ICT, Strategy & Policy
www.tno.nl

TNO 2024 - 14 June 2024

Establishing semantic interoperability across data spaces: a solution for sharing vocabularies

Author(s)	Jelte Bootsma, Jan Pieter Wijbenga, Linda Oosterheert, Michiel Stornebrink, Wouter van den Berg
Editor	Wouter van den Berg
Roles & affiliations	• Consultants semantic interoperability at TNO unit ICT, Strategy & Policty • Members of the Semantic Treehouse development team

Short description	This paper is focused on developing a standardized approach for exchanging vocabularies across data spaces to foster interoperability, alignment, and reuse of vocabularies. It addresses the challenge of semantic interoperability when data needs to be shared across varied data spaces, which may develop their own vocabularies and tools. It proposes utilizing the Data Catalogue Vocabulary Application Profile (DCAT-AP) as a model for describing and exchanging vocabularies and builds on the concept of the vocabulary hub as a component of a data space to utilize this standardized approach. The goal is to enable better discoverability of vocabularies and facilitate negotiations for common vocabularies or mappings between different ones

Release notes	This document is an abbreviated version of a previous publication with the same title published by the Dutch 'Centre of Excellence for Data Sharing & Cloud', which mission is to develop a generic data sharing infrastructure that enables seamless data sharing between existing data spaces. For more information, please visit https://coe-dsc.nl/

1 Introduction

We first clarify the foundational concepts for the vision presented in this paper.

1.1 Vocabulary

In this paper we use the term 'common vocabulary' for any kind of specification that establishes a standard language for consistently describing, interpreting and annotating data by offering a structured sets of terms, concepts, and definitions. These vocabularies are like dictionaries that help a data provider and data consumer speak the same language when it comes to exchanging data. And they come in many forms. For example, consider a JSON schema specification that clarifies the structure and usage of a particular JSON document. If it includes terms and perhaps some explicit definitions or labels then it can be considered a light-weight vocabulary. If it is adopted by a group, it can be considered 'common'. When data is shared with an unambiguous and clear understanding this is called *semantic interoperability* [3].

In the International Data Space Association (IDSA) framework, the main responsibility for this common language lies with an intermediary role called a *vocabulary provider*. This party manages and offers vocabularies that can be used to annotate and describe data. To do effectively they often utilize tools that fall into a category called *vocabulary hubs*.

1.2 Vocabulary hub

To make vocabularies findable, accessible, and usable for the data space participants, a data space deploys a *vocabulary hub*. According to the IDS RAM 4 [4], the *vocabulary hub* is a service that stores, maintains, and publishes the vocabularies and enables collaborative management of the vocabularies. It is a service supporting vocabulary publication, editing, browsing, and maintenance. The vocabularies itself semantically describe the data that is exchanged between data space participants. Beyond accessibility, vocabulary hubs foster community collaboration on common vocabularies through features such as version management, issue tracking and a co-creation process. In essence, a vocabulary hub acts as a hub for efficiently managing and publishing the vocabularies used to specify the meaning of data within and across data spaces.

1.3 Federated data spaces

Data spaces enable collaboration by facilitating controlled exchange and sharing of data. Data spaces are currently being developed in numerous sectors and regions, and by individual consortia. To unleash the true potential of data sharing, there is a growing need to enable the exchange of data across different data spaces, i.e. a federation of data spaces. In a federation of data spaces, each individual data space instance has a high degree of autonomy in developing and deploying its own internal agreements and ICT landscape. However, jointly the individual data space instances pursue a common goal of being able to share data in a trusted manner. Therefore, interface agreements and specifications are the essential design artefact for a federation of data spaces to manage and coordinate the information flows between federated data spaces [5].

2 Federated vocabulary hubs: an architecture

We present the architecture of a vocabulary hub in a federated data space and illustrate how vocabularies can be annotated with metadata using DCAT-AP for effective exchange.

2.1 Components of a vocabulary hub

A vocabulary hub provides users with a clear overview of available vocabularies and how these vocabularies can be effectively utilized. The key components of a vocabulary hub include:

- Vocabulary creation and editing component; offering flexibility for starting from scratch or integration of existing vocabularies.
- Vocabulary repository; providing storage of all the distributions of vocabularies
- Catalogue component; enabling metadata descriptions and easy access to vocabularies within the vocabulary repository, facilitating their discoverability and reuse.

Typically, the development of vocabularies is organized by business communities and is delegated to standards development organizations (SDOs), which include small organizations serving a particular sector in a certain region, and not just ISO, ETSI, et cetera.

A vocabulary hub assists users in creating and publishing these vocabularies. It offers various functionalities for *vocabulary creation*, whether by reusing existing vocabularies or starting from scratch. All the distributions of vocabularies are stored within the *vocabulary repository* of the vocabulary hub.

To ensure the utilization and reusability of these vocabularies, a vocabulary hub provides functionalities for incorporating metadata about each vocabulary. This collection of metadata improves the discoverability of each vocabulary. All the metadata and a link to the vocabularies within the vocabulary repository are published in the *catalogue component* of the vocabulary hub.

Figure 1 illustrates the relationship between the types of data and the repositories of a vocabulary hub. The vocabulary hub is not involved in the actual data transaction between a data provider and a data consumer. Instead, the vocabulary hub only includes the vocabularies to which the data in the transaction must adhere. The catalogue component assists users in easily finding their desired vocabulary. Once the appropriate vocabulary is found by exploring the metadata, users can retrieve the specifications and details on the vocabulary itself from the vocabulary repository. Both the vocabulary repository and the catalogue are usually accessible via an API endpoint.

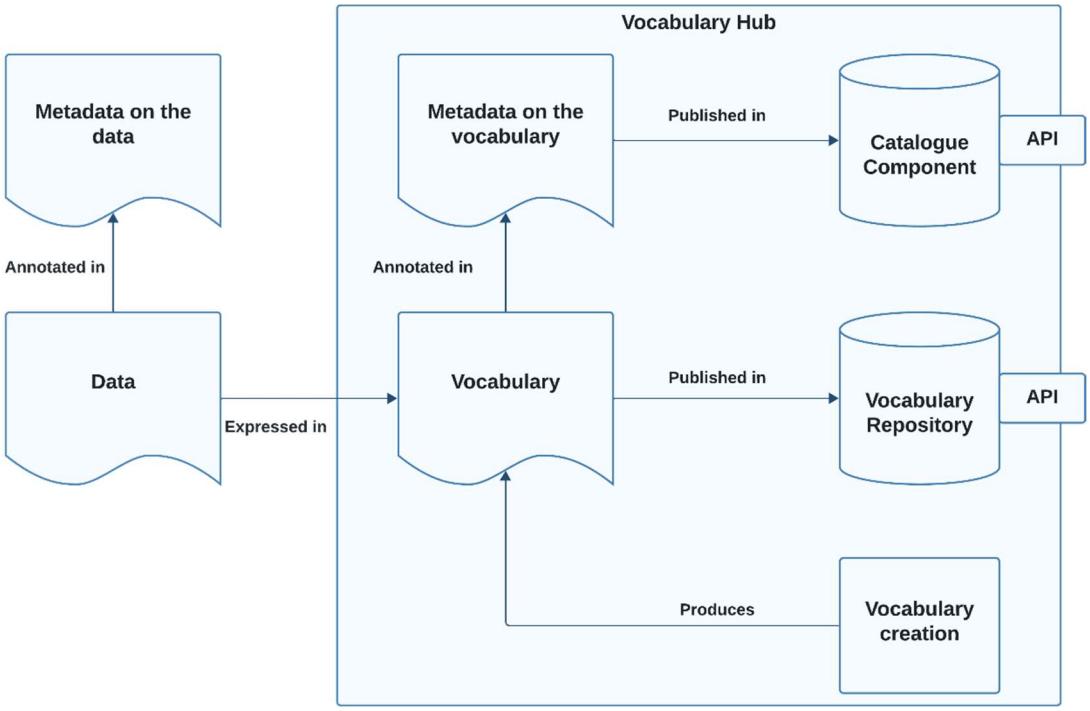

Figure 1: Components of a vocabulary hub

2.2 Federation through DCAT-AP

DCAT-AP can be used to facilitate the exchange of vocabularies between different vocabulary hubs by standardizing the metadata of vocabularies. By encouraging all vocabulary hubs to adopt DCAT-AP for metadata descriptions on vocabularies, we unlock the potential for federated searches for vocabularies across diverse vocabulary hubs.

The central notion in DCAT-AP, as can be read in the DCAT specification [6], is a Dataset, described as "a collection of data, published or curated by a single source, and available for access or download in one or more formats.". A Data Catalogue on the other hand is described as "a catalogue or repository that hosts the Datasets or Data Services being described". The definition of datasets and catalogues in DCAT-AP are broad and inclusive, aiming to embrace data types arising from diverse communities. Therefore, we view the catalogue component of a vocabulary hub as a DCAT Catalogue that hosts DCAT Datasets, representing vocabularies which can be considered as datasets. These vocabularies are available for access or download in one or more formats in our vocabulary repository.

DCAT-AP standardises information on dataset attributes, including descriptions, publishers, and version control. It allows a vocabulary to be any specification, from spreadsheets, ontologies, and JSON Schema to specialized formats. DCAT-AP does not assume anything about the vocabulary specification format but distinguishes between its various distributions. Therefore, DCAT-AP can be used to specify metadata on the vocabulary with reference to one or more distributions of the vocabulary in a vocabulary repository.

Figure 2 illustrates this idea with a basic example. This figure shows how to use DCAT-AP to describe the metadata of vocabularies.

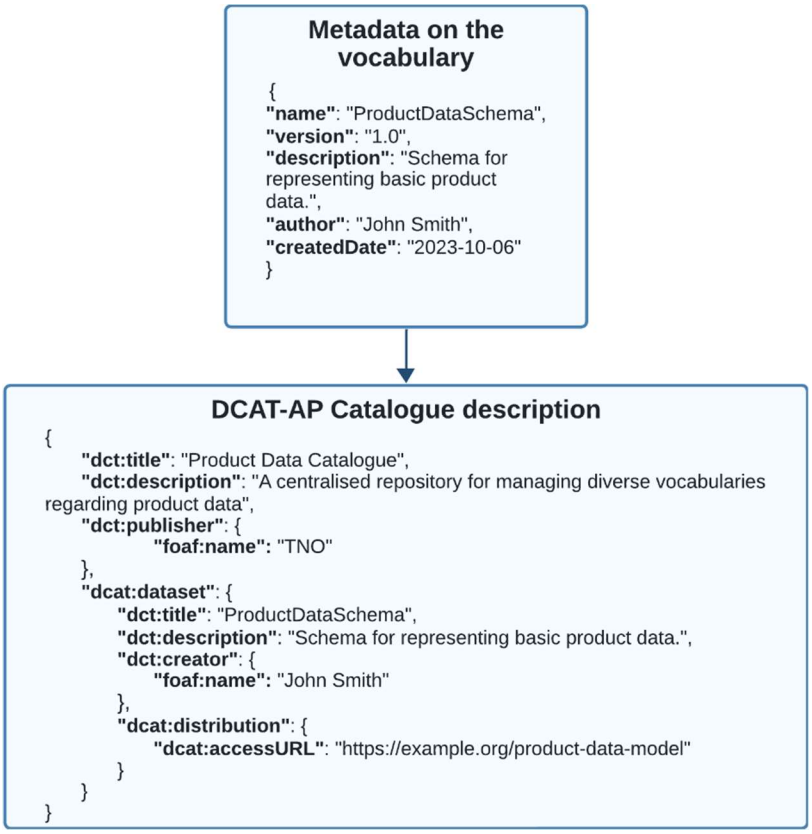

Figure 2: DCAT-AP Catalogue description including a vocabulary as DCAT Dataset

Once a vocabulary and a vocabulary hub are described using DCAT-AP, this representation can be used in the federation of data spaces. Figure 3 shows how the example from Figure 2 is exchanged between vocabulary hubs. The data contains "dcat:accessURL" that specifies how to access the distribution of the actual vocabulary. This can be an URL that enables retrieving a vocabulary via the API of the vocabulary repository in a vocabulary hub.

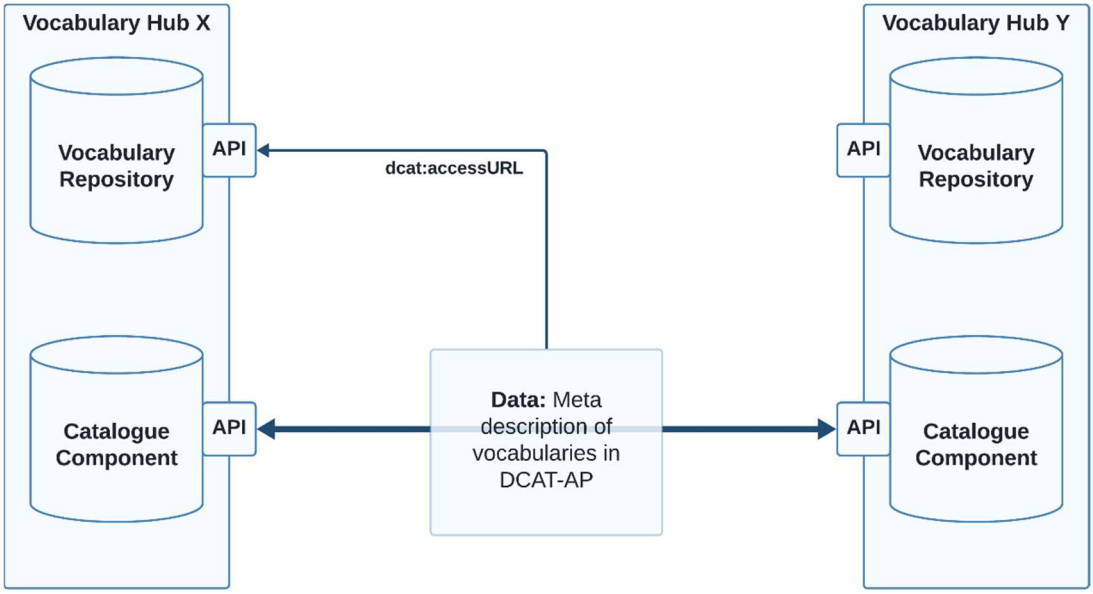

Figure 3: Exchanging vocabularies using DCAT-AP

3 Conclusion

Data spaces can be expected to employ their own vocabularies and vocabulary hub(s). A standardized solution is necessary to exchange vocabularies, bridging the gap in semantic interoperability across various data spaces. This paper has illustrated that vocabularies can be considered as datasets, enabling them to be described and shared using DCAT-AP.

If vocabulary hubs adopted this use of DCAT-AP as standardized approach it would allow data space participants to explore vocabularies from other hubs as if they were present in their own. This facilitates the creation of negotiating on semantics across data spaces, reducing the need for each data space to independently create and maintain its own. In essence, this marks the beginning of fostering semantic interoperability in federated data spaces.

3.1 Future work

Appendix A includes our demonstration of the ideas laid out in this paper. One of the current deployments of the Semantic Treehouse [7] vocabulary hub served as the testing environment. The test has proven the operational feasibility of describing and exporting vocabularies in DCAT-AP.

To facilitate the sharing of vocabularies between vocabulary hubs and other data space components (e.g., connectors and metadata brokers), it will be necessary to further develop this prototype by including importing functionality for DCAT-AP representations of external vocabularies.

As we conclude the current phase of our research and implementation, it is essential to look further ahead and outline potential directions for future work. The proposed approach currently focuses solely on DCAT-AP; however, considerations of other standards should be explored in the future. Some data space specifications already address contractual negotiations within data spaces, such as the data space protocol defining the contract negotiation protocol. In the future, attention should be paid towards determining the feasibility of agreeing upon semantics using this protocol or other specifications.

Additionally, addressing access control over non-public specifications is crucial, considering that some vocabularies are not freely downloadable.

Another aspect to address is the establishment of conditions and criteria for vocabularies intended for exchange. Vocabularies should meet specific criteria, including well-written documentation and permanent accessibility through a permanent URL.

4 References

[1] European Commission, “A European strategy for data,” 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

[2] European Commission , “DCAT Application Profile for data portals in Europe,” [Online]. Available: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/200.

[3] International Data Spaces , “Meaningful data,” [Online]. Available: https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4/context-of-the-international-data-spaces/2_1_data-driven-business_ecosystems/2_5_meaningful_data.

[4] International Data Spaces, “Vocabulary Hub - IDS Knowledge Base,” [Online]. Available: https://docs.internationaldataspaces.org/ids-knowledgebase/v/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.

[5] The Netherlands AI Coalition, “Towards a Federation of AI Data Spaces: NL AIC reference guide to federated and interoperable AI data spaces,” The Netherlands AI Coalition, 2021.

[6] W3C, “Data Catalog Vocabulary (DCAT) - version 3,” 07 March 2023. [Online]. Available: https://www.w3.org/TR/vocab-dcat-3/.

[7] TNO, “Semantic Treehouse,” TNO, [Online]. Available: https://www.semantic-treehouse.nl. [Accessed 14 June 2024].

[8] SETU, “SETU standaarden voor flexible arbeid,” [Online]. Available: https://setu.nl/.

[9] EC DG DIGIT Interoperability Test Bed, “BRegDCAT-AP validator,” EC DG DIGIT, [Online]. Available: https://www.itb.ec.europa.eu/shacl/bregdcat-ap/upload. [Accessed 14 June 2024].

Appendix A
Proof of concept

We have built a prototype to demonstrate how our vision for describing and exporting vocabularies in DCAT-AP can be operationalized. We have used TNO's vocabulary hub Semantic Treehouse [7] as starting point. As the testing environment we used the Semantic Treehouse deployment for SETU [8], a small-sized SDO serving the Dutch flexible staffing sector. This chapter describes details about the current implementation, shows example output, validation details and describes final work to be done.

A.1 Mapping to DCAT-AP

This section details the mapping from the current structure describing vocabularies in Semantic Treehouse to the DCAT-AP standard version 2.1.0.

As introduced earlier, a vocabulary is any specification that can be used to consistently describe a set of concepts or data. In Semantic Treehouse, all specifications fall into one of six types, such as message model specification (traditionally XSD messages) or ontologies. Multiple versions can exist for each specification, representing different releases or iterations of the same specification. These specification types and their versions can be distributed in seven types of export formats, including JSON Schema and OpenAPI specification. Overall, an implementation of Semantic Treehouse is called an implementation of a vocabulary hub and contains one or more projects that serve as organized catalogues for related specification and their versions.

We choose to map the Semantic Treehouse structure to DCAT-AP version 2.1.0. as depicted in the following table:

Table 1: Mapping from Semantic Treehouse structure to DCAT-AP v2.1.0.

Semantic Treehouse structure	DCAT-AP
Project	DCAT Catalogue
Specification	DCAT Dataset
Version of a specification	DCAT Dataset
Distribution of a specification	DCAT Distributions

A.2 Example of DCAT-AP export output

For the proof of concept, export functionality is built into Semantic Treehouse, which makes it possible to export each project to DCAT-AP, and to export the entire vocabulary hub with all projects to DCAT-AP. Two export options are currently provided: a button triggering the browser to download the content in DCAT-AP, and an API is available for machine-to-machine implementation, offering the same export functionality.

The export output is a ttl (Turtle) file that contains all specifications, their versions, and references to their distributions. Within each project catalogue, all the specifications are

bundled in a DCAT Catalogue. When exporting the content of the entire vocabulary hub, it is represented as a DCAT Catalogue, including all project catalogues expressed as DCAT Catalogues.

The screenshot below illustrates an exported 'project' containing a specification version named Human Resource Message version 1.3.1. This example represents a vocabulary used in the flexible staffing industry, where the Human Resource Message is used to match a human resource to an open position at an employer.

The first part of the export provides metadata about the specification, which facilitates discoverability by providing sufficient information to grasp the subject of a specification. In the second part, all distributions of the specification version are included. In this case, an XML schema for the Human Resource Message version 1.3.1 is included, with references to access or download the distribution of a specification.

```
<https://setu.staging.semantic-treehouse.nl/specversions/MessageModelVersion_35478627-b2c
  a dcat:Dataset ;
  dc:description "The SETU standard for Ordering and Selection is used for matching a hum
  dc:title "SETU HumanResource v1.3.1"@en ;
  dcat:distribution <https://setu.staging.semantic-treehouse.nl/specversions/MessageModel
  dc:publisher <https://setu.staging.semantic-treehouse.nl/groups/SETU> ;
  dc:identifier "https://setu.staging.semantic-treehouse.nl/specversions/MessageModelVers
  dcat:landingPage <https://setu.staging.semantic-treehouse.nl/#/Message_32_model/Message
  dc:issued "2015-06-16"^^xsd:date ;
  owl:versionInfo "1.3.1 (RELEASE)" .

<https://setu.staging.semantic-treehouse.nl/specversions/MessageModelVersion_35478627-b2c
  a dcat:Distribution ;
  dcat:accessURL <https://setu.staging.semantic-treehouse.nl/api/v1/fit/message/Property_
  dc:description "Distribution of type XSD for STH specification version with id MessageM
  dc:format <http://publications.europa.eu/resource/authority/file-type/SCHEMA_XML> ;
  dcat:downloadURL <https://setu.staging.semantic-treehouse.nl/api/v1/fit/message/Propert
  dcat:mediaType <https://www.iana.org/assignments/media-types/application/xml> ;
  dc:title "XSD schema distribution for MessageModelVersion_35478627-b2d0-4bce-baec-5776b
```

Figure 4: A cropped fragment of an export of the "Human Resource" message model as a DCAT Dataset and an XSD distribution.

A.3 Validation

As a means of testing our proof of concept, a generated DCAT-AP Catalogue containing all types of specifications were exported to a file and that file was uploaded to a European validation service for DCAT-AP [9]. The generated code passed the most important validity checks. The limited checks that failed had to do with the specific requirements of the validation service that focuses on a dialect of DCAT-AP with additional business rules, like the rule: "Catalogue Publishers need to be from a list of recognized authorities".