University of Stuttgart - Institute for Parallel and Distributed Systems / AS

From Data Asset to Data Product – The Role of the Data Provider in the Enterprise Data Marketplace

Rebecca Eichler, Christoph Gröger, Eva Hoos, Holger Schwarz and Bernhard Mitschang

In: Proceedings of the 16th Symposium and Summer School On Service-Oriented Computing (SummerSoc 2022)

BIBTEX:

```
@inproceedings{Eichler2022,
        author = {Eichler, Rebecca and Gr{\"{o}}ger, Christoph and Hoos, Eva and
        Schwarz, Holger and Mitschang, Bernhard},
        booktitle = {Proceedings of the 16th Symposium and Summer School On
        Service-Oriented Computing (SummerSoc 2022)},
        title = {From Data Asset to Data Product – The Role of the Data Provider in
        the Enterprise Data Marketplace.},
        year = {2022}
}
```

From Data Asset to Data Product – The Role of the Data Provider in the Enterprise Data Marketplace

Rebecca Eichler[1], Christoph Gröger[2], Eva Hoos[2], Holger Schwarz[1], and Bernhard Mitschang[1]

[1] University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany
{Firstname.Lastname}@ipvs.uni-stuttgart.de
[2] Robert Bosch GmbH, Borsigstraße 4, 70469 Stuttgart, Germany
{Firstname.Lastname}@de.bosch.com

Abstract. In the big data era companies have an increasing volume of data at their disposal. To enable the democratization of this data so it can be found, understood and accessed by the majority of employees, so-called data providers must first publish the data and provide provisioning options. However, a lack of incentives and increased effort for the data providers to share their data hinders the democratization of data. In this work, we present the current state and challenges of a data provider's journey, derived from a literature study as well as expert interviews we conducted with a globally active manufacturer. To address these challenges, we propose the use of an enterprise data marketplace, a platform for sharing data within the company. By presenting a functionality framework for such a marketplace and by highlighting how it can integrate with a company's data catalog, we outline how a marketplace can support the data provider. We implemented a prototype of an enterprise data marketplace and determined the feasibility of three scenarios to relieve the data provider. Finally, an assessment based on the prototype yields that the data marketplace supports the provider throughout the provider's journey, addresses major challenges, and thus, contributes to the overall goal of data democratization within enterprises.

Keywords: Data Marketplace, Data Catalog, Data Sharing, Metadata Management.

1 Introduction

Data contain the potential to provide companies with important knowledge, for example, to optimize processes or develop new business models [1]. Therefore, data democratization initiatives with the goal of empowering and motivating employees to find, understand, access, use and share data across the company [2], are gaining importance. To drive democratization aspects such as data sharing across the company, the use of enterprise data marketplaces has been proposed [3].

In general, data marketplaces are metadata-driven self-service platforms for trading data and data related services [3, 4]. Enterprise data marketplaces are specifically designed to facilitate the exchange of data and data related services within a compa-

ny [5]. Within enterprise data marketplaces company employees take on the roles of data marketplace administrators, data consumers and data providers. In the context of providing data, we distinguish three roles: Firstly, data is created by a data producer [6] which can be both a person or system, e.g., a manufacturing machine. The person responsible for this data is called the data owner [6]. Responsibility includes various aspects such as legal or technical topics, yet the owner may delegate the realization thereof to other employees such as data stewards. The data owner may be the data producer. Lastly, the employee that makes the data available is the data provider [7]. The data provider may or may not be the data owner or data producer, for example, they could be the owner but not the producer.

In an external data marketplace, in which data is exchanged across institutions, the main incentive for data providers to supply data is the monetization of data and the resulting profit [8, 9]. Within a company, however, monetization inhibits data democratization, as money presents a barrier to the data consumer. Therefore, data monetization is only envisaged to a limited degree within the enterprise data marketplace.

However, without monetization one main motivation for data providers to share data within the company disappears. Fernandez et al. [9] also discuss that providers lack information on how consumers require data and are disincentivized to share data which may leak confidential information. In addition, providers may be reluctant to share data as releasing data implies revealing own processes and quality standards [8]. Ultimately, the provider has the additional effort but no advantage by offering the data. This lack of incentives and increased effort on the part of the data providers therefore hinders the democratization of data. For this reason, in this paper we examine the role of data providers in the enterprise data marketplace, the challenges and efforts they face and how they can be supported in sharing data.

To this end, we offer the following contributions: Based on a literature study and expert interviews conducted with a globally active manufacturer, we developed *(1) a data provider's journey* which reflects the steps and roles involved in publishing and provisioning data, presented in Section 2. From this journey, we derive *(2) current challenges* the provider is faced with in Section 2 and propose the use of an enterprise data marketplace to address these. To investigate how a data marketplace assists the provider, *(3) we have developed a functionality framework* which also differentiates the marketplace functionality from other corporate data related tasks in Section 3. In the same section, we also *(4) introduce a distinction between data assets and data products* in order to leverage the existent enterprise tool landscape, particularly existing data catalogs, to support the provider. Lastly, we have developed a prototype to *(5) assess the extent to which an enterprise data marketplace supports the data provider* and addresses the challenges in Section 4. Section 5 addresses related work and Section 6 concludes this paper.

2 Providing Data in the Enterprise

In order to identify the data provider's assignments and associated processes within an enterprise, we conducted a literature study including [3, 9–12]. Yet, we found that

many articles focus on the consumer perspective as opposed to the provider perspective or only describe very abstract insight into the provider's processes. Therefore, we also conducted expert interviews with employees of a global manufacturer to gain a more detailed and practical perspective.

The manufacturer is active in a variety of sectors like the mobility or industrial sector and operates a global manufacturing network. A lot of data are collected and stored across the industrial value chain, e.g. by internet of things (IoT) devices or operational systems like enterprise resource planning (ERP) systems. The manufacturer's business strategy is to become a data-driven Industry 4.0 company and they aim to create an environment where data can be shared freely and efficiently within the company. As part of these efforts, the IT system landscape for handling data is already enhanced with tools such as data catalogs and enterprise data marketplaces are being actively investigated (see our previous work [13] for more details on the manufacturer's case and [14, 15] for details on data-driven manufacturing and Industry 4.0). The exchange with experts from various key data-related roles in an industrial enterprise, including enterprise and solution architects, as well as data scientists and business analysts, gives us a representative view of current processes for publishing data in industrial enterprises from different perspectives.

Based on the conducted interviews, we derived the data provider's processes within industrial enterprises, which we have merged into an overarching data provider journey as presented in Section 2.1. It entails the essential steps and the parties involved in the journey. Thereupon, the data provider's challenges in this journey are examined in Section 2.2. Building on this, the following sections in this paper examine the extent to which an enterprise data marketplace can address these challenges.

2.1 The Data Provider Journey

The data provider's journey of making data available in the company, as illustrated in Fig. 1, consists of two processes: *publishing* the data, i.e. making it public within the company so it can be found, and *provisioning* the data, i.e. making it available to consumers so these can work with it. These parts contain a set of steps which are carried out by various roles in the company such as the data provider, the data owner, IT or operations, legal experts or management. To illustrate the journey by way of an example, we demonstrate it with the scenario of a data steward, whose job it is to maintain data on behalf of the data owner [6]. The data steward wants to provide machine sensor data from running productions lines proactively to support machine maintenance use cases, e.g., predictive maintenance. For warranty cases concerning the machines, the sensor data are collected and stored in a database for up to 15 years.

Part one, the publishing process, comprises two segments: *documenting* the data, and *making it known* to other employees so the data can be found, understood and requested. To begin with, the data provider must *assemble documentation*. This is essentially metadata on various aspects of the data such as descriptions of the content, the data's quality or lineage, the underlying data model, technical descriptions, and lifecycle specifications. Basically, this constitutes all the information which a data consumer will require to understand and work with the data. For example, these

metadata could be descriptions on the machines which provide the sensor data, their semantics, e.g., machine temperatures or torque, and lifecycle information that these are stored up to 15 years. If the data is documented sufficiently, the data provider will next specify the *legal framework* in which the data may be used. This entails topics such as specifying access rights, the allowed usage or the data's security class, defining the data's sensitivity, e.g. whether it is ranked as public, internal or confidential. Specifications such as these are relevant for ensuring personal related data privacy and compliance to legal regulations such as the General Data Protection Regulation (GDPR) [16]. In this step, the provider may seek the assistance and guidance of legal experts. In our example, the sensor data is not personal data and ranked as internal, so all employees can access it, and there are no limitations to the usage.

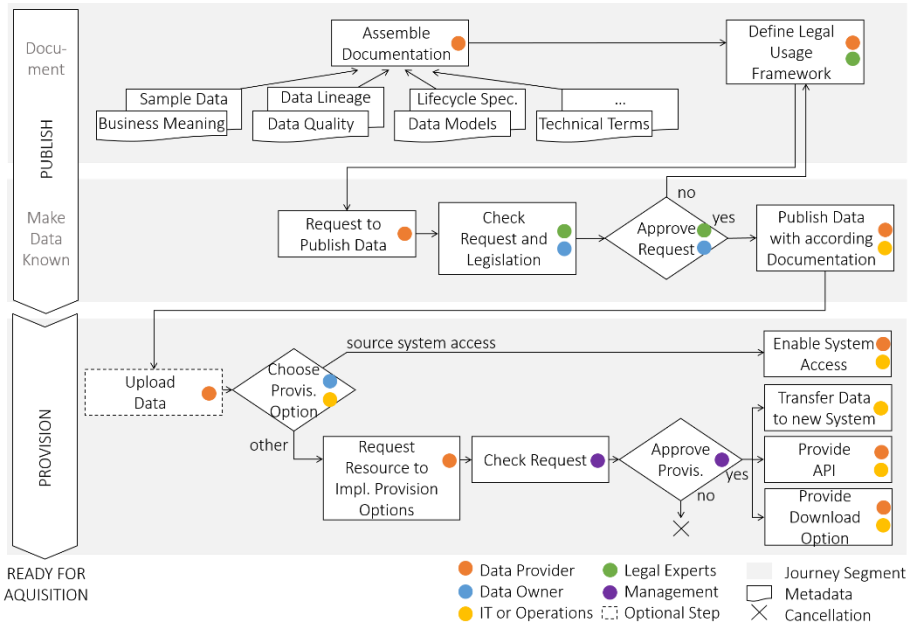

Fig. 1. The image depicts the steps and parties involved in the journey for publishing and provisioning data within an enterprise.

Having clarified the legal issues, the data must next be published, so the data can be found, understood and requested by company employees. To begin with, the data provider must issue a *request to publish data*. On the one hand, the provider must attain the consent of the data owner to publish the data. On the other hand, legal experts have to verify the authorization of the requester to release the data and whether the publication of this data is compliant with legal regulations. If the request is rejected, the legal framework must be adjusted. If it is approved, the provider may *publish the data*. This entails entering the data into an enterprise data inventory through which the data becomes discoverable. For this, a minimum amount of metadata must be provided, such as the name and location of the data source, the type, e.g., oracle data-

base, a short description what it contains and who owns the data. To enable better understanding, the metadata assembled in step one can be added as well. Assistance from IT may be required for integration into the inventory as this may require technical expertise. In continuation of our example, the steward contacts the data owner and legal experts, e.g., by email asking for the permission to publish the data. Given approval, the steward registers the data in an inventory adding information where the data source, in this case an oracle database with the machine sensor data, is located, a description of the data on the production lines, the name of the owner and so on.

At this point, employees can find and understand data through the inventory and provided metadata. The data provider now enters the second part of the journey and must provide a *provisioning* option for the data in the event that there is an access-request. As the data might currently be hosted on a local machine, it may first have to be *uploaded* to some system through which it can be made accessible, e.g., a data lake. If data is uploaded to a different system the inventory must be updated. Next, the *provisioning option must be chosen*. As the responsible person for the data, the owner has to decide with potential help from IT if *direct access to the source system* can be granted through a user account. This might, for instance, not be desirable due to a potential system overload or the risk of data manipulation in operational systems. For instance, in the case of the sensor data, direct access is not possible for risk that it may be manipulated and jeopardize the machine warranty. Based on the decision, the provider either enables system access for the data consumers or implements an alternative access method. Providing another access method may be resource intensive, e.g., by requiring a team of developers, and therefore *resources must be requested*. Given admission by management, access methods like the *transfer* of the data into another system like a data lake, or the implementation of an *API* for access without a specific user account or *download options* can be carried out by IT. By way of example, the steward requests resources to provide an API through which the machine sensor data can be accessed. As machine maintenance is of high relevance, the resources are granted and the API development approved. If management rejects the request and no provisioning option can be guaranteed, it would be useful to indicate this circumstance in the inventory or to remove the data from it accordingly. Subsequent to performing these steps, employees can find, understand and receive access to the data.

2.2 Challenges in the Data Provider Journey

Within this section, the challenging aspects for the provider in terms of cost and circumstance are derived from the provider journey, presented in the previous section.

The *(1) assembly of metadata* is the first challenge for the data provider. Although documentation is a best practice in many processes, it is often neglected. To ensure the usability of the data, however, a certain degree of documentation is indispensable. Since the provider is not necessarily an expert for the data, he has to rely on other employees such as the data producer or a data steward to provide this documentation.

Besides assembling documentation, *(2) supplying provisioning options* apart from direct system access, is also costly. This task requires an IT project, e.g., for the implementation and realization of pipelines for moving data or developing an API. This

may be a useless expense as it is unknown whether the data is of interest to other employees and hence, the provisioning options may not be required.

In practice, there are tools for publishing data which are based on a data inventory. One of these tools is a data catalog such as Alation[1] or the Collibra Data Catalog[2]. These are tools for maintaining inventories of data with discovery, administration, governance functionality and more [12, 17]. Catalogs support finding and understanding data, however, are not built to access data. For this reason, there are further publishing tools such as enterprise data marketplaces through which the data can be requested and accessed. Examples are Snowflake[3] or the Dawex Data Exchange Platform[4]. Companies are in the process of building a tool landscape for finding, understanding and accessing data using tools such as these [13]. For the provider, this means that the data must be registered in several tools such as the data catalog and the enterprise data marketplace. Therefore, challenge three refers to the effort of *(3) registering data in several publishing tools* which partly require the same metadata.

Finally, *(4) the process involves several parties* which need to be found, contacted and coordinated. With each new party the process becomes more complex and time-consuming as each introduces latencies when processing their tasks.

3 Providing Data Through the Enterprise Data Marketplace

As data marketplaces are platforms for exchanging data, they have functionality for making data available as required in the data provider journey [18]. Therefore, we examine how a data marketplace built to exchange data within a company, i.e., an enterprise data marketplace [5], can support the data provider in their journey and to what extent it addresses the providers' challenges. To examine data marketplaces' ability to support the data provider, it is necessary to understand what functionality a data marketplace offers. Therefore, we present a marketplace functionality framework in Section 3.1. Based on this framework, Section 3.2 discusses how an enterprise data marketplace can be built on existent tools in the company, such as a data catalog. In Section 3.3, we outline three provisioning scenarios in this platform tool constellation, which advantages it confers and how it works in the favor of the data provider.

3.1 Data Marketplace Functionality

The content of this subsection refers to data marketplaces in general, i.e., not explicitly to enterprise data marketplaces. We conducted a literature study to examine existing functionality lists for data marketplaces such as those presented in [5, 18–22]. According to literature and reports, marketplaces provide a wide range of functionality such as the up- and download of data [18], functionality for selling and buying data, governance topics like license management, monetization aspects like pricing,

1 https://www.alation.com/
2 https://www.collibra.com/us/en/platform/data-catalog
3 https://www.snowflake.com/workloads/data-sharing/?lang=de
4 https://www.dawex.com/en/data-exchange-platform/

revenue allocation and sharing functions [9, 22], as well as functionality for data cleansing and preparation [5, 18–20, 22] or integration [18] and analytics [21].

When comparing these lists, it is noteworthy that the extent of functionality differs. Common features include the trading and exchange aspects of data, such as buying and selling of data. Differences arise around features like data preparation which is listed on occasion or data analytics functionality which is listed in individual cases. It is also noticeable that functionality is described at different levels of detail. For example, Meisel and Spiekermann [18] provide a very detailed list of functionality, e.g., with specifics that data cleansing functionality includes duplicate and pattern recognition or plausibility check features, whereas that of Wells [5] is at a higher level of abstraction in which, e.g., data curation and preparation are the granular listed functionality. The structure also differs, with some articles listing functionality by role, i.e., data provider and consumer [19, 20] and others breaking it down by functional group [18], such as marketplace infrastructure, interfaces and security and so on.

To assess how the marketplace supports the data provider, we need an attribution of functionality by role. Furthermore, we considered the range of the above mentioned functionality in the light of the company's experience. In this context, we notice that some lists go beyond the scope of the marketplace as we understand it. From our point of view, a data marketplace is purely a broker for data and data related services. It is a platform on which data providers can publish data and services and data consumers can find, understand and gain access to these. How the data is, for instance, prepared or integrated with other data is, in our opinion, beyond the scope of the marketplace as a broker. Finally, we noticed that literature devotes little detail to the topic of metadata in the context of data marketplaces. Since finding and understanding data is a crucial feature of the marketplace, and this is dependent on metadata supplied by the data provider, we consider metadata management to be a relevant underlying and role independent functionality in the marketplace. Therefore, we have created a functionality framework that takes these three aspects into account: the division by role, the delineation of functionality that lies within and outside the marketplace, and the metadata management which is the basis for the role-specific features.

To create the functionality framework, we incorporated the common parts from the different lists and allocated them to the roles of the data consumer, data provider and administrator. The functionality that was only partially represented, such as rating or data cleaning, were examined whether they belonged in the functional scope of a broker and were accordingly included or excluded. In this context, we also had to find a common level of abstraction that subsumes the more detailed tasks. Additionally, we extended functionality such as the necessary metadata features derived for the role-based functionality. The resulting functionality framework is illustrated in Fig. 2.

The Functionality Framework. The functionality framework, shown in Fig. **2**, displays the marketplace's functionality in the blue box and other functionality outside of it. Data governance and data management including data quality or data lifecycle management take place outside its functional scope as these concern the management as opposed to the exchange of data. Equally, activities that follow the acquisition of data such as data preparation or exploration are out of scope as these

involve data processing which goes beyond data sharing. To enable an integrated data processing toolchain, the marketplace nevertheless provides interfaces to tools that perform tasks outside the marketplace context, such as preparation tools.

Within the marketplace, we distinguish between *consumer-side functionality*, *provider-side functionality* and *administration functionality*. This functionality is accessible through a *portal*, i.e., a graphical user interface, and an API. The *metadata management functionality* and *privacy, security and compliance* extend across these areas.

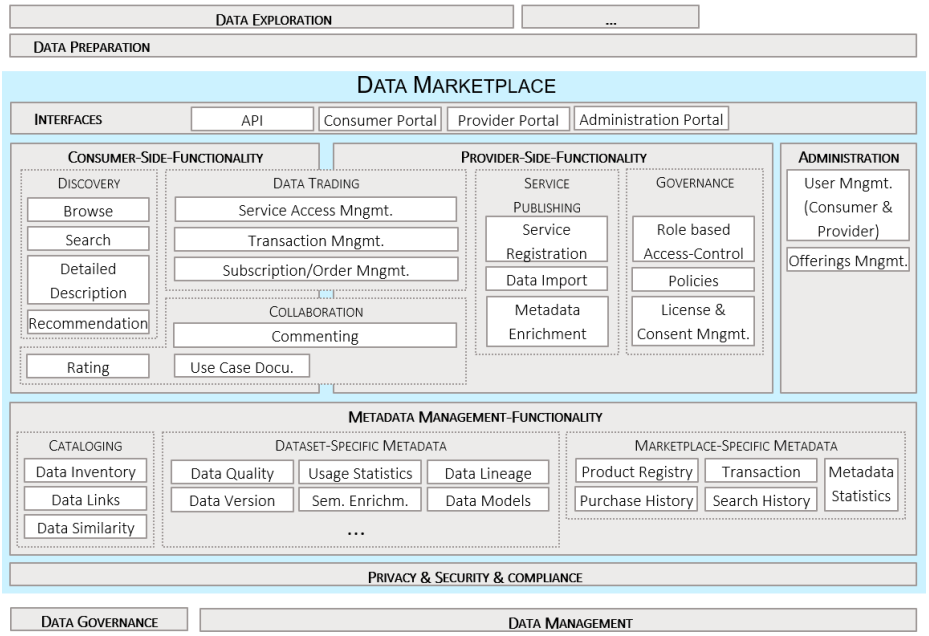

Fig. 2. Data Marketplace Functionality Framework

Consumer-Side-Functionality. The consumer-side functionality includes *discovery*, *data trading* and *collaboration* features. The consumer can *browse* or *search* for data and services in the marketplace, like the machine sensor data provided by the data steward in the provider journey example. For each search result, there is a *detailed description* with an integrated view of all available metadata. For example, the detailed description could contain a description on the machine and the according production line with technical details how and where the data is stored or operational information such as the data's lineage. The marketplace can also offer service *recommendations* based on the conducted search, previous acquisitions and those of similar users. In order to support data democratization in the sense of collaboration and knowledge sharing [2], the marketplace also offers functionality such as *commenting* to both the consumer and producer. Furthermore, consumers can *rate* data and *document their use case*, thereby enabling other users to see if the data has been used for similar use cases. In our example, a user could specify how they used the data in a

machine maintenance use case. These functionalities are only available to the consumer and thus, placed in the consumer-side box in Fig. 2. The data trading functionality like the collaboration functionality are overarching in Fig. 2, as they are available to both the consumer and producer. On the consumer-side *service-access-management* signifies the ability to request and receive access and access credentials, e.g., by ordering data through a shopping cart. Additionally, the consumer can *manage transactions* related to reimbursements for services, as well as active, expired and pending orders through the *subscription and order management*. In continuation of the example, the user can order the machine sensor data through the shopping cart and thereby request access, then pay for this data through the transaction management and after receiving access, view the active subscription on this data with according details.

Provider-Side-Functionality. The provider functionality involves *publishing*, *governance* and *data trading* functionality. For publishing services such as data-as-a-service or professional services like courses for data preparation, the provider uses the *service registration*. In this step, the marketplace adds the service, mostly a data source or a specific dataset, to the marketplace service inventory so it can be found via the search. In our example the data steward registers the machine sensor data in the marketplace which is thereby added to the marketplace inventory. Although the marketplace will reference most data as opposed to storing it, it does have a *data import* feature for cases like the upload of a locally stored singular csv file. In our example, the database in which the machine sensor data is stored is registered instead of uploading the data directly. The *metadata enrichment* allows the provider to add additional metadata to ensure a better understanding of the data. This may include a variety of metadata such as a content description, technical details as well as information about the data provenance. These could for instance be descriptions on the machine and production line, the database and the lineage showing how the data originates in the machine and is moved to the database by a specific script. Besides service registration, the marketplace offers governance functionality which supports the specification as well as compliance to aspects defined within the data's legal usage framework explained in the provisioning journey. A provider can define *role-based access control*, usage rights within *policies* and package these in *licenses* for specific services. This functionality does not replace the underlying data governance, it merely enables the implementation of the marketplace-specific governance aspects. For instance, the decision who is allowed to do what with the machine sensor data is part of the governance outside the marketplace. Within the marketplace the steward in our example merely specifies that only people from department x are allowed to use it for maintenance use cases. Like the consumer, the provider has functionality to support them in the trading of data. For instance, the provider can manage access requests, i.e., receive notifications, consult an overview and accept or decline these requests. If monetisation or other forms of reimbursement are included in the marketplace the provider can monitor transactions for the offerings. For example, the steward can view closed and outstanding invoices for the sensor data. Having provided access to the data, a provider can then handle the subscriptions and orders on the offered services. This includes an overview of who is subscribing to which data, options to contact all subscribers or

functionality to terminate subscriptions and revoke access rights. In terms of collaboration, the provider can also enter into a comment dialog with the data consumers.

Metadata Management Functionality. A data marketplace requires a variety of metadata to support the above-mentioned functionality. This comprises general metadata for *cataloging* as known from a data catalog tool like an inventory, *dataset-specific metadata* such as a business description and *marketplace-specific metadata* like the purchase history. Metadata for cataloguing can refer to a range of datasets and helps to provide an overview of existing offerings. It includes a *data inventory*, e.g., a list of contained data sets, *data links* that indicate whether data sets are related as well as *data similarity* information, which reveals replicated and similar data sets. Dataset-specific metadata refers to a specific dataset and helps users to understand and trust this data. Amongst others, this covers *data quality, lineage* and *versions*. It is important to understand that the maintenance of the dataset-specific metadata is not part of the marketplace, merely, that it is relevant for the consumer in the sense of finding and understanding data. Therefore, this metadata has to be supplied by the provider and the marketplace must support some form of indexing and integrated processing and presentation of these. The marketplace-specific metadata comprises a *product registry, purchase history, transaction history* and *search history*. *Metadata statistics* are also marketplace-specific metadata and indicate to what extent the metadata is complete or contain user statistics such as an indication how often a service has been viewed. The regulation of privacy, security and compliance, and the administration features are not discussed in detail due to lack of space. With this framework we have gained insight which explicit functionality a marketplace should offer to the user and which implicit functionality like the metadata management is required.

3.2 Data Catalogs as a Foundation for Enterprise Data Marketplaces

When examining the listed metadata management functionality within the functionality framework, it becomes apparent that there is a considerable overlap with functionality offered through a data catalog. Besides data asset inventories these have discovery, administration, governance, collaboration functionality and more [12, 17]. They contain a large part of the metadata also required in the data marketplace. Since a marketplace also requires an inventory, a data catalog is thus a component of the data marketplace [5]. Inversely this means, the data marketplace builds on a data catalog and extends it with further functionality like data trading features.

Nowadays, companies have one, or are in the process of building data catalogs [13]. Accordingly, when an enterprise data marketplace is developed, this platform can be built on top of the existing data catalogs and use them as their data inventory. Thereby, functionality is reused and extended as opposed to duplicated. Furthermore, the catalog metadata can be reused in the marketplace. This means, the marketplace can read the catalog entries so these are found in the marketplace search.
In order to enable access to the data, the marketplace requires metadata which is not part of a data catalog. This includes, for instance, details on the *provisioning options* such as an API, download or source system access and the according *access proce-*

dures that go with these options. Also, *contractual information* such as the *price*, if data is monetized, the *license*, or *subscription options* as well as the *terms of use* such as the *permitted usage* or *conditions of use*. We call these metadata product metadata. They include all information which is required to sell or make data available for access and use. In this sense, we distinguish between *data assets* and *data products*. The distinction is displayed in Fig. 3. As the term suggests, data as an asset has a potential financial value for a company [23]. They are registered and maintained through a data catalog and are therefore, enriched with a minimum of metadata for finding and understanding them, such as the *content description*, *lineage* or *data owner* [12]. Data products are data assets which have been enriched through the data marketplace with product-specific metadata and are thus ready to be accessed and provisioned. Metadata to both of these types belong to the dataset-specific metadata in the functionality framework, in Fig. 2. To conclude, this means that data assets can be found through the data catalog and data products through the marketplace. If the marketplace builds on the data catalog and uses it as its inventory, then data assets can also be found in the marketplace, even if they are lacking product metadata. In order for consumers to gain access to these assets, the provider must however, first turn them into a data product by enriching these with product metadata.

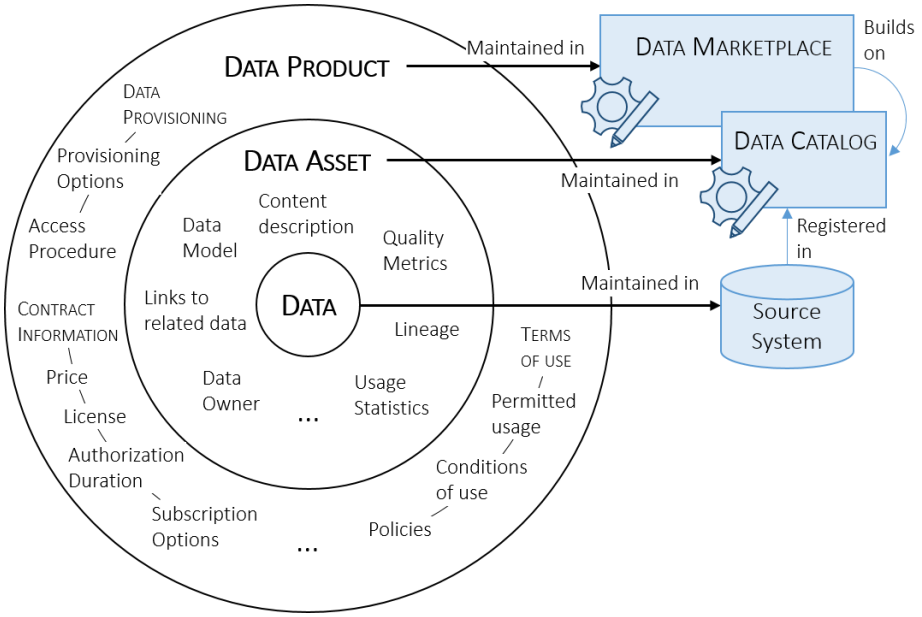

Fig. 3. The figure illustrates the distinction of data assets and data products with exemplary metadata, as well as the systems in which these are maintained. Metadata which are connected through dashes belong to a specific topic that is portrayed though capital letters.

3.3 From Data Asset to Data Product

Data assets can be transformed into a data product in different ways, therefore, we illustrate three main transformation scenarios. Within the first scenario, the provider *explicitly registers the data in the marketplace* and directly specifies all the product metadata, such as the permitted usage, license etc. By implication, the data is then also registered in the catalog. In the second scenario, the data provider *registers the data within the catalog* and does not concern himself with the data marketplace. The data is therefore a data asset. Now, some employee, e.g., a data scientist, can search for data in the marketplace and finds the data asset. The employee can then send a request to access this data to the provider who is then prompted to specify the product metadata. Having turned the data asset into a data product, access can be granted to this data. The third scenario assumes that *another employee can fill in the required product metadata* and send a request for asset-product transition to the provider. For instance, a data steward may know the product metadata and can fill this in for the data owner. The owner is notified and can accept or reject the proposed metadata. If accepted the data is turned into a product, if not it remains a data asset which cannot yet be accessed. As underlined by these three scenarios, the distinction of data assets and data products yields several advantages:

- The marketplace references data even if it has not been explicitly registered in it, but only in the data catalog.
- Consequently, the providers initially only have to register the data in the catalog so it can be found and understood within the enterprise.
- The provider only has additional effort for adding product metadata and providing provisioning options when the data are actually relevant and are requested.

4 Assessing how an Enterprise Data Marketplace Assists the Role of the Data Provider

In this section, we examine the extent to which an enterprise data marketplace supports the provider in making data available and whether the marketplace addresses the challenges (1-4). Existent solutions such as the mentioned Snowflake or Dawex Exchange Platform do not support a seamless integration with a company's existent tool landscape through out of the box loose coupling with existent data catalogs. Therefore, we developed an enterprise data marketplace prototype to demonstrate and assess the feasibility of the ideas presented in the Section 3. The prototype is an extension to our work presented in [24]. It is built with the Spring framework[5] and based on a micro services architecture including a search service, product service and order service. It is implemented on the open-source data catalog Apache Atlas[6] which registers the data assets. Product metadata is stored in the marketplace's metadata reposito-

[5] https://spring.io/
[6] https://atlas.apache.org/

ry, set up with a Neo4J[7] graph database and the metadata is modeled according to our metadata model HANDLE [25]. The enterprise data landscape is simulated by a variety of databases and a data lake which are registered in the data catalog.

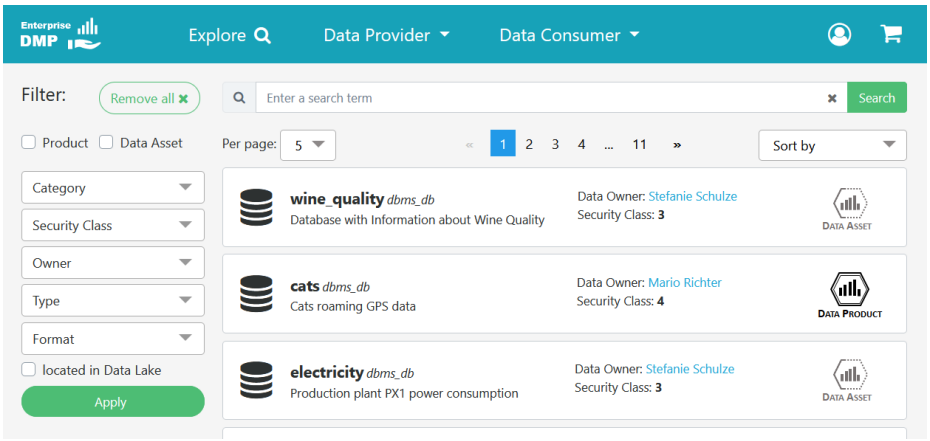

Fig. 4. Data Asset and Data Product Distinction in the Dataset Search

4.1 Prototypical Demonstration – From Data Asset to Data Product

As argued in Section 3.2, it is beneficial to build the enterprise data marketplace on a company's established data catalog. In this case, we built the enterprise data marketplace on Apache Atlas and use it as the data catalog for the marketplace. If a search query is issued in the marketplace it is forwarded to Atlas. The corresponding search results are displayed in the marketplace search results view. The marketplace can identify the data assets for which it contains the corresponding product metadata and labels these as data products, as can be seen in Fig. 4 on the right hand side of each search result. As explained in the following, the prototype supports several scenarios as to how data is provided and an asset becomes a product.

Scenario one. As specified in Section 3.3, the data provider can register the data product in the marketplace. To do this, the provider can select "Add new Product" in the menu under "data provider" and is directed to a data registration wizard. The wizard guides the provider through 3 steps, as displayed in Fig. 5. The first step prompts the provider to specify whether the data is already registered as an asset and if so, to enter the asset-id. In the second step the provider is led to a form for either registering or editing the asset if it already exists. In this case, the form fields are prepopulated with the metadata loaded from the catalog. If the data is not registered as an asset, the provider fills out the form with according metadata such as a data description, data owner, security class and so on. The metadata is sent to the catalog in which it is registered as an asset. The provider is then guided to the third wizard page for adding

[7] https://neo4j.com/

product metadata, as illustrated in Fig. 5. This step is optional as the data can already be found through the marketplace search. To register the product, the *terms of use* are specified in which the provider indicates if it is *personal data*, the *permitted usage*, *conditions of use* and a *license*. For example, the GDPR allows people to influence how their personal data is processed [26]. By way of example, the data is personal and the processing has been restricted to the evaluation of user statistics which is specified in the field *permitted usage*. Also, only persons from a specific team may access this data, therefore this is also specified in the field *conditions of use*. If none of the licenses fit the requirements, the provider can also create a customized license. Besides the terms of use, the *data delivery options* are specified. This includes information on the data's *update cycle*, the *provisioning options* and *description of the access procedure*. Having specified this information, the provider clicks on the button "Add As Product" so the metadata is stored in the metadata repository, creating a data product.

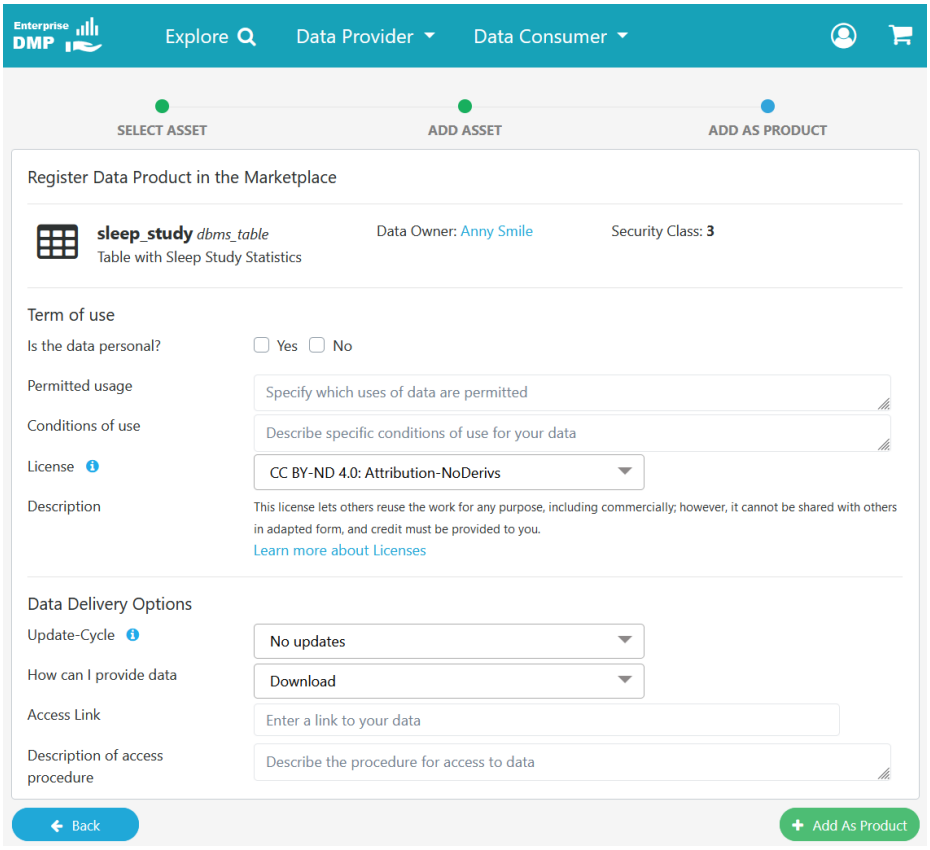

Fig. 5. Wizard for Enriching Data Assets with Product Metadata to Create Data Products

Scenario two. In the second scenario, data is only registered through the data catalog. For this purpose, the provider dials into the Atlas GUI and fills in the corresponding

form fields for registering data assets. As explained above, this data can then be found in the marketplace and is flagged as an asset. If this dataset is then requested in the marketplace by a data consumer, the provider receives an access request and is prompted to add the product metadata and is automatically forwarded to the corresponding form fields in the wizard as depicted in Fig. 5.

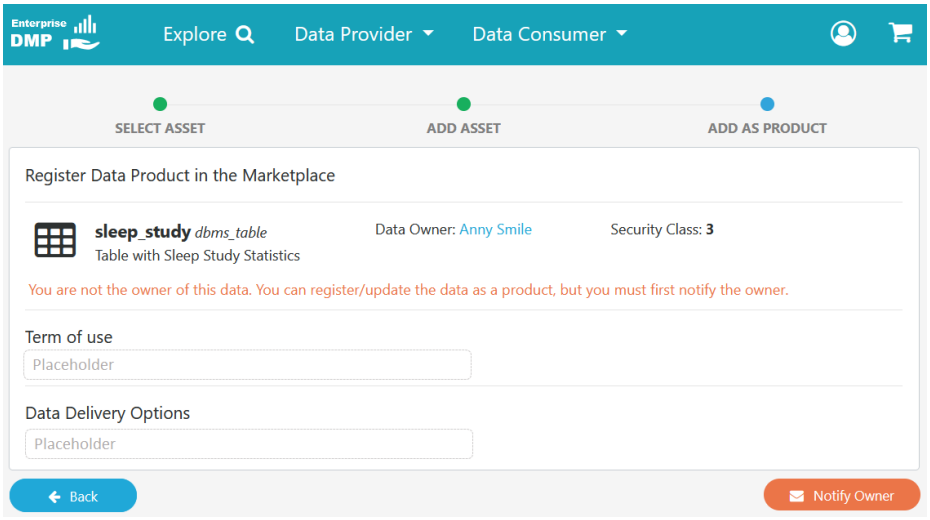

Fig. 6. This image depicts the form for creating data products which is shown if the user is not the data owner. For space reasons the form fields are replaced by placeholders. The form is identical to the form displayed in Fig. 5.

Scenario three. The third scenario entails that another employee can fill in the required product metadata for the provider. If a user selects a search result as shown in Fig. 4, they are taken to a detailed page with overview metadata providing detail on the dataset. This detail page also specifies whether the result is an asset or a product. In the case that it is an asset, a button is displayed “add product metadata”. Clicking on this button will take the user to the registration wizard where the user can navigate to the “add product” form. If the user is not the data owner, this is displayed with a message as shown in Fig. 6. The form is submitted using the “Notify Owner” button. The owner then receives the request and can accept, reject or edit the metadata before creating the product. As shown in this chapter, a data marketplace can be built on top of an external data catalog and the three scenarios as described in Section 3.3 are supported in this catalog-marketplace constellation.

4.2 Addressing of the Challenges in the Data Provider Journey

In this section, we discuss to which extent the enterprise data marketplace addresses the challenges in the data provider journey, as given in Section 2.2. The first challenge signifies the *assembly of documentation*, i.e., metadata. In effect, this task is

supported to a certain extent through tools which can automatically capture metadata. For instance, the data catalog Alation uses AI to suggest business glossary terms and suggests links to relevant data [27]. Since this concerns the first step of the provider journey and the marketplace is only utilized throughout later steps, the assembly of metadata is not supported though the marketplace.

Challenge two refers to the *effort of supplying provisioning options*, even if these may not be required. This issue is addressed by the enterprise marketplace through the differentiation of data assets and data products. It is dealt with by allowing the provider to supply product metadata and thereby make provisioning options available only when a request is made for a data asset. Therefore, the effort relating to provisioning options is only undertaken if this data is actually relevant for other employees.

Challenge three deals with the necessity of *registering data in several publishing tools*, namely the data catalog and the data marketplace. Whether this challenge is addressed by the marketplace depends on the implementation approach that is chosen. The marketplace can be built as a standalone platform with its own inventory. In this case, challenge three is not addressed, data must be registered in both tools, and some metadata must be maintained twice. As explained in Section 3.2, the implementation alternative involves integrating the marketplace with a company's existing data catalog. If data catalogs are used as an inventory for the marketplace, so it can find the data assets that are registered in them, the provider only has to register data in the data catalog. This avoids the need to register data in more than one tool. In addition, as the marketplace reads metadata from the data catalog, the duplication of the same metadata and the duplicate administration of these is avoided. Hence, this implementation option addresses the challenge of double data registration and metadata maintenance. It can also be added that users who are not data owners or dedicated data providers can suggest product metadata, which eliminates the need for the data provider to do this. In this case, the data provider only has to accept or reject the request.

That the data provider's journey *involves several parties* which have to be found, contacted and coordinated constitutes challenge four. There are two steps, which involve a request to third parties that can be partially automated through the marketplace. This includes the *request to publish data*. For this, however, the owner and the legal experts must be known and specified. If this is the case, the marketplace represents a platform via which a workflow for the request and approval of such processes can be implemented. The same is true for the *request for resources.* If the people from management are known and can be identified in the marketplace, then the marketplace can also ensure a regulated workflow for the resolution of this subject matter.

Consequently, all of the challenges are addressed through various tools and the marketplace specifically addresses the challenges two through four.

5 Related Work

Data marketplaces have been investigated in various contexts. Some research focuses on the overall picture and identifies the main characteristics of data marketplaces [23, 28–31], trends and emerging markets [23, 28, 30], as well as challenges [23, 28] and

research fields [32] around data marketplaces. Other research examines the marketplace in a more specialized context such as marketplaces for open data [33] or marketplaces for data of the internet of things [34–38]. The application of different technologies in the data marketplace, such as the use of distributed ledger technology is also studied [35, 39]. These research articles cover a wide range of topics, but do not discuss the distinguishing features of enterprise data marketplaces.

A report published by Wells [5] distinguishes internal marketplaces, which are synonymous to enterprise data marketplaces, and external markets. Within this report the characteristics, components and services, and involved technologies of enterprise data marketplaces such as data catalogs or data lakes are highlighted. However, how these can be used to an advantage within the enterprise data marketplace is not discussed. The same is true for the article by Gröger [3] in which he presents the enterprise data marketplace as a central element in the data ecosystem of industrial enterprises. Like Wells, Fernandez et al. [9] differentiate internal marketplaces in their work based on the data exchange boundaries and the incentive to share data and discuss challenges and the research agenda for constructing data marketplaces. However, they do not take into account the specifics of embedding a marketplace into a company's tool landscape. How the data marketplace can be integrated into a company's system landscape and how it can be used to its advantage has not been explored.

Data marketplaces are metadata-driven platforms [3] and the necessity for metadata management in data marketplaces is expressed in several research articles. For instance, metadata is discussed in the context of data trading challenges and provenance [19], data integration [18] or decentralized marketplaces and storing it in the blockchain [35]. The functional frameworks presented in [18] and [5] also list metadata management as a required feature. Most of the research around data marketplaces, however, only provides a high-level view on metadata management. In contrast, [40] introduces a detailed metadata model for describing data goods, to facilitate the selection and trading of data and Fernandez et al. [9] describe a metadata engine for maintaining the lifecycle of datasets. To the best of our knowledge, none of the work provides a detailed overview of necessary metadata management features in marketplaces or take peculiarities of enterprise data marketplaces into account.

As discussed in Section 3.1 several research articles provide a list of data marketplace functionality [5, 9, 18–22]. However, as already explained, the aspect of metadata management functionality is not addressed sufficiently, and the delimitation of the tasks of the data marketplace as a data broker is not clearly defined.

The role of the data provider is differentiated in a number of research articles such as [18, 23, 32, 41]. For instance, Lange et al. [32] differentiate the role of the data provider and derive a provider challenge, such as the difficulty of pricing data when lacking knowledge on the data value for the consumer. Furthermore, they introduce several marketplace types based on different data providers such as commercial, public or private data providers. Yet, these research articles do not examine the provider in the enterprise context. In contrast, Wells [7] discerns three types of providers for the enterprise data marketplace, the internal providers, people and systems in the company, open data providers which supply free external data and commercial data providers that offer fee or subscription-based external data. But he does not look at

the processes and specific challenges that data providers face in the enterprise. Fernandez et al. [9] consider the provider in internal and external marketplaces and tackle the provider challenge that sharing data is hard as the providers lack information and incentives to make data available so it increases the consumer's utility. They propose bonus points or time as an incentive of internal providers to share data. However, further specifics of the enterprise data marketplace and the provider's processes therein are not examined. Hence, this article has covered this gap in existing literature by examining the current processes for providing data within an enterprise and the corresponding challenges and how these are addressed by an enterprise data marketplace.

6 Conclusion

Data democratization initiatives with the goal to facilitate a broader availability and accessibility of data within a company are becoming increasingly important. The data provider journey we presented illustrates the current processes for providing data within an industrial enterprise and the challenges a provider faces which impede data democratization. In this work, we propose the use of an enterprise data marketplace to support the data provider throughout this journey. Our marketplace functionality framework illustrates the supplied functionality, including functionality for the provider and shows that an enterprise data marketplace is based on metadata management functionality. Through a prototypical implementation we demonstrate the integration of a marketplace with an existent data catalog, the differentiation of data assets and data products, and how this enables several application scenarios which support the data provider in publishing and provisioning data. Consequently, we have demonstrated how the enterprise data marketplace can leverage the existent tool landscape to ease the publication and provisioning of data and is, therefore, a platform which enables data democratization within enterprises. In future, we intend to investigate incentivation mechanisms for data providers to share data within the enterprise and how the marketplace can leverage further tools and systems in the enterprise system landscape such as a data lake, a business glossary or knowledge graph.

References

1. Cao, L.: Data science: A comprehensive overview. ACM Comput. Surv. 50, 1–42 (2017).
2. Lefebvre, H., Legner, C., Fadler, M.: Data democratization : toward a deeper understanding. In: Proc of the International Conference on Information Systems (ICIS) (2021).
3. Gröger, C.: There is no AI without data. Commun. ACM. 64, 98–108 (2021).
4. Stahl, F., Schomm, F., Vossen, G., Vomfell, L.: A classification framework for data marketplaces. Vietnam J. Comput. Sci. 3, 137–143 (2016).
5. Wells, D.: The Rise of the Data Marketplace: Data as a Service. Eckerson Gr. (2017).
6. Abraham, R., Schneider, J., vom Brocke, J.: Data governance: A conceptual framework, structured review, and research agenda. Int. J. Inf. Manage. 49, 424–438

(2019).
7. Wells, D.: Dynamic Data Marketplace Fast Data for Fast Business. Eckerson Gr. (2018).
8. Trauth, D., Van Ouwerkerk, N., Mönckemeyer, F.: Putting a Price Tag on Data. Springer Berlin Heidelberg (2021).
9. Fernandez, R.C., Subramaniam, P., Franklin, M.J.: Data Market Platforms: Trading Data Assets to Solve Data Problems. Proc. VLDB Endow. 13, 1933–1947 (2020).
10. Gröger, C., Hoos, E.: Ganzheitliches Metadatenmanagement im Data Lake: Anforderungen, IT-Werkzeuge und Herausforderungen in der Praxis. In: Proc of the 18. Fachtagung für Datenbanksysteme für Business, Technologie und Web (BTW 2019) (2019).
11. Zeng, J., Glaister, K.W.: Value creation from big data: Looking inside the black box. Strateg. Organ. 16, 105–140 (2018).
12. Labadie, C., Legner, C., Eurich, M., Fadler, M.: FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs. In: Proc of the IEEE 22nd Conference on Business Informatics (CBI). pp. 201–210 (2020).
13. 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. In: Proc of the 24th International Conference on Business Information Systems (BIS). pp. 269–279 (2021).
14. Hoos, E., Gröger, C., Kramer, S., Mitschang, B.: ValueApping: An analysis method to identify value-adding mobile enterprise apps in business processes. In: Proc of the Conference on Enterprise Information Systems (ICEIS). pp. 222–243 (2014).
15. Gröger, C., Schwarz, H., Mitschang, B.: The manufacturing knowledge repository consolidating knowledge to enable holistic process knowledge management in manufacturing. Proc. 16th Int. Conf. Enterp. Inf. Syst. 1, 39–51 (2014).
16. General Data Protection Regulation (GDPR), https://gdpr.eu/tag/gdpr/, last accessed 2022/03/28.
17. Zaidi, E., De Simoni, G., Edjlali, R., Duncan, A.D.: Data Catalogs Are the New Black in Data Management and Analytics. Gartner. (2017).
18. Meisel, L., Spiekermann, M.: Datenmarktplätze - Plattformen für Datenaustausch und Datenmonetarisierung in der Data Economy. Fraunhofer ISST. (2019).
19. Koutroumpis, P., Leiponen, A., Thomas, L.: The (Unfulfilled) Potential of Data Marketplaces. ETLA Work. Pap. (2017).
20. Roman, D. et al.: ProDataMarket: A data marketplace for monetizing linked data. Proc. ISWC 2017 Posters Demonstr. Ind. Tracks co-located with 16th Int. Semant. Web Conf. 1963, (2017).
21. Saxena, S.: Enterprise Data Marketplace: Democratizing Data within Organizations. Tata Consult. Serv. (2018).
22. Spiekermann, M., Lehmann-Brauns, S., Tontsch, R., Otto, B., Hoffmann, M.: Datenmarktplätze in Produktionsnetzwerken. Plattf. Ind. 4.0. (2020).
23. Spiekermann, M.: Data Marketplaces: Trends and Monetisation of Data Goods. Intereconomics. 54, 208–216 (2019).
24. Eichler, R., Gröger, C., Hoos, E., Schwarz, H., Mitschang, B.: Data Shopping - How an Enterprise Data Marketplace supports Data Democratization in Companies. In:

[Submitted] to the 34th International Conference on Advanced Information Systems Engineering (CAiSE) (2022).
25. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., Mitschang, B.: Modeling metadata in data lakes—A generic model. Data Knowl. Eng. 136, (2021).
26. Art. 6 GDPR - Lawfulness of processing - GDPR.eu, https://gdpr.eu/article-6-how-to-process-personal-data-legally, last accessed 2021/09/07.
27. Alation: Data Stewards, https://www.alation.com/solutions/data-stewards/, last accessed 2022/03/12.
28. Driessen, S.W., Monsieur, G., van den Heuvel, W.-J.: Data Market Design : A Systematic Literature Review. In: [To be Published] (2022).
29. Schomm, F., Stahl, F., Vossen, G.: Marketplaces for data: An initial survey. ACM SIGMOD Rec. 42, 15–26 (2013).
30. Stahl, F., Schomm, F., Vomfell, L., Vossen, G.: Marketplaces for Digital Data: Quo Vadis? Comput. Inf. Sci. 10, (2017).
31. Fruhwirth, M., Rachinger, M., Prlja, E.: Discovering business models of data marketplaces. In: Proc of the 53rd Hawaii International Conference on System Sciences (HICSS) (2020).
32. Lange, J., Stahl, F., Vossen, G.: Datenmarktplätze in verschiedenen Forschungsdisziplinen: Eine Übersicht. Informatik-Spektrum. 41, 170–180 (2018).
33. Zuiderwijk, A., Loukis, E., Alexopoulos, C., Janssen, M., Jeffery, K.: Elements for the Development of an Open Data Marketplace. In: Proc of the Conference for E-Democracy and Open Governement. pp. 309–322 (2014).
34. Zheng, Z., Mao, W., Wu, F., Chen, G.: Challenges and opportunities in IoT data markets. In: Proc of the 4th International Workshop on Social Sensing (SocialSense 2019). pp. 1–2 (2019).
35. Ramachandran, G.S., Radhakrishnan, R., Krishnamachari, B.: Towards a Decentralized Data Marketplace for Smart Cities. In: Proc of the IEEE International Smart Cities Conference (ISC2). pp. 1–8 (2018).
36. Alrawahi, A.S., Lee, K., Lotfi, A.: AMACoT: A Marketplace Architecture for Trading Cloud of Things Resources. IEEE Internet Things J. 7, 2483–2495 (2019).
37. Krishnamachari, B., Power, J., Kim, S.H., Shahabi, C.: I3: An IoT marketplace for smart communities. In: Proc of the 16th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys 2018). pp. 498–499 (2018).
38. Schmid, S. et al.: An architecture for interoperable IoT Ecosystems. In: Proc of the 2nd International Workshop on Interoperability and Open-Source Solutions for the Internet of Things (InterOSS-IoT 2016). pp. 39–55 (2016).
39. Roman, D., Stefano, G.: Towards a reference architecture for trusted data marketplaces: The credit scoring perspective. In: Proc of the 2nd International Conference on Open and Big Data (OBD 2016). pp. 95–101. IEEE (2016).
40. Spiekermann, M., Tebernum, D., Wenzel, S., Otto, B.: A metadata model for data goods. In: Multikonferenz Wirtschaftsinformatik (MKWI 2018). pp. 326–337 (2018).
41. Muschalle, A., Stahl, F., Löser, A., Vossen, G.: Pricing Approaches for Data Markets. In: Proc of the International Workshop on Business Intelligence for the Real-Time Enterprise (BIRTE). pp. 129–144 (2013).