologies by

Extended Semantic Web Conference (ESWC 2026): The Fourth International Workshop on Semantics in Dataspaces, May 10-11, 2026, Dubrovnik, Croatia

*Corresponding author.

ruben.dedecker@ugent.be (R. Dedecker)

https://www.rubendedecker.be/ (R. Dedecker); https://julianrojas.org/ (J. R. Meléndez); https://pietercolpaert.be/ (P. Colpaert)

0000-0002-3257-3394 (R. Dedecker); 0000-0002-6645-1264 (J. R. Meléndez); 0000-0001-6917-2167 (P. Colpaert)

[1] EU Common Data Spaces Initiative - https://digital-strategy.ec.europa.eu/en/policies/data-spaces

[2] IDS-RAM (version 4) - https://docs.internationaldataspaces.org/ids-knowledgebase/ids-ram-4

making their definition available in machine-readable formats (e.g., RDF-based) to all data space actors. The Metadata Broker provides a catalog service for discovering data resources via metadata descriptions, grounded on the IDS Information Model [3] and extensions thereof following standards such as DCAT(-AP) and ODRL.

Despite the IDS-RAM goal of providing a uniform protocol layer to enable these data exchanges, it does not automatically solve the semantic interoperability challenge stemming from the heterogeneity of underlying data models in available data resources [4]. In practical deployment scenarios, data resources are increasingly complex and frequently rely on multiple, sometimes overlapping vocabularies. This semantic heterogeneity makes it difficult to automate data interoperability across organizations, as current data space components lack the inherent logic to bridge the gap between disparate data models without significant manual intervention [5].

Regional initiatives such as the Open Standards for Linked Organizations (OSLO)[3] in Flanders [6] attempt to standardize the data models used by local stakeholders for the publishing of data. This approach could be a feasible solution for regional data exchanges, however contextual differences such as language or legislative frameworks, could prevent the direct adoption of these models in other regions, where local and specialized data models need to be define instead.

To tackle this data interoperability and integration problem in data spaces, *semantic alignments* among data models become a prime necessity to capture the global insights and translate the local details across contexts. Alignments could then become useful semantic assets for data space actors to facilitate the consumption of available datasets according to their internal data model and schema requirements. These introduces the need for proper governance of these assets, so that they can transparently serve all actors in the ecosystem, reduce duplicated work and enable collaborative management.

The introduction of semantic alignments as data space first class citizens, prompts for the definition of descriptive *dataset profiles* of published data resources, as a necessary base for consumers to create and reuse semantic alignments. These profiles define the underlying specifications, ontologies, and structural constraints of compliant data resources, enabling semantic alignment and improving search and discovery capabilities.

In this work we examine an expanded role for the vocabulary hub within dataspace architectures. Beyond the conventional role as a mechanism for managing and exchanging ontologies, we assess its ability to support richer artifacts in the form of dataset profiles that extend beyond ontology-level information, as well as the publication of semantic alignments that enable consumers to map datasets to their local data models.

We describe possible interactions patterns of the proposed extended vocabulary hub and its practical utility, through a real-world use case from the DeployEMDS project [4], focusing on the interoperability and reuse of traffic measurement data. This implementation illustrates how we can effectively lower technical barriers, allowing data consumers to discover and manage semantic alignments, that enable them to integrate disparate traffic data sources according to their own local schemas and functional requirements.

In section 2 we give an overview of related work on the role of vocabulary hubs in data spaces and dataset alignments. In section 3, we define the modeling of the semantic artifacts for the dataset profile, documenting the structural and semantic constraints to which the dataset entities adhere, as well as the alignments of datasets from their source to a target profile. In section 4 we present the implications of these artifacts on the architecture, and the interactions generating and processing these artifacts for publishing, searching and aligning datasets to local data models. section 5 presents the use-case, for which the demonstrator implementation is explained and evaluated in section 6. Finally, section 7 and section 8 respectively provides our conclusions and future direction of the proposed extension to the vocabulary hub.

[3]OSLO is the data working group of Flanders - https://data.vlaanderen.be/

[4]https://deployemds.eu/

2. Related work

The Vocabulary Hub is a core component of the data space architecture, both as a part of the International Data Spaces Reference Architecture Model (IDS-RAM) [7] as well as in the Data Spaces Support Center (DSSC) blueprint[5] for the Common European data spaces. Its designated role is to host, maintain, publish, and document the ontologies used in the dataspace[6].

In previous work on vocabulary hubs for semantic interoperability, David et al. emphasized the importance of mappings datasets to semantic formats, for which the vocabulary hub can play a role in the storage of these semantic artifacts, including used vocabularies and performed mappings using the Poolparty Semantic Suite[7]. We second the conclusion that the vocabulary hub should support the cataloging of mappings from semi-structured data resources to semantic formats, which would enable alignments to be materialized external to the data publisher. However, we argue to further include such alignments as managed semantic artifacts, to aid in the governance and discoverability of dataset conversion mechanisms that bridge structural differences of data models built on different ontologies and specifications. Catering for both these features, could enable data spaces to provide a broader integration of semantically rich data and further automating interoperability.

In follow up work, David et al. defined challenges for vocabulary hubs as a core building block for the semantic layers in data spaces, emphasizing the importance of flexibility in collaboration, uniformity of interfaces and standardization of discovery and lookup. They indicate a requirement for closer linking between the semantic level of the vocabulary hubs, and the data level of the catalogs and metadata brokers. Strengthening further the relation between semantic artifacts published by the vocabulary hub and the related data resources listed in the data space catalogs, could lead vocabulary hubs to take the role of a semantic entry point towards automating discovery and integration of datasets based on their linked semantic artifacts.

Work by Bootsma et al. advocates for the use of DCAT as an interoperability layer for cataloging published semantic artifacts detailing used ontologies and enabling collaboration between data spaces with diverging implementations of the vocabulary hub. These catalog entries may provide metadata about the published ontologies, yet they fall short on describing in detail the content of data resources (e.g., constraints).

Data spaces tackle the technical interoperability challenges needed to exchange data resources between participating actors via the data space protocol [8]. But for actual data resources to be integrated in practice, consumers must be able to align the data models used by dataset entities both on a structural and semantic level to their internal data model. The vocabulary hub definition currently restricts itself to the governance of terms and ontologies within a data space, limiting the potential for alignments to ontological alignments, mainly taking the form of OWL, RDF or SKOS mappings between ontology definitions [11].

Multiple implementations exist which fulfill the functional requirements defined by the IDS-RAM for the Vocabulary Hub. As shown in Table 1, the available tools extend on the core requirements of vocabulary maintenance and term discovery by including additional functionalities, in the form of systems of governance, collaborative editing, search and discovery, and alignments between ontologies or supported data modeling approaches. The alignment capabilities of these implementations focus primarily on semantic-level alignments through SKOS mapping properties, OWL/RDFS equivalence assertions, and similar mechanisms for establishing conceptual correspondences. However, they cannot bridge extensive structural differences between particular uses of data models, as they do not specify how to transform data representations from one schema to another.

[5]The Data Spaces Support Center (DSSC) includes the concept of a vocabulary hub in their blueprint - https://blueprint.dssc.eu/?pane=technical&technical=how-a-data-plane-and-control-plane-work-together

[6]IDS-RAM defines a vocabulary hub component - 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

[7]The Poolparty Semantic Suite - https://www.poolparty.biz/

[8]https://eclipse-dataspace-protocol-base.github.io/DataspaceProtocol/2025-1/

PoolParty and Semantic Treehouse extend on this, by providing capabilities for consumers in data spaces to transform datasets into specified data models, through mapping approaches such as RML [12]. Once mapped to an internal data model concept, Semantic Treehouse enables the mappings between their internal data model concepts using SKOS mappings [9], where the Poolparty suite conceptualizes all loaded data as a centralized knowledge graph representation, on which views can be generated using SPARQL Construct queries.

We follow the proposition made by David et al. [8], of storing mapping assets in vocabulary hubs, and further extend it with storage and management of alignment artifacts, grounded in concrete *dataset profiles*. This works steps towards expanding integration capabilities within data spaces, with the provision of necessary resources to automate the creation and execution of data transformation pipelines for semantic alignment.

Table 1
Comparison of Tooling for the Vocabulary Hub role in data spaces

Tool	Ontology models	Application Profiles	Alignment Options
VocBench[10]	SKOS/SKOS-XL, OWL/OWL2, RDFS, Ontolex-lemon	Hybrid OWL / SKOS models	SKOS mappings, INRIA Alignment, OWL/RDFS equivalence
OntoPortal[11]	OWL, RDFS, SKOS, OBO, UMLS-RRF	Ontology descriptions through uniform metadata model	SKOS mappings, interportal mappings, bulk JSON import
Semantic Treehouse[12]	OWL, RDFS, SKOS, SHACL	Message model combining semantic and structural constraints	Loading data into message models, SKOS mappings between message models
PoolParty[13]	SKOS (primary), OWL2, Custom Schemes	Custom Schemes for ontology subsets, SKOS/OWL integration	SKOS mappings, Linked Data alignments, reasoning-based inference

3. Artifact definitions and interactions

The data space architecture manages the technical alignment between actors necessary for the publishing, discovery, negotiation and exchange of data [7]. To ensure data consumers in a data space can integrate available data resources, they must be able to retrieve the published data and perform the alignments necessary to integrate the data in their local systems. For this, structural information such as schema and shape resources for both source and target data models become necessary to enable consistent definitions of alignments that can be reused within the data space. Additionally, semantic alignment needs to be ensured to prevent incorrect processing of structurally aligned data. Here, the availability of used vocabularies, ontologies, schema and specifications adhered to by a retrieved dataset provide a basis on which alignments can be defined and validated.

[9] Semantic treehouse is looking into adding mapping support according to their blogpost at https://www.semantic-treehouse.nl/blog/mapping-specifications/
[10] vocbench.uniroma2.it – Documentation: vocbench.uniroma2.it/doc
[11] ontoportal.org – Documentation: ontoportal.github.io/documentation
[12] semantic-treehouse.nl – Documentation: semantic-treehouse.nl/docs
[13] poolparty.biz – Documentation: help.poolparty.biz

3.1. The dataset definition

Datasets published on the data space must be indexed to enable for search, discovery and retrieval. This indexing is modeled as a catalog of dataset entries, as shown in Listing 1. For the IDS framework, this catalog is included in the IDS Information Model[14], for which it builds upon the W3C DCAT ontology [13] to model the catalogs describing the published datasets by data space connectors catalogs and metadata brokers. Within Europe, these catalogs are extended through specific DCAT Application Profiles (DCAT-AP)[15] that provide an interoperable basis for both data space initiatives[16] and (inter)national open data portals. In the DeployEMDS project, we build on the mobility application profile for DCAT[17].

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
<https://meta-boker.org/datasets/catalog> a dcat:Catalog ;
   dcat:dataset <https://iot.hamburg.de/v1.1/Datastreams(26110)> .

<https://iot.hamburg.de/v1.1/Datastreams(26110)> a dcat:Dataset ;
   dct:title "Hamburg traffic counting dataset 26110"@en ;
   dct:publisher <https://iot.hamburg.de/id/organisation/urban-data-platform> ;
   dct:conformsTo <https://docs.ogc.org/is/18-088/18-088.html> ;
   dcat:distribution <https://meta-boker.org/distributions/hamburg-stapi>.

<https://meta-boker.org/distributions/hamburg-stapi> a dcat:Distribution ;
   dct:conformsTo <https://w3id.org/ldes/specification> ;
   dcat:accessURL <http://193.190.127.148/ldes/> ;
   dct:format <https://www.iana.org/assignments/media-types/text/turtle>.
```

Listing 1: Dataset catalog entry for Hamburg traffic observations published as LDES

3.2. The dataset profile

The dataset metadata included in DCAT dataset definitions depends mainly on the application profiles adhered to in the spaces in which the catalog is defined. These application profiles tailor to the metadata requirements of specific domains as the basis of interoperability. Building on this, we define the concept of the *dataset profile* as a generic publishable artifact in data spaces, that defines the specifications to which a dataset (distribution) adheres. These serve as a flexible extension to the application profiles for DCAT dataset definitions, providing more in-depth structural, semantic and functional constraints through associated resource descriptions.

Using the Profiles Vocabulary [14], these artifacts model a compliance profile to a set of specifications, extended with comprehensive resource descriptors referencing used vocabularies, structural constraints, schema definitions, usage examples, and transformation specifications, covered in subsection 3.3, within a unified profile declaration.

This provides an additional layer of flexibility on top of the dataset descriptions that contrasts with the structural constraints of specific DCAT application profiles that enforce restrictive, predetermined schemas which limit expressiveness to a fixed set of metadata properties. Where the current approach of structural rigidity benefits existing integration flows, the flexibility provided by more extensive dataset profiles provides a broader basis for performing dataset alignments in data spaces, through direct linking of vocabularies, schemas and validation resources.

In addition, these profiles provide a more stable foundation for automated agents to perform data discovery and integration within data spaces. Similar to how the Model Context Protocol (MCP)[18]

[14]The IDS Information Model - https://w3id.org/idsa/core
[15]DCAT Application Profiles (DCAT-AP) - https://semiceu.github.io/DCAT-AP/drafts/latest/
[16]Semantic Interoperability in Data Spaces - https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-Semantic-Interoperability-in-Data-Spaces-V1.1-1-5.pdf
[17]Mobility extension for DCAT-AP - https://w3id.org/mobilitydcat-ap/drafts/latest/
[18]https://modelcontextprotocol.io/docs/getting-started/intro

embeds examples and contextual information directly within the protocol specification to enable autonomous agent reasoning, the dataset profiles provide new integration pathways, through evaluation of the resources linked in these profiles.

Although these profiles can be published in combination with the dataset metadata in the connector catalogs and metadata brokers of the data space, we make the choice to publish the dataset profile catalogs in the vocabulary hub component. The interlinking of the profiles to the relevant datasets can be handled by both the dataset catalogs published by connectors and metadata brokers with the profile catalog published by the vocabulary hub, for which we use the *dct:relation* property. The *dct:conformsTo* property is used to define the individual compliance of datasets and their profiles.

The inclusion of these artifacts in the vocabulary hub component is done both as a practical consideration of restricting the required extensions to a single component, making the existing profiles externally discoverable to prevent duplication of profile definitions and improve discoverability. However, we acknowledge that including these profiles in the connector or broker catalogs might be also a practical choice.

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix prof: <http://www.w3.org/ns/dx/prof/> .
@prefix role: <http://www.w3.org/ns/dx/prof/role/> .
<https://voc-hub.org/profiles/catalog> a dcat:Catalog ;
   dcat:resource <https://voc-hub.org/profiles/stapi-ldes> ,
              <https://voc-hub.org/profiles/oslo-verkeersmetingen> .

<https://voc-hub.org/profiles/stapi-ldes> a prof:Profile ;
   dct:title "SensorThings API LDES Profile"@en ;
   dct:publisher <https://urbandataplatform.hamburg> ;
   prof:isProfileOf <https://docs.ogc.org/is/18-088/18-088.html>,
      <https://w3id.org/ldes/specification> ;
   prof:hasResource [
      prof:hasRole role:vocabulary ;
      prof:hasArtifact <https://w3id.org/stapi>
   ] , [
      prof:hasRole role:validation ;
      prof:hasArtifact <https://voc-hub.org/shapes/sensorthings-api-shape.ttl>
   ] .
<https://meta-boker.org/distributions/hamburg-stapi>
   dct:conformsTo <https://voc-hub.org/profiles/stapi-ldes>.

<https://voc-hub.org/profiles/oslo-verkeersmetingen> a prof:Profile ;
   dct:title "OSLO Verkeersmetingen Profile"@en ;
   dct:publisher <https://data.vlaanderen.be/id/organisatie/OVO000009> ;
   prof:isProfileOf <https://data.vlaanderen.be/doc/applicatieprofiel/verkeersmetingen/
       erkendestandaard/2024-04-17/> ;
   prof:hasResource [
      prof:hasRole role:vocabulary ;
      prof:hasArtifact <https://data.vlaanderen.be/ns/verkeersmetingen>
   ] , [
      prof:hasRole role:validation ;
      prof:hasArtifact <https://voc-hub.org/shapes/oslo-verkeersmetingen-shape.ttl>
   ] .
```

Listing 2: Profile catalog with SensorThings LDES and OSLO Verkeersmetingen profiles

3.3. Alignment artifacts

With the goal of simplifying the integration process of published datasets for data consumers, the data space should support the alignment of relevant published dataset into the data models required by their

internal systems. Such alignments require not only the structural conversion of the source dataset into the target data model, but also align the semantic meaning of their terminologies. In practice, this can be achieved by means of e.g., SPARQL Construct queries, RML mappings, N3 rules or dedicated vocabularies. Additionally it should possible to validate the compatibility of the alignment outcome in the local system (consumer) compared to the source system (publisher).

For this reason, the alignments are published as separate semantic artifacts, that include the mapping resource that performs the necessary structural and ontological mappings between datasets from the source to target profile, and additionally are able to include supporting information, such as describing the quality of the performed alignment and potential supporting resources through the same profile mechanism used for the dataset profile artifacts.

The separation of the mapping artifacts from the profile definitions follows their decentralized nature in the data space ecosystem. Since mappings between profiles are envisioned as publishable by ecosystem participants to the vocabulary hub, the same requirements for governance, iterative design, and requirements for trustworthiness based on the publisher credentials and associated metadata, provides a strong requirement in viewing these as their own artifacts rather than as part of the profile definitions.

To model the mapping profile, we introduce the *pmap:sourceProfile* and *pmap:targetProfile* properties, with an associated *pmap:ProfileAlignment* class to denote the direction of the profile mapping. This was already considered in the past but was deemed out of scope for the Profile Vocabulary[19].

```
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix prof: <http://www.w3.org/ns/dx/prof/> .
@prefix role: <http://www.w3.org/ns/dx/prof/role/> .
@prefix pmap: <https://w3id.org/pmap#> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
<https://voc-hub.org/alignments/catalog> a dcat:Catalog ;
   dcat:resource <https://voc-hub.org/alignments/stapi-to-oslo> .

<https://voc-hub.org/alignments/stapi-to-oslo>
   a pmap:ProfileAlignment , prof:Profile , dcat:Resource ;
   dct:title "SensorThings to OSLO Traffic Alignment"@en ;
   dct:publisher <https://urbandataplatform.hamburg> ;
   prof:isProfileOf <https://voc-hub.org/profiles/stapi-ldes> ;
   pmap:sourceProfile <https://voc-hub.org/profiles/stapi-ldes> ;
   pmap:targetProfile <https://voc-hub.org/profiles/oslo-verkeersmetingen> ;
   dqv:hasQualityMeasurement [
      dqv:isMeasurementOf dqv:completeness ;
      dqv:value "0.87"^^xsd:double
   ] ;
   prof:hasResource [
      a prof:ResourceDescriptor ;
      prof:hasRole role:mapping ;
      dct:title "SPARQL CONSTRUCT transformation"@en ;
      dct:format <https://www.iana.org/assignments/media-types/application/sparql-query#
          Resource> ;
      dct:conformsTo <http://www.w3.org/TR/sparql11-query/> ;
      prof:hasArtifact <https://voc-hub.org/alignments/stapi-to-oslo.rq>
   ] .
```

Listing 3: Alignment catalog entry transforming SensorThings to OSLO Verkeersmetingen

[19]Introduction of from and to properties to define mapping resources - https://github.com/w3c/dx-prof/issues/15

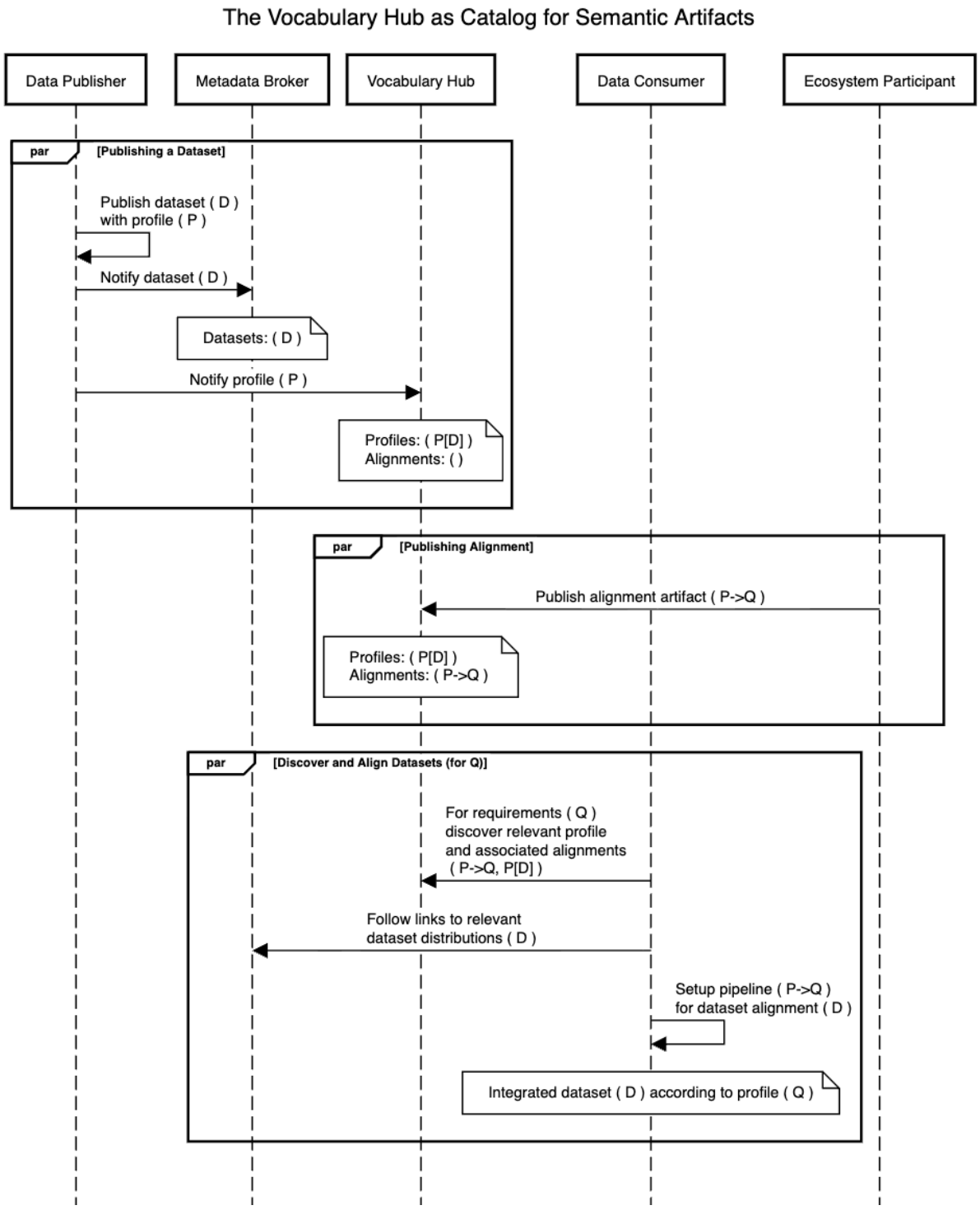


Figure 1: The interactions afforded by the addition of dataset profiles and alignment queries

4. Architectural implications and interactions

We closely follow the architectural design of both the IDS-RAM and the DSSC blueprint[20]. In this section, we detail the envisioned interaction patterns for data publishers and data consumers in their participation with the data space, going from dataset publishing to the data integration step, skipping over the orthogonal aspects of the data space for the proposed approach, such as authentication and

authorization. The resulting interactions are showcased as a sequence diagram in Figure 1.

4.1. Dataset publishing

Aiming at wide adoption for published datasets, data spaces benefit from the availability of rich semantic information to maximize the alignment potential for data consumers to integrate these datasets into their local systems. The use of semantic data formats for data publishing provides a rich basis both for alignment and integration requirements.

Where local tools such as the PoolParty suite enable mapping of structured data models into semantic formats through internal workflows, this alignment capability can be extended to the data space through the publishable artifacts defined in section 3. Publishers can define a source dataset profile linking the core specifications to which it adheres, with an extending profile defining the extensions, constraints and examples of their native API. Alignments can now be published to the data space as artifacts that include mappings from such profile into a target profile detailing the resulting semantic format constraints through SHACL validation resources, used ontologies and conformance specifications.

In the same way data space connector catalogs and metadata brokers maintain metadata of published datasets, the vocabulary hub catalogs both the dataset profile artifacts and alignment artifacts, as discussed in subsection 3.2. This positions the vocabulary hub as a entry point for discovery and search over dataset based on the structural and semantic definitions of their included resources, such as vocabulary, schema, validation and specification resources.

4.2. Publishing alignment artifacts

Alignments between dataset profiles function as collaborative semantic artifacts within data spaces, analogous to shared ontologies, supporting data integration across heterogeneous sources. The extended dataset profile introduced in subsection 3.2 provides the flexibility to define dataset-specific application profiles from which alignment artifacts are generated. These alignment artifacts may originate from multiple sources. Dataset publishers can decide to provide alignments to semantic formats, or alignments to external data standards, while external integrators may publish alignments to more specialized or localized data models.

This distributed alignment ecosystem requires consumers to make informed decisions when selecting and composing alignments for their integration workflows. To support such decision-making, alignment artifacts incorporate provenance information, publisher attribution, and Data Quality Vocabulary (DQV) descriptions of the alignment process, leveraging the metadata capabilities of DCAT and the Profiles Vocabulary as detailed in subsection 3.3. These metadata elements establish a confidence basis for consumers, similar to published dataset and profile information, based on which integration decisions can grounded. How exactly this is integrated in the decision making process is beyond the scope of this contribution, but we note that this can build upon the existing frameworks employed to ground confidence for dataset and ontology integration choices.

4.3. Dataset discovery and alignment

With the addition of the extended dataset profile and alignment artifacts published by the vocabulary hub in the data space, data consumers are presented with additional flexibility in how they pursue the integration of published datasets.

In addition to the filtering of datasets based on specific DCAT application profiles, the dataset profiles introduced in subsection 3.2 and published as a DCAT catalog by the vocabulary hub, enable the filtering of datasets based on the availability of resource descriptors for specific roles and the contents of these descriptions. Filtering can happen based on availability of schema resources, used ontologies, specification documents, validation resources, as well as on filtering logic for specific implementations, such as JSON-schema matches, SHACL shape matching or matching on used ontological terms.

[20]Data Spaces Support Center blueprint for Data Spaces - https://blueprint.dssc.eu/

Upon making a decision for a target set of profiles T, that can be integrated into the consumer local systems, the consumer can further look for alignment artifacts published by the vocabulary hub. Based on the alignment artifacts, and an evaluation step deciding the consumer confidence in the alignment quality and trustworthiness, the consumer can extend their integration capabilities to all source profiles linked by the trusted alignments for which the target profile is in T. The links between the profile definitions and the datasets that correspond to these profiles, can be linked back from the profile definitions in the vocabulary hub, and alternatively be discovered form the catalog or metadata broker catalogs in the data space. Based on this discovery step and assuming all source datasets are available to the consumer, the consumer can now setup the necessary integration pipelines, based on the source data formats and alignment mappings retrieved from the vocabulary hub, to integrate published datasets into their local data systems.

Where the approach described above details an interaction flow based on the processing of published DCAT catalogs, the availability of tooling for the management of these additional published artifact catalogs could streamline their integration in the same way they achieve for published ontology information in data spaces currently.

5. Use case

This work was performed in the context of the DeployEMDS project, for continuing the deployment of mobility data spaces in the European Union, with a focus on how the vocabulary hub can serve to improve interoperability and the data integration capabilities of actors in the data space.

5.1. Data sources

With the goal of exchanging and integrating mobility data across countries, we included data from Germany and Belgium, both from official sources and industry partners in the setup of our demonstrator. We retrieve data from the Hamburg IOT data platform [21], which conforms to the OGC SensorThings API specification [15]. Here, we defined the SensorThings API ontology (stapi)[22], and setup an integration pipeline using RDF-Connect[23] [16]. It automatically keeps up to date with a data stream on the Hamburg IOT data platform, and executes an RML mapping that converts the data from its original JSON format into a Linked Data Event Stream (LDES) [17] using the created ontology, which is then published as a set of interlinked resources according to the LDES specification [24]. For Belgium, we included traffic counting data both from both the Straatvinken[25] and Telraam[26] initiatives, that respectively include manual traffic counting information, and citizen data sensors that automatically count passing traffic in streets across Flanders. These traffic counts are also published as LDES for both Straatvinken and the Telraam datasets, both using the Verkeersmetingen ontology[27] published by OSLO. Finally, we include traffic count prediction datasets synthesized by an industrial partner, that is be made available using the SSN/SOSA ontology [18].

5.2. Alignment mapping pipelines

We provided two alignment artifacts in the form of SPARQL Construct queries, one for adapting the profile based on the created stapi ontology into a profile representing the SOSA/SSN data model, and a second query from the Flemish OSLO Verkeersmetingen profile into the SOSA/SSN profile. These

[21] https://iot.hamburg.de/

[22] The created SensorThings API ontology (stapi) is published at https://w3id.org/stapi

[23] The RDF-Connect pipeline setup to convert data from the SensorThings data source into the created stapi ontology is published at https://github.com/rdf-connect/hamburg-to-ldes-pipeline

[24] https://w3id.org/ldes/specification

[25] The Straatvinken initiative performs manual traffic counting in Flanders - https://www.straatvinken.be/

[26] The Telraam initiative collects sensor data by citizen traffic sensors - https://telraam.net/

[27] The Flemish data model for describing traffic counting data - https://data.vlaanderen.be/ns/verkeersmetingen/

queries can be found on Github[28]. The execution of these queries on the client is done using an RDF-Connect pipeline setup via a docker container. Based on the data source, a fetch component is setup to retrieve materialized resources, or an LDES client is setup to retrieve a source Linked Data Event Stream. The dataset entities are then piped through a SPARQL engine that evaluates the retrieved queries if alignment is needed. Finally, the mapped entities are added to a local graph store, from which local applications can access the retrieved data.

6. Implementation

The implementation of a prototype setup for our proposed architecture can be found at https://github.com/IDLabResearch/vocabuary-hub/ [29] It provides a demonstrator setup, that initializes up the necessary components with vocabulary hub and metadata broker functionalities locally using a docker setup, and provides a user Web Interface for showcasing the interactions that are made possible with the addition of dataset profiles and alignments as semantic artifacts in the data space. The demonstrator assumes that all datasets are available to the consumer, and all catalogs are writeable by the actors. Governance challenges are left out of scope for the demonstrator.

6.1. Setting up the data space components

The demonstrator includes a setup for all data space components that are needed to showcase the functionality in the Web interface. Primarily, this includes setting up the materialized DCAT catalogs, as they would be exposed by the metadata broker and vocabulary hub components. These include the catalogs defining the datasets, profiles and alignments. Additionally, an RML mapping service is started which, given a data source and YARRRML mapping configuration (editable in the Web interface), performs the RML mapping over the target data source, and publishes the resulting dataset distribution in RDF format at a given location, while updating the DCAT catalog with the new distribution for the given dataset.

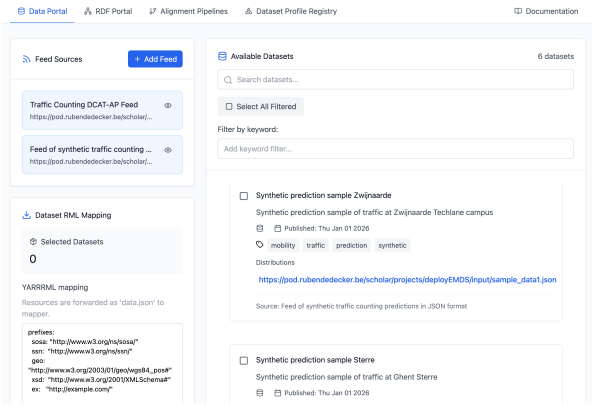
6.2. The Web user interface

The Web user interface showcases the functionalities for the publishing of datasets, and the consuming of published datasets. Such tasks are the responsibility of data space connectors, but we simplify and enable them for demonstrating how they can be assisted by the proposed semantic artifacts. The interface consists of four sections, that interact with the DCAT catalogs setup in subsection 6.1, as they would be published by the data space metadata broker and vocabulary hub components. The four sections of the interface are showcased in Figure 2.

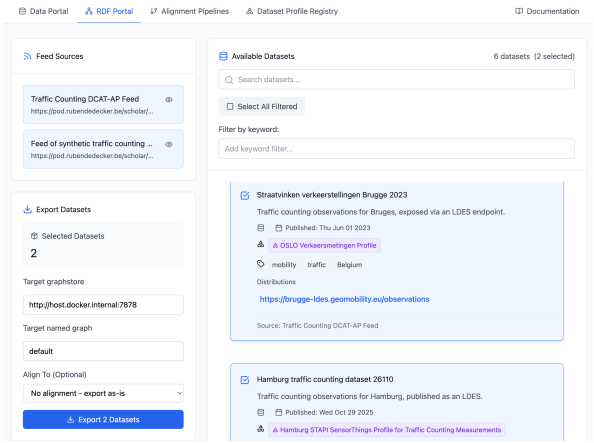
The Data Portal section includes functionality that is aimed at the interaction with the data space catalogs and metadata brokers, consisting of three components. The *Feed sources* component provides an overview of the currently loaded DCAT catalogs form which dataset entries are included. It provides functionality for loading DCAT catalogs, and experimentally for DCAT-AP-feeds [30]. The *Dataset browser* component enables the filtering of the loaded datasets and their distributions, through filter-based selection. When selecting non-RDF datasets and providing a YARRRML mapping [31], the *Dataset RML Mapping* component calls the external mapping service to perform an automated mapping of the selected datasets. The resulting RDF dataset is stored at the provided location, and the resulting metadata entry is added as a distribution of the dataset entry in the DCAT catalog.

The RDF Portal section is structurally similar to the Data Portal page, but filters the loaded datasets based on their distributions in an RDF format, as this is the basis on which we perform semantic

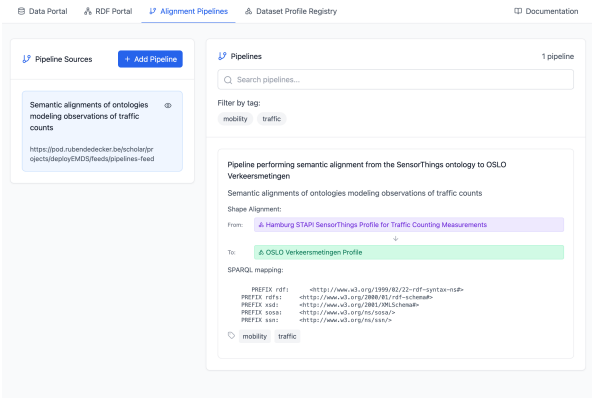
[28]The conversion queries are published at https://github.com/deployEMDS/vocabulary-hub-demo/tree/main/examples/queries
[29]A live instance is available at https://idlabresearch.github.io/vocabuary-hub/
[30]https://semiceu.github.io/LDES-DCAT-AP-feeds/index.html
[31]https://rml.io/yarrrml/spec/



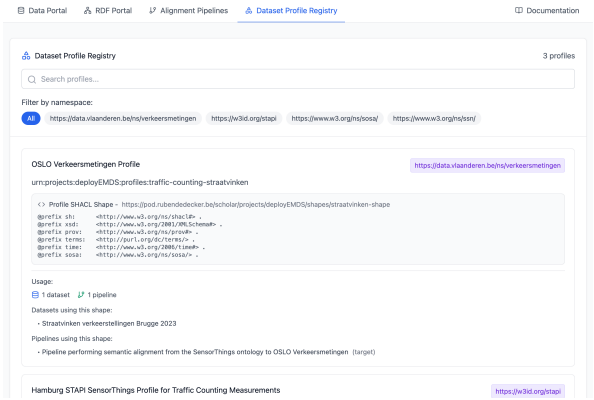
(a) The Data Portal page loads dataset catalogs, filters dataset and performs RML mappings.



(b) The RDF Portal exports RDF dataset distributions and alignments for a target profile



(c) The Alignment Pipelines page manages the alignment artifacts loaded in the interface



(d) The Dataset Profile Registry page shows the loaded dataset profiles

Figure 2: The Web interface showcasing the data space functionality afforded through the published dataset, profile and alignment catalogs

alignments. Here we bring together the dataset entries loaded from the dataset catalogs, with the profile entries from the profile catalogs, and map these profiles on the alignments loaded from the alignment catalogs, as shown in the *Retrieving Dataset for profile* section of Figure 1. Finally, the *Export Datasets* component exports the current selection of RDF distributions of datasets, according to a target profile that can be selected in the interface. The selected distributions are checked on their profile information, and the interface returns the set of selected data sources that can be directly integrated in the target profile, and the set of data sources that can be aligned to the target profile by providing the accompanying profile alignment queries. This is currently implemented in the form of a docker compose configuration, that automatically sets up the correct loader, SPARQL Construct-based alignment and publishing components to load data from either an RDF Resource or LDES Source, align the dataset entities using a SPARQL engine with the relevant alignment query, and outputs the resulting mapped entities to a local graph store.

The Alignment Pipelines section is used to manage the alignment pipelines. The *Alignment sources* panel shows the loaded alignment catalogs, and includes functionality for the adding of additional alignment queries to the loaded catalogs. The *Alignment browser* panel displays the loaded alignment artifacts, showing the source and target profiles, and the SPARQL Construct query that maps dataset entities between the profiles, linking to the relevant entries in the Dataset Profile Registry page.

The Dataset Profile Registry section shown in Figure 2 provides an overview of the dataset profiles

from the loaded profile catalogs. The *Profile browser* lists the available profiles, showcasing the resources of which it consists, and includes their links to the related dataset distributions and alignment artifacts loaded in the interface, showcasing the interlinking between the catalog entries.

6.3. Implementation considerations

The current architecture setup supports streaming data sources through approaches such as LDES, enabling live data integration via streaming RDF and SPARQL construct mappings. We make use of this approach both for the mapping of streaming data sources of traffic counting data into semantic data formats at the publisher side, and similarly for performing semantic alignments through SPARQL construct queries at the data consumer side.

Using RDF-Connect as a pipelining framework, these streams can be parallelized during import into local graph stores, facilitating scalable data ingestion workflows. As the LDES protocol is based on HTTP, binary streaming formats such as Jelly [19] can be considered where higher throughput is required.

Discovery and search work on the basis of fully retrieving available catalogs, yet for more extensive data spaces, the feasibility of loading complete catalogs on a Web interface at the client side may be infeasible. Where the catalogs serve as the base interoperability layer, services and tooling available in the data space can help relieve scaling issues of this nature, through the provision of uniform search interfaces e.g., via SPARQL endpoints.

7. Conclusion

With this work, we tackle some of the challenges stated in the W3C Dataspace Challenges Community Group report[32], focused on data discovery and enabling schema alignment and semantic transformations in data spaces.

Where current data spaces initiatives align on specific DCAT application profiles as a structural agreement for exchanged datasets, the inclusion of *dataset profiles* as semantic artifacts provides an extension to the application profile datasets can advertise based on a uniform data model. The linking of conformity to ontologies, schemas, validation resources and specifications imposes specific structural and semantic constraints, providing a strong basis for discovery and search mechanisms in data spaces that surpasses the rigid metadata definitions of DCAT application profiles.

The inclusion of alignments as artifacts published to the data space, envisions a collaborative basis, similar to ontologies and other semantic artifacts, that can build on the same systems of decentralized publishing, systems of governance and trust, that provides tangible benefits for enabling alignments in data spaces. By including mappings through languages such as SPARQL construct and RML, the mappings provided by these alignment artifacts are directly executable on retrieved datasets, enabling both structural and semantic alignment requirements. The advertised quality, authorship and other information associated to the mappings, provides a confidence basis on which consumers can decide to include mappings in their integration pipelines, or choose between different available options.

Our demonstrator for the DeployEMDS project use case of aligning traffic counting data over country and industry borders, shows that publishing these extended semantic artifacts facilitates the integration of heterogeneous data streams. Although the demonstrator was not yet implemented for scalability, the nature of the approach through artifact catalogs and interlinking of artifacts and dataset descriptions, opens the door for tooling integration that can resolve some of the performance and scaling considerations of the proposed approach. Additionally, publishing semantic artifacts as DCAT catalogs in the data space through the vocabulary hub, provides a basis for more extensive capabilities in federated implementations.

Finally, with data spaces predating the arrival of Large Language Models, the restrictiveness of the DCAT application profiles, although beneficial for uniform integration in existing architectures, may

[32]Dataspace Challenges report - https://w3c-cg.github.io/dataspaces/

miss opportunities for more flexible integration pathways enabled by LLM agents. The dataset profile, in its extension of the application profiles, provides a straightforward approach for annotating datasets with specifications, structural and semantic constraints, and usage examples that can be picked up in discovery, search and integration tasks of such agents.

8. Future work

Future work will focus on integrating the extended semantic artifacts published by the Vocabulary Hub in the form of dataset profiles and alignment queries in the data space component connector implementations. Here, will focus on further automating semantic interoperability through profile-based content negotiation mechanisms that allow data consumers and their corresponding data space connectors to automatically identify compatible dataset representations and alignments in the data space. Additionally, we aim to validate the prototyped approach in larger and more diverse deployment scenarios, to more clearly assess scalability challenges, automation opportunities, and interoperability across a broader range of domains.

Acknowledgments

This work was performed in the context of the DeployEMDS project, co-funded under the Digital Europe Programme[33], contributing to the further development of the common European mobility data space.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-5.2 in order to: Grammar and spelling check and managing sentence structure. The author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] O. M. van der Valk, M. Ryan, Data for the common good in the common european data space, Data 38; Policy 7 (2025) e32. doi:10.1017/dap.2025.5.

[2] A. N. Lam, R. Avogadro, F. Martin-Recuerda, B. Elvesæter, X. Ma, E. J. Nystad, D. Roman, A. J. Berre, Towards a toolkit for semantic interoperability in data spaces, in: 2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS), 2025, pp. 1–7. doi:10.1109/ICPS65515.2025.11087848.

[3] S. Bader, J. Pullmann, C. Mader, S. Tramp, C. Quix, A. W. Müller, H. Akyürek, M. Böckmann, B. T. Imbusch, J. Lipp, S. Geisler, C. Lange, The international data spaces information model – an ontology for sovereign exchange of digital content, in: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web – ISWC 2020, Springer International Publishing, Cham, 2020, pp. 176–192.

[4] J. Theissen-Lipp, M. Kocher, C. Lange, S. Decker, A. Paulus, A. Pomp, E. Curry, Semantics in dataspaces: Origin and future directions, in: Companion Proceedings of the ACM Web Conference 2023, WWW '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 1504–1507. URL: https://doi.org/10.1145/3543873.3587689. doi:10.1145/3543873.3587689.

[5] L. Sánchez-González, A. Iglesias-Molina, Ó. Corcho, M. Poveda-Villalón, On the governance of semantic artefacts in dataspaces., in: SDS@ ESWC, 2024.

[6] D. V. Lancker, S. Logghe, J. A. Rojas, A. D. Craene, Z. Vanlishout, P. Colpaert, Semantic andnbsp;technically interoperable data exchange innbsp;thenbsp;flanders smart data space, in: The Semantic

[33] Digital Europe Programme - https://digital-strategy.ec.europa.eu/en/activities/digital-programme

Web – ISWC 2024: 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11–15, 2024, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2024, p. 289–303. URL: https://doi.org/10.1007/978-3-031-77847-6_16. doi:10.1007/978-3-031-77847-6_16.

[7] I. D. S. Association, Industrial Data Space – Reference Architecture Model (IDS-RAM), Version 4.0, Technical Report, International Data Spaces Association, Dortmund, Germany, 2022. URL: https://internationaldataspaces.org.

[8] R. David, P. Ivanov, V. Alexiev, Raising the role of vocabulary hubs for semantic data interoperability in dataspaces, in: Workshop on Semantic Interoperability in Data Spaces, Budapest, Hungary, 2024.

[9] R. David, V. Alexiev, P. Ivanov, W. van den Berg, J. P. Wijbenga, M. Stornebrink, Federated vocabulary hubs as a building block for semantic layers in data spaces (2025).

[10] 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, Technical Report, TNO, 2024. URL: https://coe-dsc. nl/knowledge-base ..., ????

[11] A. Miles, S. Bechhofer, Skos simple knowledge organization system reference, 2009. URL: https://www.w3.org/TR/skos-reference/.

[12] A. Iglesias-Molina, D. Van Assche, J. Arenas-Guerrero, B. De Meester, C. Debruyne, S. Jozashoori, P. Maria, F. Michel, D. Chaves-Fraga, A. Dimou, The rml ontology: A community-driven modular redesign after a decade of experience in mapping heterogeneous data to rdf, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 152–175.

[13] R. Albertoni, D. Browning, S. Cox, A. N. Gonzalez-Beltran, A. Perego, P. Winstanley, The w3c data catalog vocabulary, version 2: Rationale, design principles, and uptake, Data Intelligence 6 (2024) 457–487. URL: https://doi.org/10.1162/dint_a_00241. doi:10.1162/dint_a_00241. arXiv:https://direct.mit.edu/dint/article-pdf/6/2/457/2458974/dint$_{a0}$0241.*pdf*.

[14] R. Atkinson, N. Car, The Profiles Vocabulary, W3C Working Group Note, World Wide Web Consortium (W3C), 2019. URL: https://www.w3.org/TR/dx-prof/.

[15] S. Liang, T. Khalafbeigi, H. van der Schaaf, OGC SensorThings API Part 1: Sensing, Version 1.1, OGC Implementation Standard 18-088, Open Geospatial Consortium, 2021. URL: https://docs.ogc.org/is/18-088/18-088.html.

[16] Vercruysse, Arthur and Pots, Jens and Rojas Melendez, Julian and Colpaert, Pieter, RDF-Connect : a declarative framework for streaming and cross-environment data processing pipelines, in: Joint Proceedings of the 1st Software Lifecycle Management for Knowledge Graphs Workshop co-located with 23th International Semantic Web Conference (ISWC 2024), volume 3830, 2024, p. 15. URL: https://ceur-ws.org/Vol-3830/.

[17] D. Van Lancker, P. Colpaert, H. Delva, B. Van de Vyvere, J. R. Meléndez, R. Dedecker, P. Michiels, R. Buyle, A. De Craene, R. Verborgh, Publishing base registries as linked data event streams, in: M. Brambilla, R. Chbeir, F. Frasincar, I. Manolescu (Eds.), Web Engineering, Springer International Publishing, Cham, 2021, pp. 28–36.

[18] A. Haller, K. Janowicz, S. Cox, D. L. Phuoc, K. Taylor, M. Lefrançois, Semantic Sensor Network Ontology, W3C Recommendation vocab-ssn, World Wide Web Consortium, 2017. URL: https://www.w3.org/TR/vocab-ssn/, published 19 October 2017.

[19] P. Sowiński, K. Bogacka, A. Danilenka, N. Kozlov, Jelly: a fast and convenient rdf serialization format, arXiv preprint arXiv:2506.11298 (2025). URL: https://arxiv.org/abs/2506.11298.