Data Product Metadata Management: an Industrial Perspective

Anonymous Authors

No Institute Given

Abstract. Decentralised data exchanges are promising alternatives to monolithic data lakes and warehouses which are typically emerging around complex service solutions. At the same time, novel frameworks such as data meshes and markets as well as community data spaces, shift the responsibility of providing data from central data offices towards domain-oriented data providers. In theory, this removes some of the bottlenecks of traditional data management solutions. In practice, the road towards achieving such goal is a long way ahead. In this work, we provide an industry perspective on the implications for such work, with a focus on metadata management; the work in question draws from an in-vivo action research approach we enacted at a major German automotive company that is transitioning to an internal decentral data market. Our results provide insight into an industry perspective on the requirements for metadata management. Additionally, we propose and validate a solution design for metadata management in decentralised data exchanges based on semantic web service technology.

Keywords: Data Mesh · Data Market · Data Product · Metadata · Semantic Web

1 Introduction

Despite the promises of big data to revolutionise the way companies do business, many organisations still grossly fail to fully capitalise on the data they are generating [1]. For example, several surveys and market analyses show that 60% to 85% of data analytic and data science initiatives never make it to production [18]. Critics have blamed this inability to make full use of data inside an organisation or company on the monolithic service-oriented approaches typically exploiting such data—e.g., data lakes and warehouses—that are nowadays the standard architecture approach for storing and exchanging data [15,11]. The main downside of these monolithic approaches is that they fail to scale with the number of data sources on the one hand and data science and analytics use cases on the other [15,25].

To address these shortcomings, grey and white literature is showing increasing interest in decentralised data exchanges, such as data markets [6], data

[1] https://www.sisense.com/blog/why-businesses-fail-to-capitalize-on-their-data/

meshes [9], and data spaces [19]. Despite some minor differences in how these platforms approach the sharing and exchange of data, their approaches all focus on offering data as a product/service and are heavily inspired by the microservice paradigm [17,16]. Whereas monolithic approaches rely on a central data office to facilitate data management, in decentralised data exchanges, data providers[2] are responsible for taking (operational) data from their domain and providing it in a manner that is fully optimised for consumption by data consumers from across the organisation.

Despite the theoretical promises of decentralised data exchange platforms, many challenges currently impede their effective implementation and migration. As far as we know, no company or organisation claims to have successfully organised its data exchange in accordance with any proposed theoretical framework. Currently academic studies, [11,15,9], and grey literature [25,23] focus on high-level architectural concerns, as well as categorising the challenges and solutions associated with migration and design [6,3]. In this paper we explore one such challenge, namely metadata management for achieving interoperability between data products (i.e. relating disparate data sources) in a data exchange. To supplement the existing work, we take the industrial perspective of a large German automotive manufacturer who is in the process of transitioning from monolithic architectures towards a decentralised data exchange.

The rest of this paper is organised as follows: In the next we section we discuss relevant works on metadata management for decentralised data exchanges. Section 3 introduces the industrial context provided by the company where our research took place. Section 4 describes how we leveraged design science research to arrive at the above contributions. Then, in section 5 we present the results of our research. Finally, in section 6 we discuss the threats to validity in our research approach, the implications of our work and suggestions for follow-up research.

2 Related Literature

We observe that metadata management for internal data exchange platforms is still very much in its infancy. Indeed, Eichler et. al. discuss state-of-the-art metadata management and conclude that there is a research gap in metadata management, especially for internal data markets [7]. Some works do exist that focus on the exchange of data *between* companies and organisations. For example, Roman et. al. present an ontology and show how it can be used to harmonise data from different organisations using well-established ontology development methods [20]. Similarly, Spiekermann et. al. present a metadata model for data products in the context of commercial data markets [21].

When it comes to *internal* data exchanges, proposed solutions for metadata management tend to focus on modelling (meta-)data in knowledge graphs using semantic web technology [26]. For example, Hooshmand et. al. emphasise

[2] Alternatively called data owners, (data) product providers, or (data) product developers [6].

the power of semantic web technology to capture and combine domain knowledge on business objects and technical information on data assets. They propose a transition towards a decentral data mesh for managing data in the product lifecycle management landscape and discuss how different domains can have separate knowledge graphs that can be mapped to achieve interoperability; however, they do not discuss explicitly what metadata management should look like [11]. Other relevant solutions for metadata management have been proposed in the context of centralised architectures. Stach et. al. note the advantages of semantic web technology in terms of ease of use for data consumers who are not experts at data modelling and propose a method for describing desired data products [22]. Similarly, Dibowski and Schmid introduce a full ontology for describing data assets on (internal) data lakes and explain how this improves the discoverability and reusability of data [4].

Despite the clear potential of semantic web technology for (internal) decentralised data exchanges, to the best of our knowledge, no investigation has explored how this metadata management approach would meet industry requirements. In fact, there is no clear overview of what industry requirements for metadata management are in the context of decentralised data exchanges. The DAUTNIVS[3] baseline usability attributes of data products proposed by Dehghani are perhaps the closest literature has come to such an overview [3]. However, these are high-level, focus on the entire architecture and do not impose any requirements on metadata management per se. This paper aims to address this gap in the literature by trying to ascertain whether and how semantic web technology can be used for metadata management for (internal) decentralised data exchanges. In doing so, we provide the following contributions:

1. We provide an industry perspective on the requirements for metadata management for data products in an internal data exchange that supplements existing theory.
2. We conceptually show how a metadata management approach based on semantic web technology can be applied in decentralised data exchanges.
3. We validate that this approach addresses all of the identified requirements in an industrial setting.

3 Industrial Context

For our investigation of interoperable data products, we engaged the IT division of a major German automotive manufacturer, which we refer to as the Data Market Implementation Team (DMIT). The company was experiencing challenges in effectively sharing data, and the DMIT was investigating new ways to tackle these challenges with an internal data market. The automotive manufacturer operates with a multi-billion euro revenue in a global market, is organised in several organisational units across multiple continents, employs more than

[3] Discoverable, Addressable, Understandable, Truthful, Natively Accessible, Interoperable, Valuable

100.000 employees, and has numerous partner companies in its business ecosystem. Importantly, the operations of this company are not limited to manufacturing but extend to different post-sales services as well. This collaboration allowed us to approach the problem in an industrial setting and get direct input from real-world data providers, data consumers and infrastructure providers.

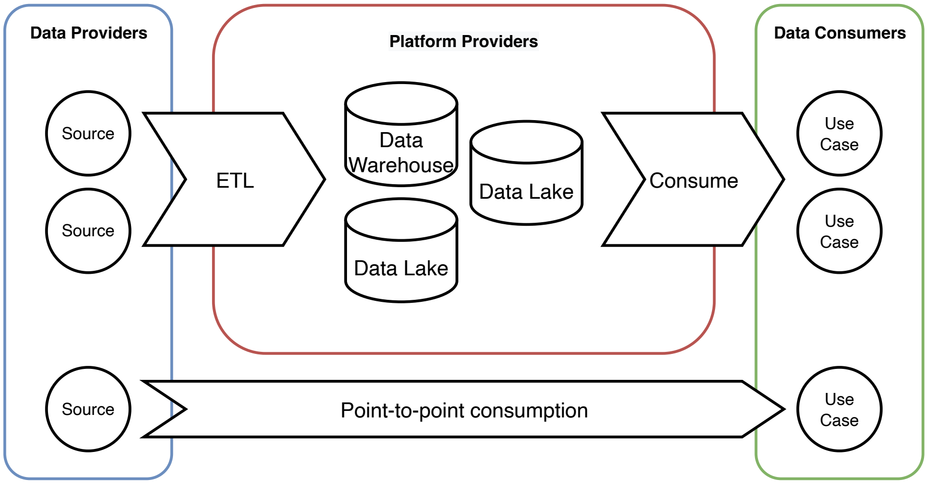

Fig. 1. A high-level overview of the original centralised data management architecture. Data providers control the operational systems that host the data sources. Platform providers build ETL pipelines to ingest the data into one or more data warehouses or data lakes that offer a single, centralised interface for consumption. Data consumers consume the data through this interface or set up direct, point-to-point connections for consumption. Metadata management is orchestrated centrally in the data warehouses and lakes controlled by the platform providers.

Figure 1 shows a high-level overview of the original centralised data management architecture. Like many large companies, the automotive manufacturer had formulated a strategy for transitioning towards a more data-driven business model, hoping to expand its analytical and data science capabilities by taking data from its operational environment and sharing it with data consumers across the company. As a result, data warehouses and data lakes were created on top of the domains' operational databases. These monolithic platforms ingest operational data through ETL (Extract, Transform and Load) operations and offer a central interface to data consumers. Interoperability between data from different sources is orchestrated inside data warehouses and data lakes by one or more centralised teams of platform providers.

However, despite efforts to standardise the infrastructure across the company, different requirements in the various domains and the existence of diverse legacy systems resulted in a heterogeneous landscape for both the data management architecture and the use cases that relied on it. Consequently, many data consumers were struggling to find and consume relevant data. In fact, the most effective use of data happened either inside domains (where the consumer is close to the provider) or through point-to-point connections (where the consumer can control the entire pipeline). Top-down efforts to improve the situation came from central management through the development of policies and standards that specify the requirements (e.g., legal compliance and responsibility management) and general guidelines for managing data (e.g., a data product life-cycle). At the same time, bottom-up initiatives came from the domains themselves, several of which had started developing their own platforms for data sharing.

These efforts notwithstanding, the problems mentioned in section 1 are also facing the existing data management architecture shown in fig. 1:

1. The platform providers struggle to keep up with the increasing amount of data that needs to be on-boarded; as for each new request from data consumers, a new ETL pipeline has to be created and maintained to integrate the data into the data warehouse or data lake. Similarly, data consumers struggle to create and maintain a large number of point-to-point connections with sources outside their domain and expertise.
2. The central platforms are not designed for highly heterogeneous use cases. For example, one use case focused on "Noise Vehicle Harshness" (i.e. the sound inside a vehicle under different conditions), but the platform was not designed with audio data in mind and making the required changes was expensive.
3. Finally, it became apparent that the separation of data providers and data consumers hindered effective data exchange. Data consumers had difficulties understanding what data was offered and how to make optimal use of it, while data providers were unaware of many use cases that required their data. Meanwhile, although highly proficient with the technical aspects of handling data, the platform providers lacked a nuanced understanding of both the data provider and the consumer's domains.

The first and last problem, in particular, are caused by the centralised metadata management used to achieve interoperability in the existing architecture: platform providers are unfamiliar with the ground truth of the source domain, which makes it challenging to create and keep up-to-date metadata. Moreover, because the data consumer is two steps away from knowledge about the data source, it is harder to understand and hence integrate the data.

In order to address these challenges on a company-wide level, the company is currently setting up an internal data exchange. Amongst other things, this involves creating the first *data products* that can work as pioneer projects and guide the transition towards a federated data exchange platform. Although the initial focus of this platform is internal (i.e. inside the company), we observe that this platform truly is a data market following the definition of Driessen et. al. [6] because it focuses not merely on data *sharing* but rather on data *exchange*. This can be seen from the fact that 1) the ultimate goal of the platform is to exchange data products also with *external* organisations, and 2) even though data products are not yet exchanged for money, the company recognises that some kind of reciprocity is desirable for data exchanging[4].

The DMIT had already implemented several core aspects of a data market, such as a corporate data catalogue and a policy engine for creating and enforcing data usage policies. However, it was still looking for a standard for creating metadata that promoted reuse and interoperability. Our goal, therefore, was to come up with *the best way to manage data product metadata such that it achieves interoperability and, by extension, reuse.*

4 Research Methodology

We employ the design science research approach, which focuses on creating and evaluating artefacts to simultaneously address industry-relevant relevant problems and contribute new knowledge to the scientific community [10]. As shown in fig. 2, our method consists of three steps in the design cycle: problem evaluation, treatment design, and treatment validation [27]. Below, we describe each of these steps individually, after which section 5 describes the results of each step and how these results informed the consecutive steps.

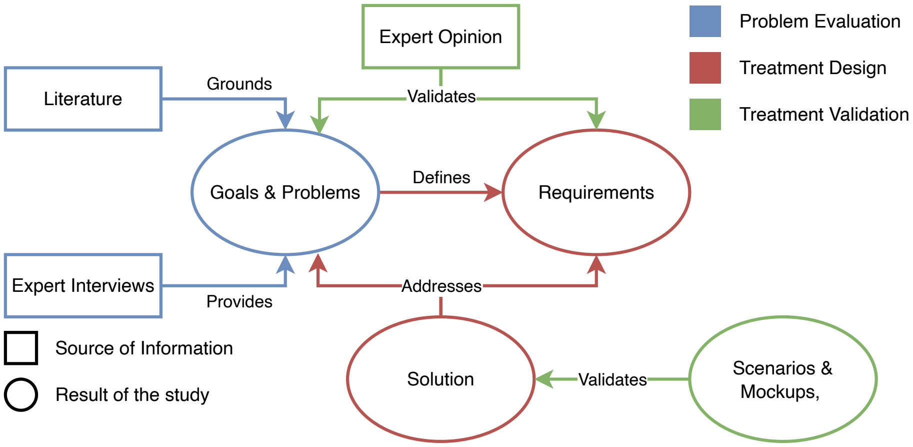

Fig. 2. A figure of showing how our methodology relates the results of the study. The rectangles show external sources of information and the ellipses show results presented with this paper. Additionally, the steps of the design cycle [27] are grouped by colour.

Problem Evaluation During the problem evaluation, we started by investigating existing literature reviews on decentral data exchanges in conjunction

[4] Several options to incentivise data providers have been considered, but these are not the focus of this work.

with repeated interviews with experts from the DMIT, to establish who would be the main stakeholders affected by the implementation of the internal data market. These efforts yielded three stakeholders. 1) The data provider, who is ultimately responsible for the data product. 2) The data consumer, who is the intended user of data products. 3) The platform provider, who is responsible for the IT infrastructure of the decentral data exchange, including the metadata management. Afterwards, we selected several experts from across the company with perspectives for each type of stakeholder, who were then interviewed to establish the different stakeholder goals and problems.

Treatment Design Based on the goals identified in the interviews and the context provided by the DMIT, we specified the requirements for our treatment and established how satisfying those requirements would contribute to the stakeholder goals. We then considered existing literature on metadata management for effective data sharing and found that an approach based on semantic web technology had the potential to address all the identified requirements.

Treatment Validation In order to validate our proposed solution, we relied on expert opinion whereby various stakeholders are asked to evaluate a potential solution by coming up with potential problems and benefits. [27,2,14]. To help validate that our solution addresses the requirements, we created and described several scenarios for creating, updating, combining and reusing data products and metadata. Furthermore, one or more mock-ups were created for each scenario to illustrate how our solution addressed the corresponding requirements. Afterwards, the scenarios and mock-ups were presented to experts from various domains. This included the interviewees from the problem evaluation step and members of the DMIT, who were asked for feedback. Finally, a workshop was organised where our findings were presented to a large audience of over 50 stakeholders at the company. At this point, feedback was solicited again from these stakeholders.

5 Results

This section discusses the result of our investigation in three steps. First, we introduce an overview of the goals, problems and requirements. Afterwards, a high-level metadata management solution for decentral data exchanges is presented. Finally, we discuss how this solution was validated in an industrial context.

5.1 Problem Evaluation

In order to accurately identify the requirements for interoperable data product metadata management, we performed interviews with ten expert stakeholders across the company. The experts were selected in consultation with the DMIT to represent the perspectives of the different stakeholders on data exchange. Additionally, as reflected by table 1, we consciously tried to interview people from various departments with varying levels of expertise and seniority. However, one challenge that arose was that there were no real data providers yet. This seems likely to occur in many organisations looking to transition towards a decentralised data exchange architecture. As mentioned in section 3, in the existing landscape of centralised architectures such as data warehouses and data lakes, onboarding data is the responsibility of a central team of platform providers. Part of moving towards a decentral data exchange entails moving these responsibilities to domains' expert data providers who work directly with the operational data [3]. In order to still get the data provider's perspective, we selected interviewees that had all been involved in previous initiatives to improve the existing data exchange infrastructure. Consequently, they were platform providers that had explicitly considered the challenge of onboarding new data assets, if not data products.

Table 1 shows the characteristics of the interviewed experts. On the one hand, the interviewees' experience with the existing data exchange infrastructure and its limitations made them ideal candidates for our investigation because they had already considered the goals and problems for their own initiatives. On the other hand, the emphasis on different perspectives (both domains and roles) ensured that we would not end up with a subset of relevant problems, goals and requirements.

The interviews themselves were semi-structured, focusing on existing processes and desired processes for data exchange and the planned internal data exchange. Each interview was recorded, after which we distilled goals and problems to supplement existing literature. The goals and problems that resulted from these interviews are described in tables 2 and 3 respectively and are dis-

Role	**Job Title**	**Work Experience**	**Department**
P	Enterprise Data Architect	23 years	Enterprise Architecture
P	Manager Enterprise Architecture	7 years	Enterprise Architecture
P	IT-Consultant/IT-Manager	23 years	IT Digital Services
C & P	Data Manager	17 years	Product Digitalisation
C	Manager Data Governance Office	25 years	Finance & Controlling
P	Technical Lead/Methods & Tools	12 years	IT Product Engineering
C	IT Project Manager	10 years	Big Data & AI
P	BI & Analytics Architect	8 years	Technical Architecture Finance Analytics
C & P	Manager Enterprise Business Architecture	29 years	Enterprise Architecture
P	Manager Technology Strategy	15 years	External Consultant

Table 1. Overview of the interviewed experts. Providers (P) gave their perspectives on both the platform providing and the data providing. Consumers (C) provided insights for the consumption of data.

cussed below per stakeholder. Following the definition provided by Wieringa [27], we consider goals to be "*the desires that stakeholders are willing to commit resources to*". In contrast, the problems describe the obstacles to achieving these goals that the interviewees expect to run into that proper metadata management can address.

Stakeholder	Goal	Problems	Freq.
Data Provider	**G1** Prioritise data assets to turn into data products.	**P1 P2**	7
	G2 Create and maintain data products in a cost-effective manner.	**P2**	8
	G3 Convincingly express the value of data products.	**P1 P3 P4**	8
Data Consumer	**G4** Discover and understand relevant data products.	**P1 P3 P4**	7
	G5 Consume & combine data products.	**P5**	8
	G6 Incentivise the creation of relevant new data products.	**P1 P2 P6**	5
Platform Provider	**G7** Create and maintain metadata tools as part of easy-to-use self-serve infrastructure.	(**P1**-**P6**) **P7**	8
	G8 Extend internal platform to external data exchanges.	**P8**	2

Table 2. The different goals and corresponding problems for each stakeholder. The frequency column shows how many of the interviewed experts (out of nine) mentioned each of the goals.

Problem	Description	Freq.
P1	There is a gap between the domain knowledge of data providers and data consumers.	7
P2	It is costly to (learn how to) create and maintain data products.	8
P3	Similar or identical business objects can lead to significantly different data products.	4
P4	It is challenging to understand data product semantics.	7
P5	Combining data from different sources is technically challenging.	8
P6	Sometimes data is not available but still desired.	5
P7	End users lack data engineering expertise.	8
P8	External organisations might use different standards.	2

Table 3. There are seven problems that the interviewees expect to run into, which describe the obstacles faced by the three main roles. Proper metadata management in an internal data exchange should try to address these problems.

Data Provider The data provider has three broad goals in the data exchange, the first of which (**G1**) is to decide which data assets would make the most *valuable* data products. However, as captured by **P1**, whenever the data consumer and the data provider come from different domains, it becomes challenging to assess what data products are relevant in the context of the data consumers domain [8,13]. Lack of data engineering expertise and the costs associated with creating and maintaining data products (**P2**) reinforce the need to prioritise when creating data products.

After identifying which data assets to turn into products, the second goal for data providers in an internal data exchange is to create and maintain these data products (**G2**). The literature describing a transition towards a decentralised data exchange platform emphasises that one of the greatest challenges in this transition is the organisational "left shift" whereby domains have to take on the new functional responsibilities of providing data [3]. Our interviews confirm that it is especially challenging to onboard new data providers, particularly if the process of creating and maintaining data products is perceived to be costly (**P2**).

Ensuring that data products are valuable by prioritising is crucial to the success of an internal data exchange. However, this is not enough for the data provider, as they must also convey *why* and *how* the data product is valuable to other actors (**G3**). First among these actors are the data consumers, who might not understand the value, leading to an unused data product. However, many interviewees also explicitly expressed the need to convince colleagues and management from the provider's domain of the necessity of spending resources on creating and maintaining data products. This is because without the support from these colleagues and managers, providing data is not a sustainable activity. The gap in knowledge between data providers and consumers (**P1**) makes it hard for data providers to convey this value. In contrast, the overall difficulty of understanding the semantics of data products (**P4**) hinders the consumers' efforts to recognise it. Finally, several interviewees reinforced an idea proposed in literature [12] that it is quite difficult to figure out which data is relevant for them whenever multiple data sources are available that describe the same business object from the perspective of different domains (**P3**).

Data Consumer The data consumer also has three goals: the first one is to understand *semantically* what is being offered on the internal data exchange so they can choose what data to consume and how to do it (**G4**). As mentioned in section 3, this includes the context of the data, the meaning of the different values and attributes, but also the relevant policies and service level agreements. This goal is similar to **G3**, only viewed from the consumer's side. The problems that can be addressed by metadata management are also the same. In particular, the fact that it is challenging to find and understand relevant data products (**P4**) is reinforced by the gap in knowledge between data providers and consumers (**P1**). Additionally, as noted above, differentiating between data products from different domains can be especially challenging (**P3**).

Once the data consumer understands the semantics of the data product, their next goal is to integrate the data in their use cases, either directly or by combining it with other data (**G5**). Even if the consumer fully understands the data provider's offering, there are still technical challenges associated with consuming and combining data products. In particular, even if it is clear what each attribute in the data product means semantically, this does not automatically lead to a way to connect it to other data (e.g. through schema matching) (**P5**) [8].

Finally, our interviews revealed that data consumers wanted the ability to discover what data assets existed in operational systems and incentivise the creation of data products offering data perceived as useful (**G6**). This goal directly tries to address the problem that sometimes data is not available but still desired (**P6**). Additionally, understanding what data exists in operational systems is challenging when these systems exist in domains that are separate from the consumer (**P1**). Furthermore, Incentivisation is hindered by the perceived costs that the data consumer incurs when creating a data product (**P2**).

Platform Providers The platform providers are charged with creating the internal data exchange on which the data products are exchanged. As such, their main concern is to provide the infrastructure that allows the data providers and data consumers to achieve their goals and overcome their problems (**P1**-**P6**). However, as data providers and consumers are generally not data engineering experts (**P7**), this infrastructure should come through the creation of an easy-to-use self-serve infrastructure layer [3] **(G7)**.

In addition to addressing the goals and problems of the data providers and data consumers, however, the interviewees from the DIT also expressed the wish to expand their internal data exchange in the long term to work with external platforms such as the Catena-X initiative [1] which connects automotive data platforms (**G8**). The problem that we foresee with this goal is that standards and metadata management initiatives that are developed internally for the automotive company might not extend easily to other platforms (**P8**).

5.2 Treatment Design

Based on the goals and problems described above, we formulate five requirements that *any* approach for metadata management in a decentral data exchange should try to meet. The requirements are shown in table 4 and are discussed in detail below.

The first requirement (**R1**) is a direct consequence of the same goals and problems that motivate a transition towards decentralised data exchanges. Understanding the semantics of the data (e.g. which business processes are involved, how it is collected and for what purpose) is essential for discovering and understanding data (**G4**). Central data offices are not as familiar with these aspects of the data as data providers from the domain, who are the only actors that can capture this information in the metadata as the number of sources increases

Req.	Description	Addresses
R1.	The metadata management tools should allow data providers to capture domain expertise (i.e. semantic knowledge) as well as technical expertise (i.e. data schemas and statistics) in their models.	**G3**, **G4**, **G5**, **P1**, **P3**, **P4**, **P5**
R2.	The resulting models should allow data products to be connected on a data level, even when crossing domain or organisational boundaries.	**G5**, **G8**, **P5**, **P8**
R3.	The resulting models should relate data products semantically, even when crossing domain or organisation boundaries.	**G3**, **G4**, **G5**, **P1**, **P3**, **P4**
R4.	Metadata should be created autonomously by data providers and this should be as easy as possible.	**G2**, **G7**, **P2**, **P7**
R5.	The metadata management tools should allow data consumers to express data product requirements.	**G1**, **G6** **P1**, **P6**

Table 4. Five requirements were identified to help the actors reach their goals and overcome their respective problems. For each requirement, the goals and problems addressed by that requirement are shown in the final column.

(**G3**). At the same time, technical information, such as the data schema and describing statistics, is still necessary to consume it effectively for a use case (**G5**). Therefore, metadata management in decentral exchanges should allow data providers to explain both types of properties in a human-readable (and possibly machine-readable) manner within the same environment.

The second requirement (**R2**) has always been a main requirement for, and indeed focus of, centralised (meta-)data management. Data warehouses, in particular, address this problem by tightly coupling schemas to a global mediated schema [5]. Such an approach allows data consumers to easily consume and combine data from different sources (**G5**). However, its reliance on a central bottleneck (the mediated schema) makes it unsuited for decentral data exchanges. Even if data products cannot be tightly coupled to a single cross-domain schema, the metadata management tools should make it easy for data providers and consumers alike to connect (the schemas of) different data products.

The next requirement (**R3** shows that, for the semantic information to be truly effective in helping data consumers find, understand and consume data products (**G4**, **G5**), it needs to relate to their domain knowledge. This helps the data consumer understand the differences, similarities and nuances between business processes that often involve similar business objects (e.g. cars) but generate vastly different data and can greatly reduce the time and efforts required to decide if- and how to use that data. Moreover, if the data provider succeeds in relating their domain knowledge to that of the data consumer, it is more likely that they can convincingly express the value of their data product (**G3**).

The fourth requirement (**R4**) takes into consideration the previously noted organisational "left shift" of responsibilities from the central data office to the data provider that accompanies the transition toward a decentral data exchange (**G2**). Reducing cognitive load for platform users is a problem from cognitive science that has been well-researched for IT artefacts [24]. Therefore tools and standards should be made available that enable data providers to create and manage metadata autonomously, without direct interference from the platform providers (G7). In this sense, data products mirror microservices, which are designed to be self-contained.

The final requirement (**R5**) is not as prominent in academic literature. Still, it becomes apparent when realising that innovation in (meta)data management in most industrial settings is driven by data consumers, who feel existing shortcomings most acutely. Allowing data consumers to express and incentivise the creation of new data products (**G6**) allows faster data product development and more accurate prioritisation by the data provider (**G1**).

5.3 Proposed Solution

Based on existing literature and tools, we believe that a metadata management approach based on semantic web technology has the potential to address all the requirements identified above. In particular, envision an approach like the one suggested by Hooshmand et. al. [11]. They create a knowledge graph using domain ontologies, as well as a global ontology to describe business objects and their relations. The ontologies can capture both technical and semantical data, and data products can be added as entities by each domain and related through the graph in one of four ways, which are shown in fig. 3:

1. A direct relation between two data products, either technical or semantic. For example, two data products share some feature(s) that allows for 'join' operations, or two data products come from the same business process or operational system.
2. When both data products are in the same domain, they should be relatable through the domain ontology.
3. When the data products are from different domains, a global ontology that connects domain ontologies can help to create a relation between them.
4. As previous relations are saved in the model, it becomes possible to connect two data products through a third data product.

It is important to note that each new connection that is found reinforces the knowledge graph and improves future interoperability and findability. For example, whenever a new direct relation is found, the platform providers can use this to update the relevant domain or global ontologies. Vice versa, every time a new relation is deduced through 2-4, this adds a new direct relation between the two data products. This combination of bottom-up and top-down interoperability removes the bottleneck of monolithic approaches that can only rely on top-down interoperability.

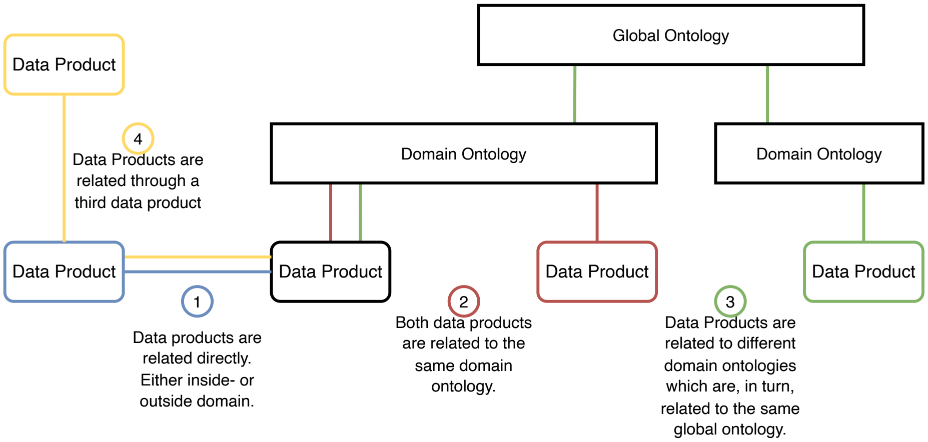
Global Ontology
Domain Ontology
Domain Ontology
Data Product
4
Data Products are related through a third data product
Data Product
Data Product
Data Product
Data Product
1
Data products are related directly. Either inside- or outside domain.
2
Both data products are related to the same domain ontology.
3
Data Products are related to different domain ontologies which are, in turn, related to the same global ontology.

Fig. 3. A high-level overview of metadata management with semantic web technology. There are four ways to connect data products, and the figure shows these in different colours. Domain ontologies define classes that can connect different data products inside a domain. Similarly, a global ontology is used to define more abstract classes of entities that persist across the company and can connect classes from the different domain ontologies. Every time two data products are connected through an ontology (2 and 3), this also creates a direct connection. Every time two data products are connected directly (1 and 4), this provides an opportunity to improve the relevant ontologies.

5.4 Validation

To validate that our proposed approach addresses all of the requirements identified in section 5.2 we developed several scenarios that encompass the problems and goals identified above. Alongside these scenarios, we created mock-ups to illustrate how our proposed solution meets these requirements. These scenarios and the mock-ups were first discussed and refined with the DMIT. Afterwards, they were presented to an audience of over 50 IT experts from across the company, whose feedback confirmed the validity of our approach.

In order to preserve space, we discuss below each of the requirements and how they can be addressed with our proposed solution. We have provided the mock-ups separately in an online repository at:
https://anonymous.4open.science/r/Data-Product-Interoperability-E633.

R1: The ability of knowledge graphs to combine technical and semantic information is well documented. To illustrate this for our experts, we created a scenario whereby data was collected from different garages that worked with the automotive manufacturer. Our mock-ups showed how the garages could create human-readable metadata using existing standards and other data products as a starting point. At the same time, they illustrated the freedom of the garages to deviate from these standards when their context or data differed from them.

R2 & R3: Although knowledge graphs are famous for connecting semantic relations, they are quite capable of connecting schemas. Our mock-ups show what these relations look like for both cases. Additionally, fig. 3 visualises how the use of domain- and global ontologies can lead to increased interoperability.

R4: Without focusing on specific tooling, we create mock-ups to show how previous information can aid the data provider with metadata creation. Our mock-ups illustrate how a data provider can find related entities in the knowledge graph (both data products and business objects). These entities, combined with the existence of standardised templates, can greatly aid the process of creating metadata.

R5: We present a scenario where a data consumer wants to consume similar data from many different data providers. By starting small, the data consumer can create a single data product in collaboration with a data provider first. Afterwards, the finished data product can work as a template for future data providers who can easily see how it will be consumed and what is expected. In addition, we note the paper mentioned in section 2 that describes how semantic web technology can be used for demand-driven data provisioning [22].

6 Discussion and conclusion

Regarding the contributions mentioned in section 2 we now discuss some of the most surprising insights that result from our work. The requirements we found were mostly consistent with those mentioned in existing literature, but the goals and problems that underlie them had not yet been discussed in detail before. Moreover, we found that an important requirement for practitioners that has been mostly overlooked in academia is the need to assign priorities to data assets that need to be transformed into data products. This prioritisation has two sides: the data providers want to validate their efforts, effectively ensuring that their created data products will be consumed. Similarly, however, data consumers want to express their needs for new data products. Literature on decentral data exchanges seems to have mostly overlooked these goals; only Stach et. al. have investigated this problem in the context of data lakes [22].

Our second finding is that creating proper data providers is a major challenge for organisations trying to transition to decentral data exchanges. Although this may not be a novel insight, the implications this challenge has on metadata management are. Ease of use has already been mentioned in academic literature. However, we find that metadata management tools should also make it easy for data providers to use existing resources (e.g. ontologies or data products) as a template. At the same time, the use of these templates should not be enforced too rigorously, and it should be easy for data providers to deviate from them whenever their ground truth demands it.

Finally, we confirm our hypothesis that semantic web technology is well-suited for metadata management in decentral data exchanges. Approaches based on knowledge graphs and ontologies work well with the decentral approach and are also capable of combining semantic and technical metadata into a single entity. Moreover, plenty of solutions exist that show how semantic web technology can be presented in a human-readable manner. Moreover, the fact that it is machine-readable opens the door for the development of automatic interoperability tools.

We acknowledge the several threats to the validity of our experiments and conclusions. First, concerning the internal validity, we note that no true data providers existed yet, in the sense that data was provided autonomously by domain teams for one or more external data consumers. We addressed this concern by interviewing extra platform providers and focusing our efforts on those who worked directly with domain teams. A threat to our findings' external validity is that they are founded on investigations inside a single company. To address this concern, we ground the findings in academic literature wherever possible. Additionally, we intend to follow up on our findings with a survey with participants across many organisations.

In this paper, we investigated metadata management for decentral data exchanges. We consider the state of the art as described in the literature and find that there has been almost no investigation of this phenomenon, especially for internal decentral data exchanges. We supplement the requirements posed by frameworks such as the data mesh with goals, problems, and requirements for industry and argue how a solution based on semantic web technology could be the way forward. In our future work, we intend to establish the external validity of our findings by surveying more professionals across different organisations. Additionally, we look forward to going beyond mock-ups by implementing and evaluating a tool for metadata creation based on our proposed solution.x

References

1. Catena-X: Automotive Network (2021)
2. Alexander, I.F., Beus-Dukic., L.: Discovering requirements: how to specify products and services. John Wiley and Sons Ltd (2009)
3. Dehghani, Z.: Data Mesh: Delivering Data-Drien Value at Scale. O'Reilly, 1st edn. (2022)
4. Dibowski, H., Schmid, S.: Using Knowledge Graphs to Manage a Data Lake. Informaitk 2020, Lecture Notes in Informatics (LNI) (January), 41–50 (2021)
5. Doan, A., Halevy, A., Ives, Z.: Principles of Data Integration. Elsevier, 1st edn. (2012)
6. Driessen, S., Monsieur, G., Van Den Heuvel, W.: Data Market Design: A Systematic Literature Review. IEEE Access **10**, 1–1 (2022). https://doi.org/10.1109/access.2022.3161478
7. Eichler, R., Giebler, C., Gröger, C., Hoos, E., Schwarz, H., Mitschang, B.: Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges. Business Information Systems (July), 269–279 (2021). https://doi.org/10.52825/bis.v1i.47
8. Fernandez, R.C., Subramaniam, P., Franklin, M.J.: Data market platforms: Trading data assets to solve data problems. Proceedings of the VLDB Endowment **13**(12), 2150–8097 (2020)

9. Gedgebuure, A.: Data mesh: Systematic gray literature study, reference architecture, and cloud-based instantiation at asml (2022), https://stefan-driessen.github.io/publication/data-mesh-systematic-grey-literature-study/
10. Hevner, A., Chatterjee, S.: Design Research in Information Systems: Theory and Practice, vol. 28. Springer (2010)
11. Hooshmand, Y., Resch, J., Wischnewski, P., Patil, P.: From a Monolithic PLM Landscape to a Federated Domain and Data Mesh pp. 713–722 (2022)
12. Koutroumpis, P., Leiponen, A., Thomas, L.: The (Unfulfilled) Potential of Data Marketplaces. ETLA Working Papers **2420**(53) (2017), http://pub.etla.fi/ETLA-Working-Papers-53.pdf{%}0Ahttp://pub.etla.fi/ETLA-Working-Papers-53.pd
13. Koutroumpis, P., Leiponen, A., Thomas, L.D.W.: Markets for data. Industrial and Corporate Change **29**(3), 645–660 (2020). https://doi.org/10.1093/icc/dtaa002
14. Lauesen, S.: Software Requirements-Styles and Techniques. Pearson Education (2002)
15. Loukiala, A., Joutsenlahti, J.P., Raatikainen, M., Mikkonen, T., Lehtonen, T.: Migrating from a Centralized Data Warehouse to a Decentralized Data Platform Architecture, vol. 13126 LNCS. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-91452-3_3, http://dx.doi.org/10.1007/978-3-030-91452-3{_}3
16. Narayan, S.: Products Over Projects (2018), https://martinfowler.com/articles/products-over-projects.html
17. Newman, S.: Monolith to Microservices : Evolutionary Patterns to Transform your Monolith. O'Reilly (2020), http://oreilly.com/catalog/errata.csp?isbn=9781492047841
18. O'Neil, B.T.: Failure rates for analytics, AI, and big data projects = 85% – yikes! (2019)
19. Otto, B., Steinbuß, S., Teuscher, A., Lohmann, S.: IDSA Reference Architecture Model. International Data Spaces Association (April) (2019), https://internationaldataspaces.org/download/16630/
20. Roman, D., Alexiev, V., Paniagua, J., Elvesæter, B., von Zernichow, B.M., Soylu, A., Simeonov, B., Taggart, C.: The euBusinessGraph ontology: A lightweight ontology for harmonizing basic company information. Semantic Web **13**(1), 41–68 (2021). https://doi.org/10.3233/sw-210424
21. Spiekermann, M., Tebernum, D., Wenzel, S., Otto, B.: A metadata model for data goods. MKWI 2018 - Multikonferenz Wirtschaftsinformatik **2018-March**, 326–337 (2018)
22. Stach, C., Bräcker, J., Eichler, R., Giebler, C., Mitschang, B.: Demand-Driven Data Provisioning in Data Lakes, vol. 1. Association for Computing Machinery (2021). https://doi.org/10.1145/3487664.3487784
23. Strengholt, P.: ABN AMRO's data and integration mesh (2020), https://www.linkedin.com/pulse/abn-amros-data-integration-mesh-piethein-strengholt/
24. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive Science **12**(2), 257–285 (1988). https://doi.org/10.1016/0364-0213(88)90023-7
25. (Thoughtworks), Z.D.: How to Move Beyond a Monothilitic Data Lake to a Distributed Data mesh (2019), https://martinfowler.com/articles/data-monolith-to-mesh.html
26. W3C: Semantic Web - Leading the web to its full potential (2015)
27. Wieringa, R.J.: Design science methodology: For information systems and software engineering (2014). https://doi.org/10.1007/978-3-662-43839-8