Journal of Big Data

RESEARCH Open Access

Introducing the enterprise data marketplace: a platform for democratizing company data

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

*Correspondence:
rebecca.eichler@ipvs.uni-stuttgart.de

[1] Universität Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany
[2] Robert Bosch GmbH, Borsigstraße 4, 70469 Stuttgart, Germany

Abstract

In this big data era, multitudes of data are generated and collected which contain the potential to gain new insights, e.g., for enhancing business models. To leverage this potential through, e.g., data science and analytics projects, the data must be made available. In this context, data marketplaces are used as platforms to facilitate the exchange and thus, the provisioning of data and data-related services. Data marketplaces are mainly studied for the exchange of data between organizations, i.e., as external data marketplaces. Yet, the data collected within a company also has the potential to provide valuable insights for this same company, for instance to optimize business processes. Studies indicate, however, that a significant amount of data within companies remains unused. In this sense, it is proposed to employ an Enterprise Data Marketplace, a platform to democratize data within a company among its employees. Specifics of the Enterprise Data Marketplace, how it can be implemented or how it makes data available throughout a variety of systems like data lakes has not been investigated in literature so far. Therefore, we present the characteristics and requirements of this kind of marketplace. We also distinguish it from other tools like data catalogs, provide a platform architecture and highlight how it integrates with the company's system landscape. The presented concepts are demonstrated through an Enterprise Data Marketplace prototype and an experiment reveals that this marketplace significantly improves the data consumer workflows in terms of efficiency and complexity. This paper is based on several interdisciplinary works combining comprehensive research with practical experience from an industrial perspective. We therefore present the Enterprise Data Marketplace as a distinct marketplace type and provide the basis for establishing it within a company.

Keywords: Data Catalog, Data Democratization, Data Market, Data Sharing, Enterprise Data Marketplace, Metadata Management

Introduction

An enormous amount of data is generated in the big data era by, for instance, the Internet of Things (IoT), social media networks, transactional processing systems, or wearables and mobile devices [7]. This data contains a potential value which may lead to, e.g., new insights. The data value can, however, only be extracted if the data is available for use in, e.g., data science and analytics projects. In this context, the data marketplace (DMP) is gaining in importance. Data marketplaces are electronic platforms for trading data as well as data-related services [35, 38]. They provide infrastructure for the data exchange by acting as a digital intermediary connecting data providers and data consumers [38]. Data marketplaces yield several advantageous outcomes. For instance, they stimulate innovation as consumers can acquire data which would have been unavailable and available data can initiate the improvement of products, services, or processes or also the development of new business models [30].

Data marketplaces are mainly considered for the exchange of data and services between organizations or private individuals. There are, however, also other relevant application scenarios for data marketplaces, such as their deployment within a company. Studies show that over half of the data goes unused within companies [50]. In this context the FAIR principle, i.e., making data findable, accessible, interoperable and reusable [34, 57], as well as data democratization are discussed in literature. Data democratization has the objective to motivate and empower the majority of company employees to find, understand, access, use, and share data within the company, in consideration of data security and compliance [5, 37]. Lefebvre et al. [37] define four data democratization dimensions. The first describes the enablement of broader access to data and tools for users with varying skill-sets, the second signifies the development of data-related and analytic skills such as data cleaning. The third dimension covers collaborative knowledge-sharing between employees, and the fourth entails the promotion of data value like communicating the importance of data assets. In this context, it has been proposed to employ the data marketplace within a company in order to address data democratization and the corresponding dimensions [15].

The term data market is, in the economic sense, the setting in which data providers and consumers meet to exchange data and related services against a form of compensation. The term data marketplace refers to the platform built to facilitate this exchange. In the company-internal context, the data marketplace is referred to as an Enterprise Data Marketplace [26, 55] or an internal data marketplace [19]. In extension of Wells [56] definition, we propose the following:

The Enterprise Data Marketplace is a type of data marketplace for the exchange of data and data-related services between company employees, and optionally invited business partners. It has the objective to democratize data within the company. This does not only involve making data available but explicitly addressing the data consumers' information needs so they can obtain access to data how they require it. To promote data democratization the Enterprise Data Marketplace (EDMP) offers the full scope of a company's data, not only selected datasets. This includes data from different domains, data in varying processing degrees, and also data insights such as reports or machine learning models. In a company's system landscape the EDMP is a mediating instance, facilitating the availability of data in data storage systems ranging from operational systems like enterprise resource planning (ERP) systems over analytical systems like data lakes and data warehouses. Apart from data storage systems, the EDMP must also be distinguished from existing tools such as data catalogs. These provide an inventory of data assets over the above mentioned storage systems and facilitate finding and understanding the contained data [59]. The EDMP complements the data catalogs with additional functionality such as features for requesting and managing access to data. In an example usage scenario a data scientist is looking for customer and product data to gain insights on which customers buy which products in order to provide targeted advertisements. The data scientist can use the EDMP to search for such data and find that there is data according to his/her requirements stored in a data lake. The data scientist requests access to this data through the EDMP and receives access to available provisioning options for this data. Optionally the data scientist can request the data with additional software or infrastructure like a virtual machine so they are directly supported throughout their use case.

The EDMP has, however, been studied very little in literature and researchers have highlighted the need for further conceptual and practical research to reveal its capabilities and value-adds [29]. Topics in this regard also include how it differs from other types of data marketplaces or how such an EDMP is built. Therefore, we make the following contributions: Besides the provided definition, (1) *we position the Enterprise Data Marketplace in a classification framework* differentiating it from other marketplaces and thereby provide a type distinction. (2) *We present requirements for data marketplaces* and highlight which are specific to an EDMP. Based on these requirements (3) *we provide an Enterprise Data Marketplace platform architecture*. In extension, (4) *we also discuss how the Enterprise Data Marketplace integrates in the existent enterprise system and storage landscape*. To demonstrate how the presented concepts can be realized, (5) *we showcase an Enterprise Data Marketplace prototype*. Lastly, we conduct an experiment based on this prototype (6) *evaluating the impact of introducing an Enterprise Data Marketplace* in a company. The content of this paper is based on several interdisciplinary works we compiled throughout assorted research projects, combining comprehensive research with practical experience from an industrial perspective [13, 15, 16, 22, 23, 26].

The paper is structured as follows: The introduction is followed by the Sect. Related work and the Sect. Classifying the enterprise data marketplace in which the EDMP is classified based on a data marketplace classification framework. Subsequently, general and EDMP specific data marketplace requirements are specified in Sect. Enterprise data marketplace requirements, followed by the Sect. Distinguishing the enterprise data marketplace from the data catalog. An EDMP architecture is presented in Sect. Platform architecture and Sect. Enterprise integration contains a discussion how this marketplace can be integrated in the enterprise system landscape. A prototypical implementation is illustrated in Sect. Prototypical demonstration and in Sect. Evaluating an enterprise data marketplace an experiment is presented, conducted to evaluate the extent to which an EDMP improves the data consumer processes in a company. Finally, Sect. Summary and conclusion concludes this paper.

Related work

The Enterprise Data Marketplace is addressed in only a few research articles. Amongst others, Gröger [26] highlights the need for this specific marketplace type, Fernandez et al. [19] consider them to bring down data silos, and Wells [55] defines and presents the EDMP in a report. Driessen et al. [12] present data marketplace types with problems and solution approaches, one of which is called *the generalist* and can be established within a single large company and thus encompasses, but is not limited to the EDMP. We also discuss the necessity and various aspects of EDMPs in our previous research [13, 15, 16]. Azcoitia and Laoutaris [6] introduce the *embedded data marketplace* type, which signifies an add-on to a data management system within a company. As large companies often build on a number of data management systems, the embedded marketplace is limited in its scope of data, and the authors point out, these are often limited in their functionality. Hence, they are similar but not equivalent to the company-wide EDMP. Jahnke and Otto [29] identify the EDMP as one class of data catalog application. They highlight the EDMP as a research gap for which further details on its capabilites and value-adds have yet to be determined. Lastly, Zasadzinski et al. [60] present how they implemented a data platform as a basis for an EDMP in a report. None of the above articles clearly highlight the specifics and differences to external marketplaces. Therefore, we close this gap by placing the EDMP in a data marketplace classification framework and provide the specific requirements and architectures.

There are several research articles that provide classification frameworks for data marketplaces. These include Schomm et al. [47] who provide an initial set of dimensions and Stahl et al. [53] that extend these. Meisel and Spiekermann [38] derive five classification characteristics and Spiekermann [49] provides economic and technological characteristics of data marketplaces. Täuscher and Laudien [54] list key business model attributes of marketplaces, which are however not exclusive to data marketplaces, and Azcoitia and Laoutaris [6] classify data exchange entities including data marketplaces through business model attributes. Fruhwirth et al. [20] provide a list of characteristics that are assigned to dimensions such as value capture, delivery, proposition and creation. So far, the EDMP has not yet been classified based on any of the frameworks, hence we provide this placement in such a framework.

Requirements for data marketplaces are listed in a range of research articles. Fernandez et al. [19] introduce requirements concerning topics such as the ability to price datasets or the ability to support markets of different types like internal and external markets. Sometimes the requirements are tailored to a specific context such as metadata management in decentralized data exchanges [11], trustworthiness through, e.g., blockchain [36] or data marketplaces in the IoT context [33, 46]. While requirements are often listed in a specific context such as IoT many still apply to data marketplaces in general, for example, requirements concerning scalability or security [2]. Requirements for the EDMP could be derived from this marketplace type's descriptions as supplied in, e.g., [15, 26, 55], and general requirements also partly apply to the EDMP. It has, however, not been clarified which explicit requirements the EDMP has and how these overlap with those of other marketplaces.

In terms of data marketplace architectures, there is a variety of architectural proposals, most of which are, however, tailored to a specific context, similar to the requirements. There are data marketplace architectures specific to the use of blockchain [45, 48], the IoT context [2, 33, 46], multilateral marketplace design [32], elements in decentralized marketplaces [43], personal data valuation [31], or also specific data marketplace aspects like a market management system or mashup builder [19]. None of these architectures reflect the specific components of the EDMP. In contrast Wells [55] provides a component overview for the EDMP, nonetheless, it is not apparent which aspects are special to the internal setting or also how the components interact. Similarly, components of an EDMP's underlying data platform are illustrated in the report [60], yet the distinction between the EDMP and data platform components is not clear, nor how it leverages and is embedded in a company's existent system landscape. In contrast, Gröger [26] places the EDMP in the data ecosystem of an industrial enterprise and Wells [55] gives an overview of required technologies in the EDMP, both do not however explain how the EDMP and its components interact with the other systems.

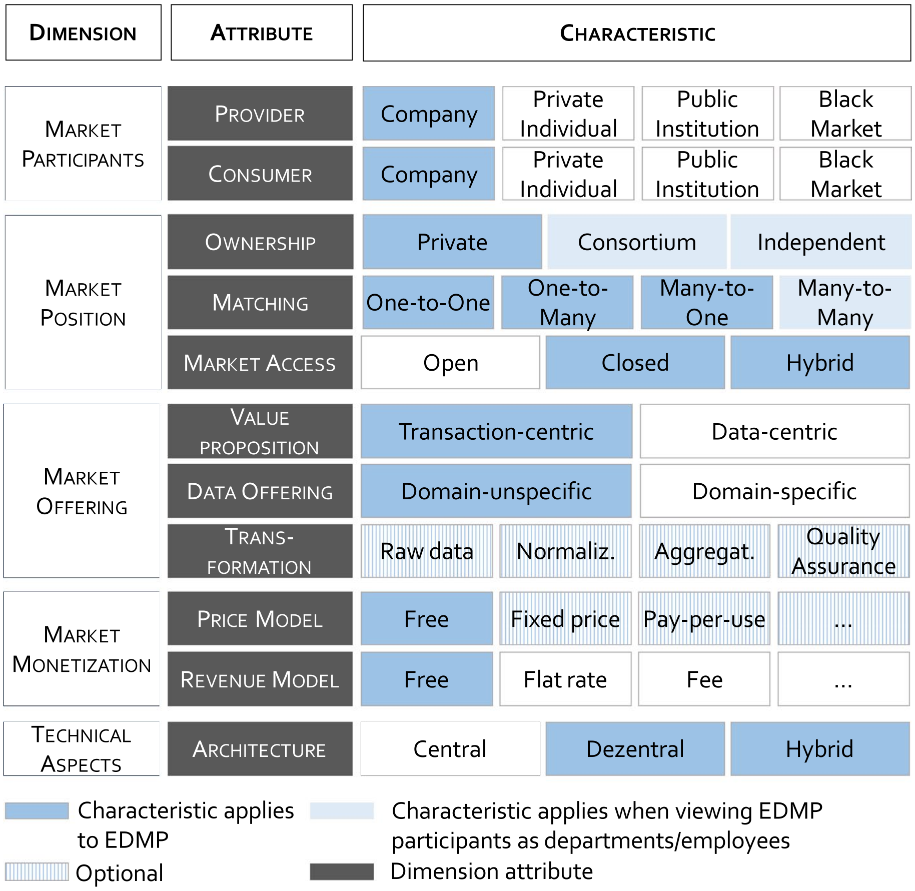
Finally, we want to point out the relation of the data marketplace to similar concepts and research areas. These include the data mesh, the data fabric, and the data space. The data mesh is a new organizational paradigm for handling analytical data [10]. It is a decentralized approach for managing, sharing and accessing analytical data at scale, mainly inside but also beyond enterprises. Like the data mesh, the data fabric is an approach to facilitate managing, sharing and accessing of corporate data across a hybrid data landscape [39]. As opposed to the data mesh, which is an organizational approach, e.g., by dividing an organization into domains, the data fabric is technology driven by combining interchangeable tools and technologies to achieve the above mentioned goals [39]. In both, the topic of self-service is addressed through some kind of data platform. We see the EDMP as a component in this self-serve data platform, offering services such as data registration, discovery and access. While the data mesh and fabric are approaches that focus on enabling data management and sharing mainly within an organization, the data space is a data-sharing ecosystem across organizations [41]. It facilitates sharing data in a secure and trusted way, based on standards and collaborative governance models and specifically focuses on preserving the digital sovereignty of data owners over their data. In the data space context, data marketplaces can constitute data providers that contribute data into the data space community [40]. Hence, a variety of data marketplaces can be connected to the data space. Relating to data space components, we also see an overlap in functionality of the data space's broker service provider role. Similar to a data marketplace, it stores and manages information, i.e., metadata, about the available data sources and acts as an intermediary connecting data providers and data consumers [40, 41]. Yet, as opposed to the data marketplaces, the broker service provider is not involved in the exchange of data [41]. Furthermore, there may be several of these brokers in a data space, whereas the data marketplace as a broker would constitute one central component.

Classifying the enterprise data marketplace

In order to identify the distinguishing characteristics of the EDMP, we position it in a classification framework for data marketplaces. The framework is presented in Sect. The data marketplace classification framework and the identified characteristics are discussed in the following Sect. Enterprise data marketplace characteristics. By highlighting the distinct features, we introduce the EDMP as a marketplace type.

The data marketplace classification framework

In this section, we present a classification framework designed to highlight the specific characteristics of the EDMP. On the one hand, the framework identifies the EDMP as a distinct type of data marketplace and on the other hand, it can be used to determine whether data marketplace solutions are suited for the use within an enterprise as an EDMP or if they are more suited as an external data marketplace for use between enterprises.



DIMENSION	ATTRIBUTE	CHARACTERISTIC			
MARKET PARTICIPANTS	PROVIDER	Company	Private Individual	Public Institution	Black Market
	CONSUMER	Company	Private Individual	Public Institution	Black Market
MARKET POSITION	OWNERSHIP	Private	Consortium	Independent	
	MATCHING	One-to-One	One-to-Many	Many-to-One	Many-to-Many
	MARKET ACCESS	Open	Closed	Hybrid	
MARKET OFFERING	VALUE PROPOSITION	Transaction-centric	Data-centric		
	DATA OFFERING	Domain-unspecific	Domain-specific		
	TRANS-FORMATION	Raw data	Normaliz.	Aggregat.	Quality Assurance
MARKET MONETIZATION	PRICE MODEL	Free	Fixed price	Pay-per-use	...
	REVENUE MODEL	Free	Flat rate	Fee	...
TECHNICAL ASPECTS	ARCHITECTURE	Central	Dezentral	Hybrid	

🟦 Characteristic applies to EDMP
⬜ Characteristic applies when viewing EDMP participants as departments/employees
▥ Optional
⬛ Dimension attribute

Fig. 1 Data marketplace classification framework highlighting the characteristic-profile of the EDMP in Blue

To classify the EDMP, we studied data marketplace characteristics. As outlined in the previous section, these are provided through various research articles such as [6, 20, 32, 35, 38, 47, 49, 52–54]. The characteristics range from aspects like marketplace ownership over the value proposition, data access methods, monetization aspects to the underlying architecture. Meisel and Spiekermann [38] provide a classification framework by combining characteristics identified through various research articles including [32, 35, 47, 52]. Spiekermann [49] also provides a data marketplace classification framework based on a taxonomy developed explicitly for classifying data marketplaces based on their business models. By combining both of these frameworks an overview covering various dimensions of data marketplace characteristics can be obtained. Hence, we developed the data marketplace classification framework as displayed in Fig. 1 by combining both of these frameworks. We extended the resulting framework with the attribute *consumer* for the sake of completeness and renamed a few attributes and corresponding characteristics. These include the characteristic *company*, which is called "commercial" in the original source. As the term commercial signifies both a business interest and cash flow, yet the cash flow does not represent the participant, we renamed it company which complements the characteristics *private individual* and *public institution*. Also the attribute "market positioning" [49] is replaced through the more expansive attribute *ownership* of [38] and the attribute "integration" [49] is renamed to *data offering*. By grouping the attributes, we receive five dimensions based on which an EDMP can be classified: the *market participants*, the *market position*, the *market offering*, *monetization and technical aspects*. The characteristics that apply to the EDMP in these dimensions are highlighted in two shades of blue.

Enterprise data marketplace characteristics

For the attributes defined in the framework, one, several or none of the characteristics may apply to the EDMP.

Market participants involve both the data and service *providers* as well as the *consumers* in the data marketplace. In the case of the EDMP the participants in both categories are employees within the same *company*, this is not immediately apparent through the classification framework. In some cases, an enterprise may choose to open their EDMP to selected business partners [56], which also classify as a company.

The *market position* signifies who *owns or operates* the data marketplace, the *matching*, i.e., the number of parties involved, together with the service orientation among these, as well as the *accessibility* of the data marketplace. As the EDMP mainly contains enterprise internal data, including classified and personal data, it is usually owned and operated by the same company, hence is *private*. In this context, the company also bears the costs of operating the EDMP. Considering not the entire company, but its departments or employees as participants, it can be argued that it is either a *consortium* or *independent* EDMP depending on whether the department operating the EDMP is an active participant. Therefore, all three characteristics are highlighted. In the same sense, it is a *one-to-one* matching, considering the entire company exchanging data and services with itself, or a *one-to-many* or *many-to-one* matching, if business partners are involved and the company is either sharing with or receiving data and services from them. The *many-to-many* matching refers to the company's departments or employees trading data amongst each other. Depending on whether the EDMP is accessible only to the company employees or also to invited guests, it is *closed* or *hybrid* respectively.

The dimension *market offering* constitutes the *value proposition*, *data offering* and *transformation functionality* in the marketplace. The EDMP's value proposition is *transaction-centric* as its core offering is the switching function of data and services, i.e., bringing data providers and consumers together. It only forwards the consumer to tools for data analysis, visualization and preparation and does not incorporate this functionality and is therefore not data-centric, according to [49]. The scope of offered data spans across all company data, hence, the data offering is *domain-unspecific*. According to Spiekermann [49] transformation refers to the data marketplace's ability to transform raw data into a normalized or aggregated state or assure data quality. While we argued in [16] that a data marketplace does not offer functionality to process data, e.g., aggregate it, the marketplace can offer data in various transformation states, e.g., data in data lake zones in varying processing degrees. Hence, these characteristics are marked as optional, as they are not essential for classifying the EDMP.

As *monetization* of data offerings would hinder the EDMP's goal of democratizing data within a company, the *price model* for most offerings is *free*. There may be instances in which a cash flow between separate business units is required for legal reasons, or if data is sold to a business partner, therefore, the EDMP may support any other form of price model as well. The *revenue model* signifies under which monetary conditions participants can use the data marketplace. As a revenue model would be a barrier for employees to use the EDMP, and therefore hinder data democratization, the revenue model is *free* in the EDMP.

With the goal of democratizing most enterprise data, it is feasible to retain data in the source systems, as opposed to storing it redundantly in a centralized data marketplace repository. Therefore, the EDMP has a *decentralized* data storage architecture. However, to support the registration of, e.g., a single report or file which should not be stored in any other storage system a *hybrid* approach with both a centralized and decentralized repository can be chosen.

Concluding, a data marketplace that meets these criteria is classified as an Enterprise Data Marketplace. By highlighting its distinct characteristics, we have exposed the EDMP as a type of data marketplace. This type of marketplace also has its own set of requirements, which we discuss in the following section.

Enterprise data marketplace requirements

Having identified an EDMP's characteristics, we now specify requirements concerning this data marketplace's offering in terms of data and services, functionality and as this marketplace is operated within a company, requirements to how the EDMP should integrate with the existent enterprise system landscape. The requirements are derived and generalized from existing literature on data marketplaces and data democratization. The practical relevance of these was also validated through a case study with a large industrial company, including an enterprise-practice point of view. The company in question, is a globally active manufacturer, striving to become a data-driven Industry 4.0 company and is therefore building a tool landscape including an EDMP (for details on the case study see [13]). In the following Sects. Required data marketplace offerings-Enterprise integration requirements, we highlight and explain which requirements are specific to the EDMP and which are relevant for data marketplaces in general. The relevance of the various requirements for the marketplace types is also shown and consolidated in Table 1.

Required data marketplace offerings

The term offerings refers to the items, or in this case to the services, which a consumer can acquire in the data marketplace. As mentioned in the introduction, it is the objective of an EDMP to address data democratization, which implicitly sets the baseline for the required offerings.

In order to facilitate the data democratization dimension of broader access to data [37], all kinds of data have to be made available within the company [26]. Therefore, the EDMP's main offer must be *Data-as-a-Service* [55]. Ultimately, the EDMP should make all corporate data available. This includes data from operational systems such as ERP systems as well as analytical systems like data lakes. Both internal company and externally acquired data are included in this. Likewise, raw data, data in various processing

Table 1 Relevance of requirements in the DMP and EDMP

Requirement		DMP	EDMP
Service Offerings	Data-as-a-Service	+	+
	Infrastructure-as-a-Service	o	+
	Software-as-a-Service	o	+
	Professional Services	o	+
Functionality	Consumer-Side	+	+
	Provider-Side	+	+
	Administration-Side	+	+
	Metadata-Management	+	++
	Privacy & Security	+	++
Enterprise Integration	Data Storage Systems	+	++
	Metadata Management Tools	-	++
	Administrative Tools	-	++

- irrelevant, o not specifically relevant, + relevant, ++ specifically relevant

degrees as well as ready-to-use data and data insights such as machine learning models or reports, belong into this scope. As explained in Sect. Enterprise data marketplace characteristics, the data is not limited to a domain such as finance or manufacturing.

The definitions of data democratization also specify that the data must be made available to all kinds of users, i.e. also non-specialist users [5, 26]. This type of users may lack the skills for setting up the required infrastructure and software, or only have skills to work with data in specific tools. Hence, the EDMP should also offer *Infrastructure-as-a-Service*, and *Software-as-a-Service* in combination with the data. For instance, a user may order data with infrastructure like a virtual machine. The EDMP could provide the virtual machine such that it contains the data as well as the required software for a data preparation or analysis task. The user could also have the data provided directly in a tool such as a Tableau[1] or Microsoft Power BI[2] instance. Thereby, the EDMP supports self-service consumption of data. Any marketplace can offer these services, yet they are relevant in the EDMP to achieve broader access to tools for users with varying skill-sets which is part of the first data democratization dimension.

The development and sharing of data skills is part of the second data democratization dimension [5, 37]. Hence, the EDMP should also offer *Professional Services*. These are services offered by users with specific skills and can, for example, involve training courses to acquire skills for processing data, dashboarding or data preparation.

While all these offerings are not exclusive to an EDMP, they are relevant for it because of the democratization objective of this type of marketplace.

Required data marketplace functionality

Based on the general functionality framework for data marketplaces we present in [16] there is role-based functionality for the *consumer, provider* and the *administration*. In addition, data marketplaces offer cross-sectional functionality which includes *metadata management* as well as handling issues of *privacy and security*. A condensed version

[1] https://www.tableau.com/.

[2] https://powerbi.microsoft.com/.

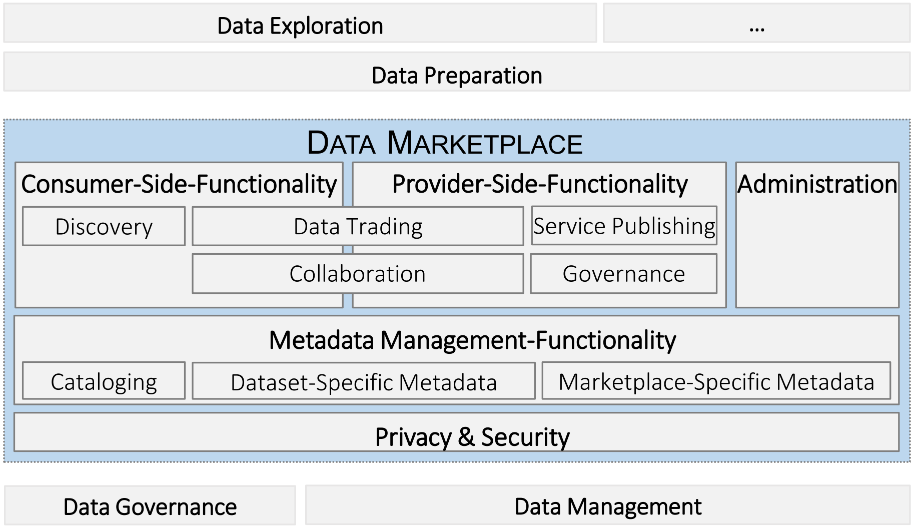


Fig. 2 Data marketplace functionality framework (Based on Eichler et al. [16])

of the framework is illustrated in Fig. 2. Besides depicting the functionality within a data marketplace it also shows which functionality is not part of the marketplace. This involves *data governance and management* topics, as these concern the management as opposed to exchange of data, as well as all topics which follow the acquisition of data, such as *data preparation* as these are beyond the exchange of data. From our point of view the data marketplace is merely a broker which offers data and can provide a stepping-stone to data-related tasks through training courses or by providing infrastructure. While most of the functionality listed is also required in other data marketplaces, we discuss in the following how some of the functionality may be specifically relevant or require specialized solution approaches in the EDMP.

The role-based functionality is not necessarily specific to EDMPs, yet also required therein. The consumer requires *discovery* features such as a search function and detailed description of the offerings. They also need access to *collaborational* features to, e.g., rate or comment on data and discuss the suitability of the offered data. Both the consumer and provider need *data trading* features. These include features like order management, e.g., to request or provide access to offered data, place requests for new data offerings, for consumers to manage their acquired data and services, and for the provider to manage the running subscriptions on their offered data and services. The provider also requires features for *service publishing*, such as a service registration, e.g., for registering data in the marketplace, or data import features for uploading data. *Governance* features are required by the provider to retain data sovereignty and offer the data compliantly. Administration requires features to manage users and offerings in the data marketplace.

In contrast, the metadata management functionality is distinctive in EDMPs. Data marketplaces are metadata-driven platforms, therefore the handling of metadata is a central aspect within these. It includes building a data *catalog* with an inventory of data and services offered, the collection of *metadata specific to these datasets* such as descriptions, quality metrics, the data model etc., and also storing *marketplace-specific* *metadata* on the marketplaces internal processes like the purchase and search history therein. Companies already have infrastructure that collects and manages a wide variety of metadata, for instance, with tools such as data catalogs or business glossaries [13]. In the company, the EDMP thus has significantly more metadata at its disposal. Furthermore, the EDMP can be tailored to reflect enterprise idiosyncrasies. For example, companies often have a company-internal "language," i.e., specific vocabulary, which is maintained through tools like business glossaries. By way of example, a company may refer to an end product as "material." Yet normally the term "material" refers to a product's elements. In an EDMP this vocabulary can be incorporated in the description of the dataset. In this sense, the EDMP is more flexible than other data marketplaces, which cannot, for instance, support a "customized" language across various companies.

Like metadata management, privacy and security aspects are especially relevant in the EDMP due to the scope and value of data that is registered in the EDMP. While selected datasets are made available through an external data marketplace, the entire scope of company data is registered in the EDMP, which includes both highly confidential and sensitive data. The ISO/IEC 27000 series [28] defines standards for information security in companies, concerning protection goals such as confidentiality, integrity, availability, and authenticity. Accordingly, these protection goals also have to be addressed in an EDMP. Due to the intrinsic properties of an EDMP, it is sufficient to use standard technologies for some of these protection goals. For instance, an EDMP is less likely to be subjected to attacks such as distributed denial-of-service attacks, as it is accessible to mainly internal and thus, more trusted users. Therefore, no special protective measures are required for such types of attacks. The appropriate protective measures for an EDMP may include data encryption for confidentiality [21], digital signatures to realize data integrity [61], proof of retrieveability to address availability [24], and attribute-based signatures to ensure authenticity [25]. Other protection goals such as privacy, are more challenging to fulfill in the EDMP, as a significantly larger amount of privacy-relevant data has to be taken into account which are requested for a variety of use cases. For instance, in order to trade personal data regulations such as the general data protection regulation (GDPR) require the consent of the data subject for this exchange [17]. The EDMP's data includes most of the personal data in the company, which was collected and approved for certain purposes. Therefore, the EDMP has to ensure that it is used and shared for these purposes only, or in an altered version to comply with the GDPR. That is, some parties may access the entire datasets, other parties may access an anonymized or distorted version of the data, and some may not be allowed to know that this data exist. However, by distorting the data the data quality may be affected [3], e.g., by removing parts of the dataset or adding noise to the dataset. Yet, a company may need to ensure privacy in accordance with GDPR in a variety of use cases without compromising the quality of its data. For these reasons, issues of remaining compliant with legal regulations like GDPR may be more challenging and significant in the EDMP. In this regard, we have investigated topics like the demand-oriented generation of data products in consideration of data privacy [51]. Data products can be generated using privacy filters for extracting privacy critical information without distorting the overall data quality. For instance, there are specialized privacy filters for location data [1, 4], images [18, 58], and time series data [42, 44]. As this constitutes an extensive research area, security and privacy aspects are subject to future work. In contrast, metadata aspects will be the focus of this paper and are discussed in more detail in the following. This metadata is also relevant to decide which privacy filters can and have to be applied to the underlying data in order to enable a trustworthy and demand-oriented handling of the data [9].

Enterprise integration requirements

External marketplaces for trading data between organizations are often stand-alone marketplaces and usually merely support a selective light-weight integration with enterprise internal systems. In contrast, the EDMP should tightly integrate with a large variety of different enterprise IT Systems in the company's system landscape, in order to incorporate existent functionality as well as existent data and metadata. In this sense, we present the following set of integration requirements.

To begin with, it should *integrate with existing data management and storage systems*. This may include operational systems like ERP systems as well as analytical systems like data warehouses or data lakes. The ability to reference data in various data management systems is not per se specific to an EDMP. An EDMP should, however, be able to reflect peculiarities of such a system or reflect data in a customized way according to the source system. For instance, it could reflect a data lakes customized zone architecture such as [22] and reference the data accordingly throughout the zones.

As mentioned previously, there are a variety of metadata management tools that are used to manage data and the understanding thereof within a company. These tools include data catalogs, business glossaries, and model repositories. Some of these tools provide functionality which is required in a data marketplace. The data catalog, for example, contains a data inventory, which is also required within a data marketplace. The business glossary and other tools contain metadata which is relevant for finding, understanding and consequently choosing data for use. This information can be reused within an EDMP. Therefore, the EDMP should tightly *integrate with the existent metadata management tool landscape*, build on existing functionality and incorporate the existing relevant metadata.

There are also administrative systems in companies such as identity management systems for managing company employees, or systems that deal with the corresponding employee access rights. By *integrating with administrative tools* single sign-on and authorization management across source systems, including the EDMP is possible. The EDMP can then also access existent information in the user profiles such as an employee's clearance level and reuse this, e.g., to filter appropriate search results.

Distinguishing the enterprise data marketplace from the data catalog

Having identified the distinct characteristics and requirements of the EDMP in the previous sections, we clarify how the EDMP is different from a data catalog in this section. These two tools are metadata-based systems [26, 59] and are very similar, due to an overlap in functionality and offerings. Furthermore, the understanding of data catalogs has evolved in the past years, and thereby, the discernment to the data marketplace has become less clear. Hence, we intend to facilitate a uniform understanding of the EDMP throughout the rest of this work by clarifying how these two tools differ.

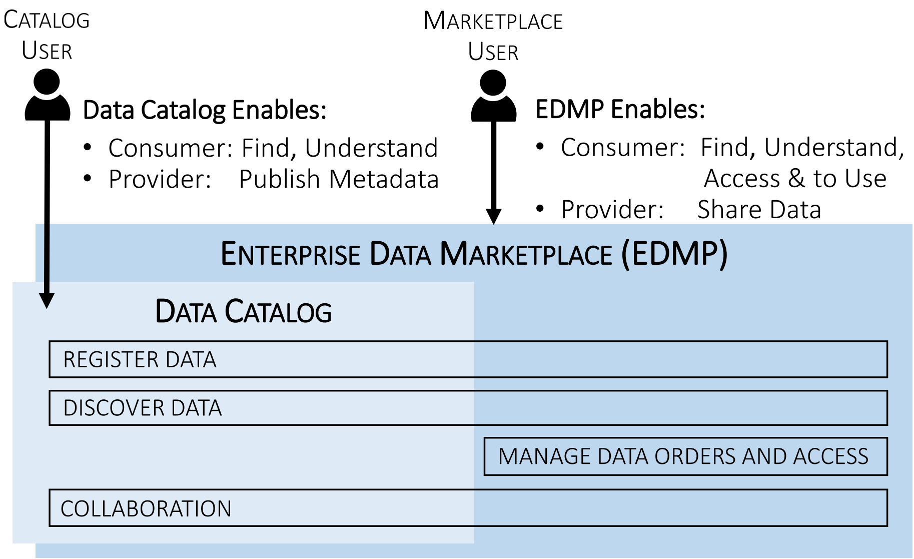

CATALOG USER
MARKETPLACE USER
Data Catalog Enables:
• Consumer: Find, Understand
• Provider: Publish Metadata
EDMP Enables:
• Consumer: Find, Understand, Access & to Use
• Provider: Share Data
ENTERPRISE DATA MARKETPLACE (EDMP)
DATA CATALOG
REGISTER DATA
DISCOVER DATA
MANAGE DATA ORDERS AND ACCESS
COLLABORATION

Fig. 3 Differentiation of the EDMP and the data catalog

Earlier definitions of data catalogs state that these are tools for maintaining an inventory of datasets that are enriched with metadata in order to enable company employees to *discover*, i.e., find and understand, data [59]. New datasets can be *registered* in the catalog by adding the according metadata to the inventory. It also offers other functionality, e.g., for *collaboration* through features like tagging, rating or commenting [34, 59]. As the EDMP is a platform to exchange data, data also has to be found and understood before a user will request access. Hence, the EDMP also has functionality to register and discover data as well as collaborative features [16]. As shown in Fig. 3, the catalog and EDMP, therefore, both provide this functionality. Having understood and selected the data, a user will want to gain access to it. In this regard, the EDMP extends the original data catalog through features for *managing data orders and access*, i.e., requesting data, checking access rights and enabling access to data. In terms of functionality the original data catalog supports data consumers to find and understand data and the EDMP additionally supports gaining access. For the data providers, the catalog enables sharing metadata whereas the EDMP enables sharing metadata as well as the data through its additional order and access functionality.

Data catalogs have, however, evolved in the past years, so these are now also discussed in the context of data access [34]. Thus, the discernment between the catalog and data marketplace becomes unclear. Jahnke and Otto [29] create a topology of data catalog applications, in which they identify the EDMP as one class of data catalog application. Therein, they identify the EDMP as a modular solution that includes the data catalog as a module and an additional brokerage module that enables describing and purchasing data products. This is conform to our understanding as also depicted in Fig. 3, wherein, the data catalog is depicted as a component of an EDMP. However, Jahnke and Otto [29] have also shown that almost 60% of data catalogs now also offer data access functionality. In this regard, we claim, that a data catalog which has evolved in such a way that it now also offers brokerage functionality for managing data orders and access has turned into a rudimentary EDMP.

Besides the offered functionality, the data catalog and EDMP also differ in terms of services offered to users. As explained in Sect. Required data marketplace offerings, the EDMP's full scope of offerings includes Data-as-a-Service, Infrastructure-as-a-Service, Software-as-a-Service, and Professional Services. Basically, the data which can be attained through the EDMP can be acquired with, or provided within infrastructure like a virtual machine or software like Tableau, and with a course, e.g., to learn how to integrate data. The data catalogs, preconditioned they support data access functionality, only offer data-as-a-service. In this regard, the scope of offerings is broader in the EDMP.

In short, some data catalogs are evolving into rudimentary EDMPs by providing functionality for data access management, yet the fundamental purpose of data catalogs remains to foster data transparency by providing a data inventory and enable connecting data supply and demand through data discovery [29]. The EDMP, in addition, aims to achieve data democratization by also supporting the data consumers in the data usage.

If data catalogs continue to progress in such a way that data brokerage becomes their focus and the previously mentioned offerings are also included, we claim that these no longer represent data catalogs, but rather the advanced form, i.e., the EDMP and should be renamed as such. Throughout the rest of this work, the data catalog is, therefore, referred to as a tool for data discovery, so that the EDMP serves as an extension through data order and access management functionality for both providers and consumers, as depicted in Fig. 3. How the catalog and the EDMP can be combined to complement each other is discussed in more detail in the following sections.

Platform architecture

As outlined in Sect. Related work, data marketplace architectures presented in literature thus far provide various perspectives on required components and the component-interactions. These include architectures that illustrate how data marketplaces can be implemented with blockchain [43, 48], architectures that position the data marketplace in IoT ecosystems [46], or architectures that focus on matching supply and demand through a so called data market management system [19]. So far, the mentioned architectures have not considered the special features and requirements of an EDMP.

Therefore, we present a platform architecture that reflects the components of an EDMP, displayed in Fig. 4. Components that are potentially distinctive in the EDMP, e.g., in regard to implementation aspects, are highlighted in grey. How this EDMP platform integrates into the existent system landscape and how the components interact therein is discussed in the following Sect. Enterprise integration.

The architecture distinguishes *frontend* and *backend* components. The frontend is responsible for offering functionality to the EDMP participants and the backend for implementing this functionality through a variety of *services*. The frontend and backend components communicate with each other, e.g., via REST through an API Gateway. In addition, there are storage components for metadata and data. Components labeled as tools or platforms may already exist as standalone solutions within an enterprise. This is a unique characteristic within the enterprise and can be

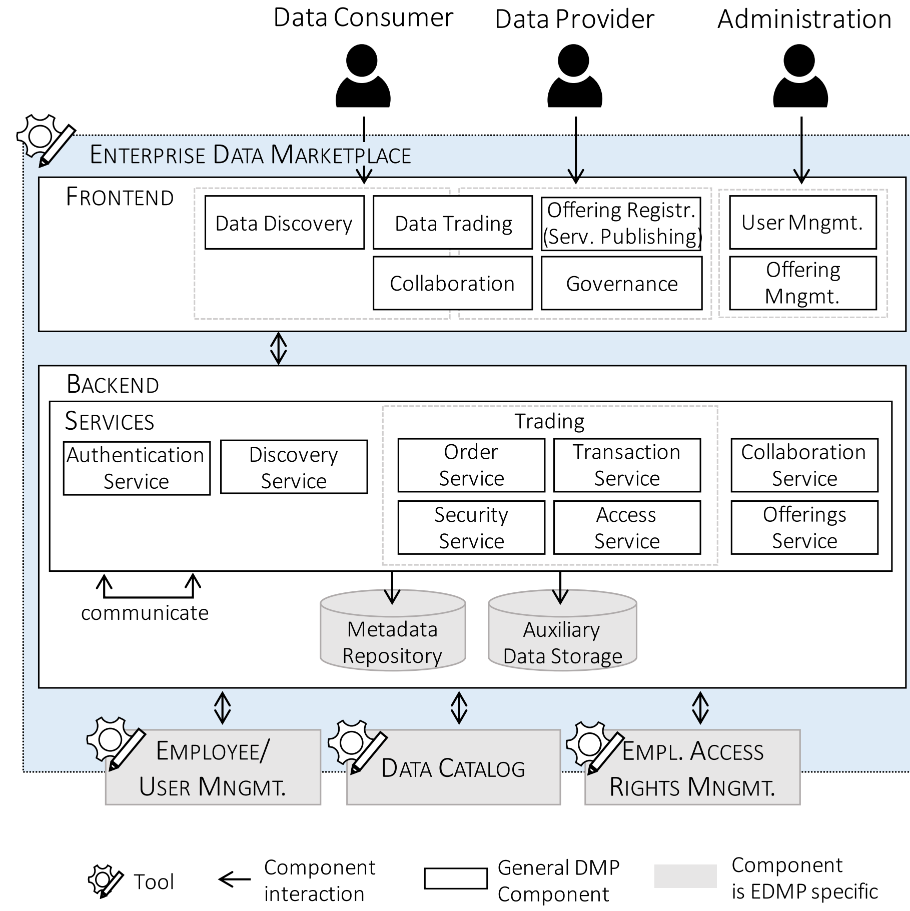


Fig. 4 Enterprise data marketplace architecture featuring a component overview

exploited by tightly integrating the EDMP with the existent solutions as specified in Sect. Enterprise integration requirements. Alternatively, the features of these components can also be implemented within the according backend services yielding a self-sufficient EDMP.

Frontend

The EDMP functionality is available to the roles, data consumer, data provider, and administrator in the frontend through, e.g., a graphical user interface and/or an API. It includes the functionality as described in Sect. Required data marketplace functionality and as listed in the functionality framework [16]. Namely, this is *data discovery, data trading*, and *collaboration* functionality for the data consumer, and complementary, *offerings registration* and *governance* functionality for the data provider, as well as *user* and *offerings management* for the administrators. Since the functionality from the cross-sectional areas, i.e., metadata management and privacy, security and compliance, is not directly accessible to users, it is not represented in the frontend. These are addressed indirectly throughout the backend services.

Backend

The backend provides a variety of services according to the functionality offered through the frontend. The services partially build on and communicate with each other, e.g., via a message broker. There are services for *authentication, discovery, order, security, transaction, access, offerings* and *collaboration* functionality. The authentication service is responsible for managing user access to the EDMP and in this sense handles the registration and login. Search functionality together with a detailed view on offerings is provided through the discovery service. To facilitate trading, several services are required. The creation, monitoring and management of orders and subscriptions is handled through the order service. The security service deals with permission and provision approvals for the orders. This entails topics such as the verification whether a user has appropriate access rights for data with a higher security class. If any form of monetary transaction is called for, this is dealt with by the transaction service, and the access service is accountable for creating and managing access methods such as data base access, or access-links to data. The offerings service is responsible for the registration of any kind of service as described in Sect. Required data marketplace offerings, i.e., data, training courses etc. It adds the data offerings to the data catalog which maintains a data inventory, with according metadata relevant for finding and understanding data and stores additional metadata which is not associated with the catalog, e.g., metadata for accessing the offerings, in the metadata repository. Lastly, the collaboration service takes care of any form of interaction on the offerings such as comments, use-case-documentations or ratings.

Enterprise data marketplace tailored components

The components highlighted in gray in Fig. 4 are required in all data marketplaces, but can be specifically tailored to the enterprise setting, and are therefore termed as EDMP tailored components. For instance, the components marked as tools can be implemented as part of the data marketplace, producing a stand-alone solution which could be used in an external context. These components can, however, already exist within an enterprise, and could therefore be reused and integrated in the EDMP.

The component *employee/user management* is responsible for the identity management and authentication of users, meaning, enterprise employees and invited business partners that have access to the EDMP. Essentially this is the user database. In terms of the data democratization goal, getting access through, e.g., a user account, should be easy and attainable for the employees. As mentioned previously, companies usually have tools to manage information on their employees, such as Employee Database Software[3] which offers a directory of employee profiles and functionality to structure and secure employee data including personal information, qualifications, skills and so on. As the EDMP will require an extract of exactly this metadata, it can be built on such an existing tool instead of recording the same information twice.

Closely related is the component *employee access rights management*, which handles the users authorization, e.g., for various tools and platforms and potentially specific actions therein. Through it users can apply for, attain and manage these rights.

[3] https://www.scnsoft.com/software-development/databases/employee.

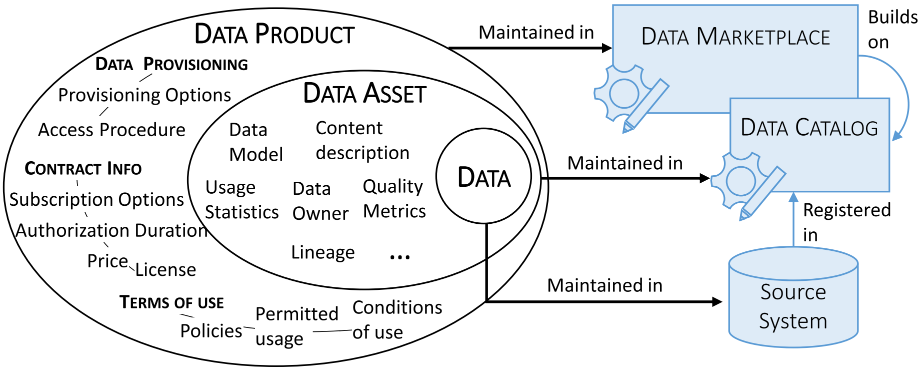


Fig. 5 The distinction of data assets and data products [16]

Like before, there are tools for this on the market that are already used within the enterprises such as Access Rights Manager[4] and could be integrated into the EDMP.

A *data catalog* such as Alation[5] is a tool for maintaining a data inventory and amongst others, offers discovery, administration and governance functionality [34, 59]. Within the marketplace this inventory would reflect the offered data and services like training courses, with according metadata like a content description, the owner, who may access and use it and so on. This inventory can be maintained as part of the marketplace's metadata repository, or could be maintained within an external tool. As companies are in the process of building and maintaining data catalogs [13] the stored information could be reused within the EDMP as opposed to doubling the inventory with collected metadata and functionality. As a data marketplace requires more metadata for data trading than is normally collected within data catalogs this requires a distinction of data which is registered in the data catalog, i.e., a data asset, and data which is explicitly registered in the marketplace, i.e., a data product [16]. This distinction is illustrated in Fig. 5. The product has been prepared for sharing and therefore, provides an extended set of metadata to the asset that explicitly enables the exchange of data, such as information on the license, price, or access options. As discussed in [16] this differentiation of data assets and products and the integration of an existent catalog supports and relieves the data provider, who has the potentially laborious task of making data known and providing provisioning options.

The *metadata repository* stores the metadata which is relevant for operating the EDMP. As data marketplaces are metadata driven platforms [26] this is an essential component. What metadata is maintained in the EDMPs varies depending on whether the above mentioned tools are integrated in the EDMP, or if it is implemented as a stand-alone solution. Besides metadata for cataloging the offerings, user information and access rights, the metadata repository may store metadata on, e.g., the order process, the purchase history, transaction history, or search history.

As explained in Sect. Enterprise data marketplace characteristics an EDMP may have a hybrid architecture with both a centralized and decentralized data storage.

[4] https://www.solarwinds.com/de/access-rights-manager.

[5] https://www.alation.com/.

Most of the offered data should be referenced in the according storage systems, in order to support the scope of most enterprise data, and is therefore, part of the decentralized storage. However, if there is no storage system that can be referenced for certain data, there is the option of loading the data directly into the integrated *auxiliary data storage* of the EDMP. This data storage may be omitted if such data can be loaded into and provided through an external system like a data lake.

The extent to which the EDMP distinctive components constitute an independent tool or have to be implemented in the EDMP also depends on the existing system and tool landscape in the company which we discuss in the following section.

Enterprise integration

In this section, we explain how the EDMP can integrate into a company's existent system and tool landscape, as depicted in Fig. 6, and how this integration can be advantageous. This is distinctive for the EDMP, as stand-alone marketplaces, for instance, for trading data between companies are usually not connected with the various data management systems within the participating companies. For one, this would be challenging for reasons of data security and privacy, but also because the participating organizations have a wide variety of system landscapes that the data marketplace would have to be able to reflect. The typical integration scenarios are derived from our previous work in [13, 15, 16].

Only a few architectures presented in literature consider the marketplace in the context of a company's internal system or tool landscape. Gröger [26] presents the core elements of a data ecosystem with an EDMP, yet states that implementation and integration aspects are yet to be investigated. Wells [56] roughly highlights which technologies are needed within the EDMP components, i.e., data lake management, data pipeline management, data catalogs and data preparation. How the EDMPs interact with existing tools that implement these technologies is not discussed. Therefore, we address this topic in this section.

Integration with data sources

To begin with, we would like to illustrate how the EDMP will be integrated with or reference data within the enterprise source systems. This does not concern the integration of data, but the exchange between the EDMP and these systems. As can be seen in Fig. 6, a wide variety of data sources, such as operational systems, e.g., ERP systems, and analytical systems, such as data lakes, are registered in a data catalog, as currently set up and maintained in many companies [13]. The EDMP references these systems via the data catalog. As discussed previously, only data that cannot be referenced in any external system is loaded and stored in the EDMP. If data cannot be provided in the source systems, there is also the option that these are transferred into another system such as a data lake. The EDMP can then grant access to this new system.

Integration with tools

As stated previously, many companies have a variety of tools that provide functionality which is partly required in the EDMP. This includes functionality in tools for managing

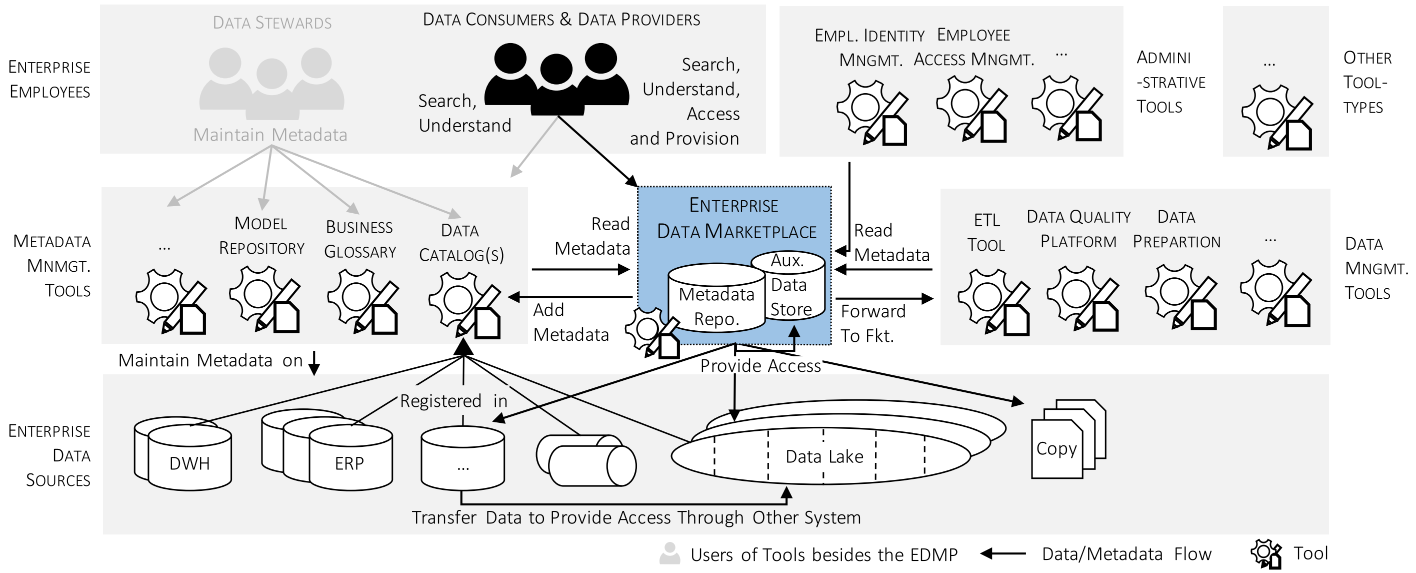


Fig. 6 An illustration of how the enterprise data marketplace integrates with a company's existent system landscape

data and metadata, or administrative tools. Figure 6 indicates how the EDMP interacts with these tool groups.

As the EDMP is a metadata-driven tool [26] most of its functionality is based on metadata. An example of this is the data inventory, which consists of metadata listing available datasets with information such as the storage location. Apart from the auxiliary data store, the EDMP does not interact with the actual data, only with the according metadata. As can be seen in Fig. 6, metadata is collected and maintained in the company through *metadata management tools* such as data catalogs, business glossaries, for defining business terms and term relations [27], model repositories with semantic data models which are integrated with the business glossary [13] and so on. These metadata are relevant in the selection process of a dataset. As described in [15], the distribution of metadata across a wide range of tools is a challenge for data consumers in the process of finding relevant data. For this reason, the EDMP requests the metadata from these tools and provides it in an integrated view. This is a read-only process on these tools. The data catalog is an exception in this context. Since an inventory of data records is already maintained in the catalog, the EDMP builds on this inventory, i.e. when new data is registered in the EDMP, it creates an entry in the existing data catalog for the new dataset, and thus performs a write operation. Although the EDMP extracts metadata from these tools, it is important to note that the metadata will continue to be maintained by the employees within the respective tools. The exception being the data catalog, which metadata is maintained through both the EDMP and catalog. Therefore, the introduction of the EDMP does not change the entire metadata management workflow and the EDMP does not need to provide the functionality of all these different tools. Also, while a consumer can find an integrated version of the metadata in the EDMP, it is still possible to view this metadata in the individual tools.

There are also *data management tools* that collect metadata. These include for instance, ETL tools that can reflect data lineage, or quality management platforms that amongst other things collect quality metrics such as a datasets completeness. As with the other tools, the EDMP can extract metadata from these tools and provide it in the integrated view if these are of interest in the data selection process. Furthermore, as explained in Sect. Required data marketplace functionality, the EDMP is a broker for data between consumers and producers, and does not provide functionality for processing data. It can however, provide the data within an instance of such a tool, e.g., in Tableau, or transfer the consumer to tools with required functionality like data preparation after data acquisition.

In addition to the data and metadata management tools, the EDMP is integrated with *administrative tools* for, e.g., identity management. Thereby, employees only need to acquire the rights to access the EDMP, and the EDMP can then extract employee information from these tools. Based on the extracted information it can for instance, display only those records that match the employee’s clearance level.

Enterprise integration advantages

Integrating the EDMP in the enterprise system and tool landscape has several advantages. For one, *existent functionality is reused*. By building on the existent tools, the EDMP does not double functionality such as access rights management which also avoids the EDMP becoming a jack of all trades monolithic application. Also, there is a *comprehensive view on metadata.* If metadata collected throughout various tools is displayed in an integrated view in the EDMP this provides holistic information on the data. It is, however, important to note that integrating the EDMP with metadata management tools, as well as the integration of the metadata itself is a complex topic which elicits a variety of challenges including the classic data integration problems. Another advantage of integrating the EDMP in the enterprise is a *reduction in metadata management effort and errors.* By reusing metadata already collected within other tools, there is no additional effort for maintaining a redundant set of metadata in the EDMP. This reduces the workload of the data providers that only have to maintain the metadata in one system and is also less error-prone. More information on this can be found in [16]. Finally, there is *less redundant data.* The same is true for the data, when referencing data within the data sources as opposed to uploading the data redundantly in the EDMP, there is less effort on behalf of the providers, reduced storage-cost, no synchronization-efforts and so on.

Prototypical demonstration

To evaluate the presented EDMP concepts, validate their feasibility and further examine the idiosyncrasies of marketplaces used within enterprises, we implemented an Enterprise Data Marketplace prototype. The prototype yields the basis for conducting an experiment for evaluating the impact of introducing an EDMP in the company, as presented in the following Sect. Evaluating an enterprise data marketplace. Mainly the aspects required for validating the presented concepts and evaluating the EDMP's impact in a company are implemented in this prototype. An overview of the prototype is presented in Sect. Prototype overview. Section Application case demonstration demonstrates three typical data marketplace application cases, namely: registering data, searching for, and then requesting access, i.e., ordering this data.

Prototype overview

We based the choice of tools for the prototype on non-commercial and open-source tools because we want to enable free usability and customization. As depicted in Fig. 7, a source system landscape is represented by a variety of database types and a data lake. The databases include the document store MongoDB,[6] the object-relational database PostgreSQL,[7] the columnar database Cassandra[8] and the key-value database Redis.[9] These databases contain a variety of structured, semi- and unstructured sample datasets. In order to explore how an EDMP can reflect the characteristics of specific system types, we have also implemented a data lake. It is realized as a conglomeration of storage systems, including the Hadoop Distributed File System (HDFS)[10] and PostgreSQL, and is based on the data lake zone model by Giebler et al. [22]. Apache Airflow,[11] a workflow

[6] https://www.mongodb.com.

[7] https://www.postgresql.org.

[8] https://cassandra.apache.org.

[9] https://redis.io.

[10] https://hadoop.apache.org.

[11] https://airflow.apache.org.

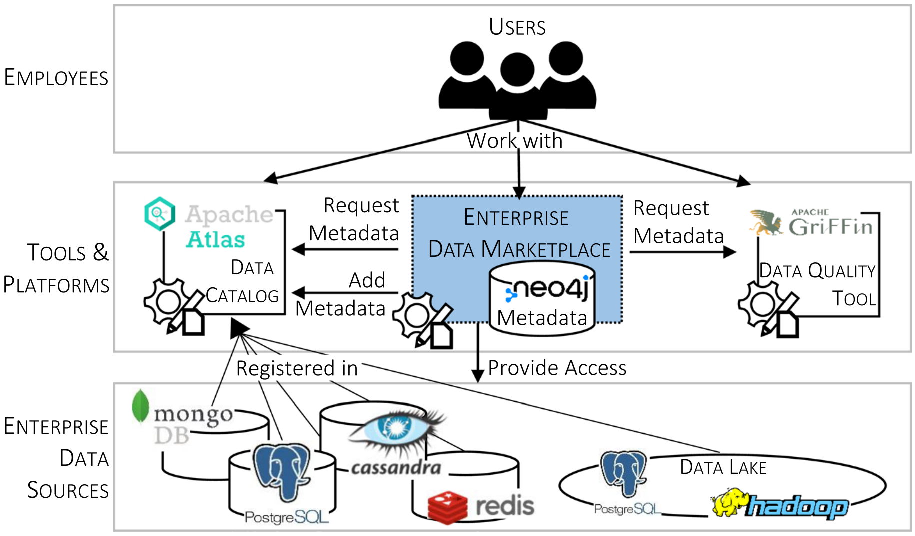


Fig. 7 Tools in the EDMP Prototype. Blue/Dotted Box Represents the Marketplace, as in Fig. 4 and 6

management tool, is used to coordinate processes for moving the data into the appropriate zones based on three exemplary use cases.

The data sources are registered in the open source data catalog apache Atlas.[12] Amongst others, it provides governance and metadata management functionality for building a catalog of data assets. Besides classic metadata such as a content description, our Atlas instance also reflects system specific metadata such as the mapping of data assets to data lake zones. Next to the data catalog we introduced Apache GriFFin[13] into our tool landscape. It is a data quality solution which can measure data quality metrics such as the completeness, accuracy or timeliness of datasets. GriFFin tracks quality metrics on a selection of datasets in our source system landscape.

The EDMP itself is implemented with the Spring framework[14] based on a micro services architecture including an authentication, discovery, order, security, access and offerings service. The services communicate via the message broker RabbitMQ.[15] EDMP specific metadata is stored in a Neo4J[16] graph database and the metadata is modeled according to our metadata model HANDLE [14].

Application case demonstration

Based on three standard application cases of data marketplaces, derived from our previous work in [15, 16], we demonstrate how the EDMP components and enterprise tools interact with each other. In this regard, we present the application case of registering data in the company, how this data can be searched for and found, and finally, ordered. Individual steps of these application cases are exemplified with screenshots of the prototype.

[12] https://atlas.apache.org.

[13] https://griffin.apache.org.

[14] https://spring.io.

[15] https://www.rabbitmq.com.

[16] https://neo4j.com.

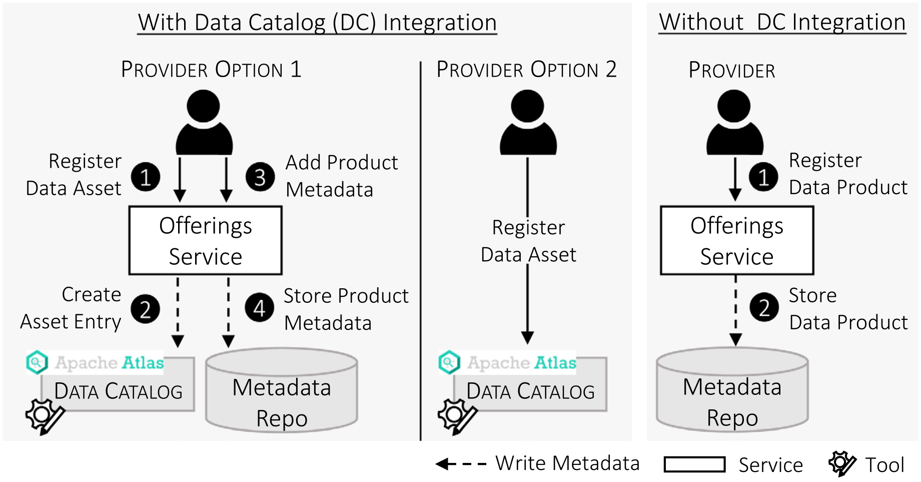
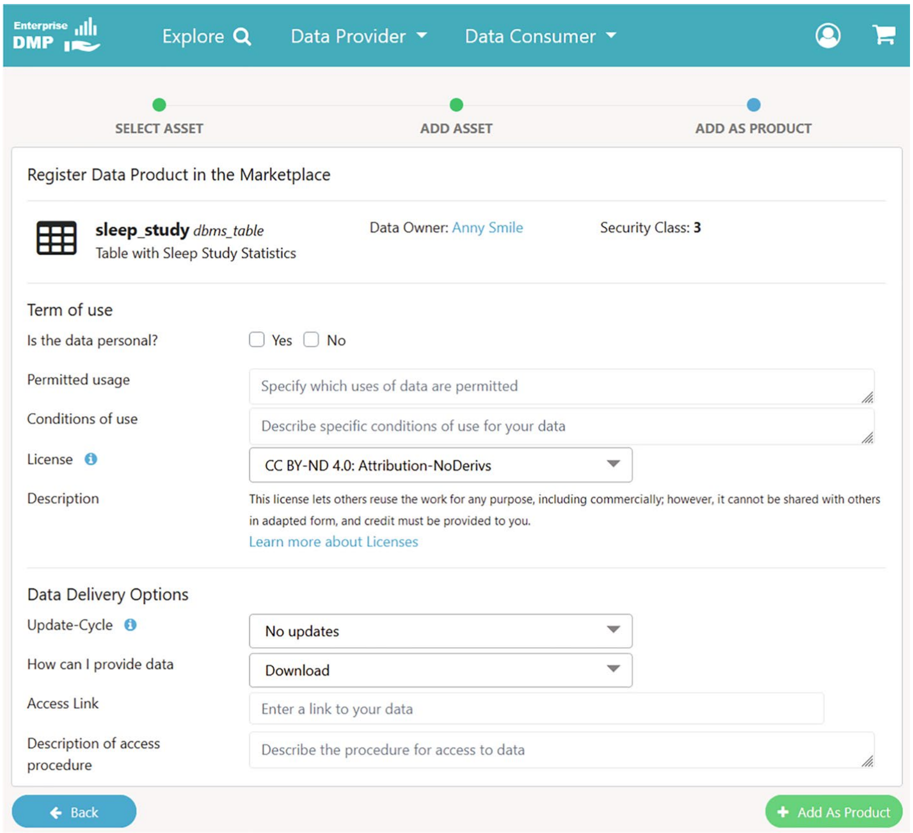


Fig. 8 Data Registration Process Variants and Possible Implementation Variants with and without a Data Catalog



Register Data Product in the Marketplace

sleep_study *dbms_table*
Table with Sleep Study Statistics

Data Owner: Anny Smile

Security Class: **3**

Term of use

Is the data personal? □ Yes □ No

Permitted usage: Specify which uses of data are permitted

Conditions of use: Describe specific conditions of use for your data

License: CC BY-ND 4.0: Attribution-NoDerivs

Description: This license lets others reuse the work for any purpose, including commercially; however, it cannot be shared with others in adapted form, and credit must be provided to you.
Learn more about Licenses

Data Delivery Options

Update-Cycle: No updates

How can I provide data: Download

Access Link: Enter a link to your data

Description of access procedure: Describe the procedure for access to data

Fig. 9 Prototype - Data Product Registration Wizard

Application case 1 - registering data

In order for an EDMP to become effective it needs to have an assortment of offerings. The registration process is different depending on whether the EDMP integrates with an existent data catalog or not, as illustrated in Fig. 8. With a data catalog, a data provider

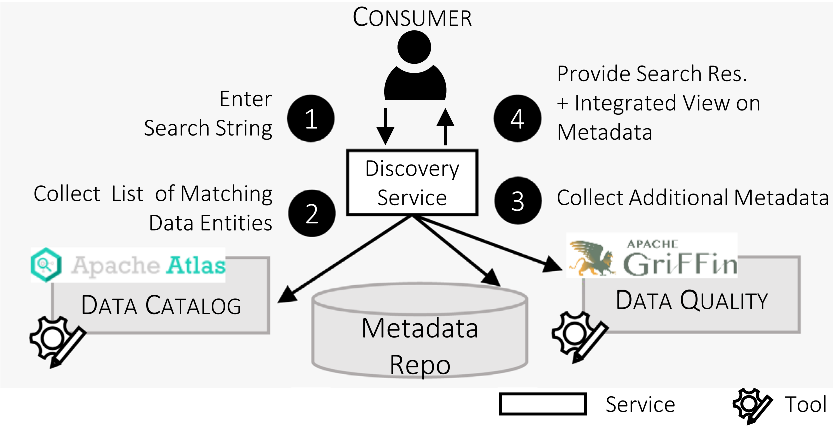
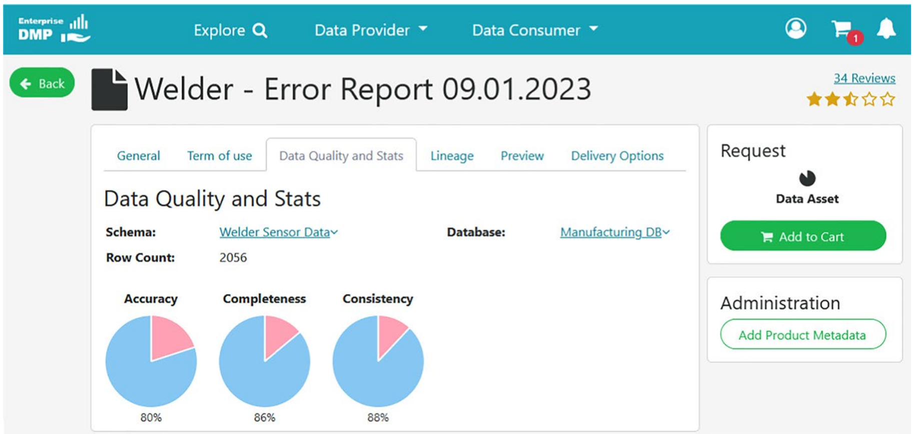


Fig. 10 Search Process for Data with Involved Tools and Components



Welder - Error Report 09.01.2023

Data Quality and Stats

Schema: Welder Sensor Data
Database: Manufacturing DB
Row Count: 2056

Fig. 11 Prototype - Data Asset Information Page

has two options. They can register data through the EDMP as shown in Fig. 8 on the left hand side. They enter asset metadata, meaning, descriptive metadata relevant for understanding the data through a form in step 1. The offerings service then creates an according entry for the data asset in the data catalog, i.e., Atlas, in step 2. At this point the provider can stop as this dataset can be found in the EDMP by potential consumers. Yet, this data is missing product metadata relevant for acquiring it [16]. This could be for instance, the allowed usage, a license, price or subscription or provisioning options. In our prototype this product metadata is added through the product registration wizard as shown in Fig. 9. This constitutes step 3. The product metadata is specific to data trading and thus stored in the EDMP by the offerings service in step 4. At this point the data is ready to be ordered and provisioned to consumers.

Alternatively, the provider can register data directly in the catalog. This is illustrated as provider option 2 in the middle of Fig. 8. As the EDMP is integrated with the catalog this entry can be found in the EDMP, yet, once the data is requested by a consumer, the provider will be prompted to add the product metadata through the EDMP, continuing option 1 at step 3.

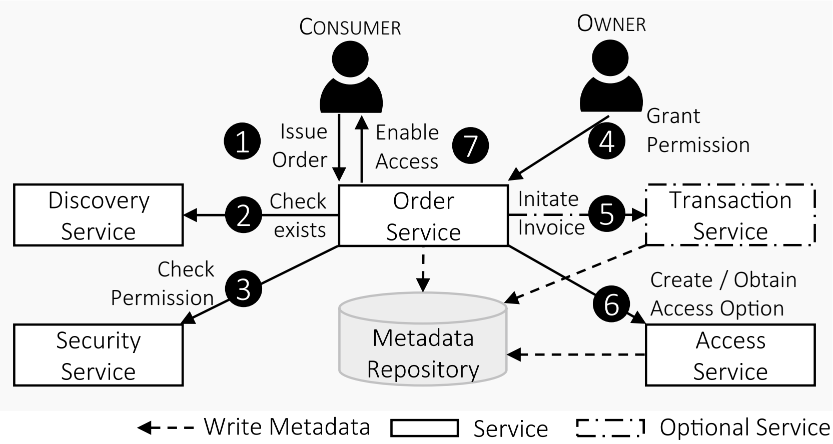


Fig. 12 Data Order Process with Involved Components and Their Interaction Patterns

If the EDMP does not build on an existent data catalog there is no distinction between data assets and data products and the provider only has the option of registering the data through the EDMP, step 1, which will store it in the EDMP's inventory in the metadata repository, step 2, as shown in Fig. 8 on the right.

Application case 2 - searching for data

After registering data in the EDMP it can be found, as displayed in Fig. 10. The consumer enters a request into the frontend search in step 1. Based on the search string the discovery service collects entries from the data catalog, in our case Atlas, in step 2. Then it collects additional metadata such as product metadata from the metadata store and according metadata from other tools such as quality metadata from GriFFin in step 3. A list of search results is returned to the consumer in step 4. As shown in Fig. 11 the single results can be expanded to provide an information page. In this case, the key-value attributes displayed in the central field are extracted from Atlas, whereas the quality metrics are extracted from GriFFin. This information page is one of the features that sets the EDMP apart from external data marketplaces, as it demonstrates how the EDMP can tightly integrate with the existent tools as described in Sect. Enterprise integration and thus, provides a comprehensive view on data assets and products by leveraging and integrating the existent metadata.

Application case 3 - ordering data

In the EDMP's detailed-view-page on data the consumer can add the data to a shopping cart and order it to gain access which is illustrated as step 1. To issue the order the consumer also specifies the intended usage and choses the provisioning option. Figure 12 demonstrates how once the order is submitted the order service checks if the chosen dataset is valid through the discovery service in step 2. After this has been verified the order service transfers the request to the security service in step 3 though which the consumer's permission to access this dataset is checked. For example, this includes a check if the consumer has an adequate clearance level for the dataset's specified security class. If all is adequate, the order services notifies the data owner that they have a new access request. The owner can then grant or deny permission in step 4 based on, e.g., the specified usage information. If monetization is involved the order service initiates the transaction process in step 5 through which an invoice is sent to the consumer. When the transactions are completed the order service forwards the request to the access service as part of step 6. The access service deals with data provisioning options, for example, depending on the chosen and available provisioning options, the access service could create and store an access link which is then forwarded to the consumer in step 7 through the order service. The consumer can now access the ordered data.

Based on the three application cases and the prototype, we have demonstrated how the platform architecture introduced in Sect. Platform architecture can be implemented, how the components interact and how different processes unfold in the EDMP as well as how the EDMP can be integrated with external tools like Atlas and GriFFin. Furthermore, the EDMP prototype, as well as second and third application case presented in this section provide the basis for evaluating the impact of introducing an EDMP into the company tool landscape, discussed in the following section.

Evaluating an enterprise data marketplace

We stipulate throughout this work that the EDMP improves the workflows of data consumers and data providers. In order to verify these assumptions, we leveraged the EDMP prototype described in the previous section to conduct an experiment. The experiment was designed to test the extent to which an EDMP supports and relieves the data consumer in the process of finding and requesting access to data, as described in the Sects. Application case 2 - searching for data and Application case 3 - ordering data.

The research question we aim to resolve reads: *Does the use of an Enterprise Data Marketplace improve the data consumer process of finding, understanding, and requesting access to data?* In this context, we hypothesize that the use of an EDMP improves the consumer process in terms of efficiency, effectiveness, and complexity. We expect that the process will be more efficient, meaning it will involve significantly less time. We also expect that more of the consumers might be effective in the sense that they request access to data that fully matches their requirements, i.e., the correct data assets or products. The complexity signifies how challenging it is for the consumers to identify and request access to data, and how intuitive, laborious and cumbersome they find the overall process. While the efficiency and effectiveness are quantitative dimensions, the complexity is concerned with the portrayal of the qualitative user perspective.

By determining these three measures within the scope of the experiment, we will be able to evaluate whether it is worthwhile to launch an EDMP based on the data consumer's point of view. In the following, we outline the Experiment design, Results, and present the Experiment discussion and conclusion.

Experiment design

To evaluate whether the EDMP improves the consumer process we want to compare the consumer processes of finding and requesting access to data with and without the use of an EDMP. In order that the participants would not already know which dataset to request after performing one of the two variants, we used two identically structured sets of data, that, however, reflect different topic domains. We therefore introduced two scenarios, one without and one with the use of an EDMP. Both scenarios were set in the same enterprise tool and system landscape, except that in one scenario one additional tool was available, i.e., the EDMP. Both scenarios were performed by the same participants, and in both scenarios the participants received the same task, i.e., to find and request access to a specific dataset based on the same set of requirements. The main difference between the two scenarios therefore is the workflow for performing the same task with a different set of tools, i.e., with and without an EDMP. In the following, we provide more details on the data, participants, and procedure involved in the two experiment scenarios, as well as how measurements were taken.

Data: For each of the scenarios, 55 datasets were entered in the prototypical system landscape introduced in Sect. Prototype overview and registered in the data catalog. As in a real-world setting quality information is not available for every dataset, the quality tool calculated different metrics on a selection of these datasets. To ensure comparability between the scenarios, the structure and relationship between the datasets within the scenarios were the same, i.e., both scenarios had the same lineage graph. Participants were only given access to metadata on these datasets during the experiment, thus details on the content of the data are not relevant at this point.

Participants: The experiment was conducted with twelve computer scientists. By choosing subjects that are active within the computer science domain we ensured that the subjects have a basic understanding of what data and data analytics constitute and that they know how to operate a variety of tools, in this case, tools in the context of data management. We thereby eliminate the issue of results being biased due to lacking knowledge of what metadata might be, what the metadata means, or a lacking skill in operating software systems.

Procedure: In both scenarios the participants were tasked to act as data consumers and to find and request access to data. They were given a set of requirements that the data should fulfill. All participants were subjected to both scenarios, from which follows that we chose a “within-subject design” [8], where each participant receives each treatment. This design was chosen so the performance of participants could be compared in both scenarios and so they could be asked to compare the scenarios. To avoid learning effects influencing the results of the second scenario, we switched the order of the two scenarios for 50% of the participants. Hence, 50% started with the marketplace scenario and moved on the scenario without a marketplace, and the other 50% vice versa. Tribal knowledge in companies is often exchanged verbally amongst colleagues. Therefore, one of the authors of this work was available for questions in the role of a colleague working on the same topics throughout the entire experiment, to simulate an environment with co-workers. The two scenarios, the specific tasks, and the tools and system landscapes used therein are presented in the following.

Scenario 1 - Without the Use of an EDMP (S1): This scenario presents the reference scenario in which no EDMP is available to the data consumers. Relating to the prototypical enterprise tool and system landscape as presented in Sect. Prototype overview the participants only get access to the data catalog Apache Atlas and the data quality (DQ)

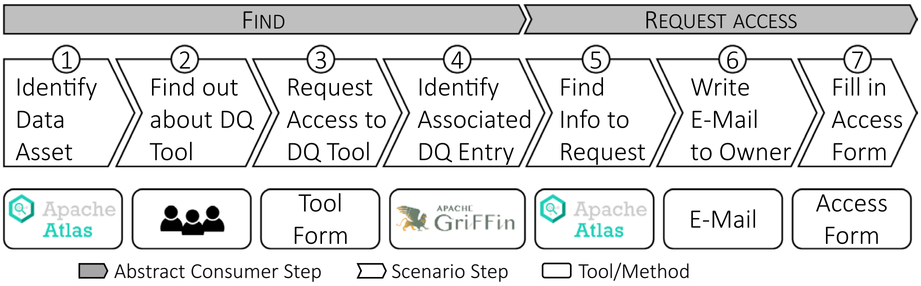


Fig. 13 The workflow and tools without the use of an EDMP

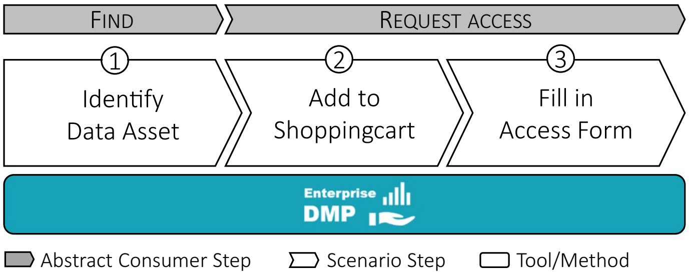


Fig. 14 The workflow and tools with the use of an EDMP

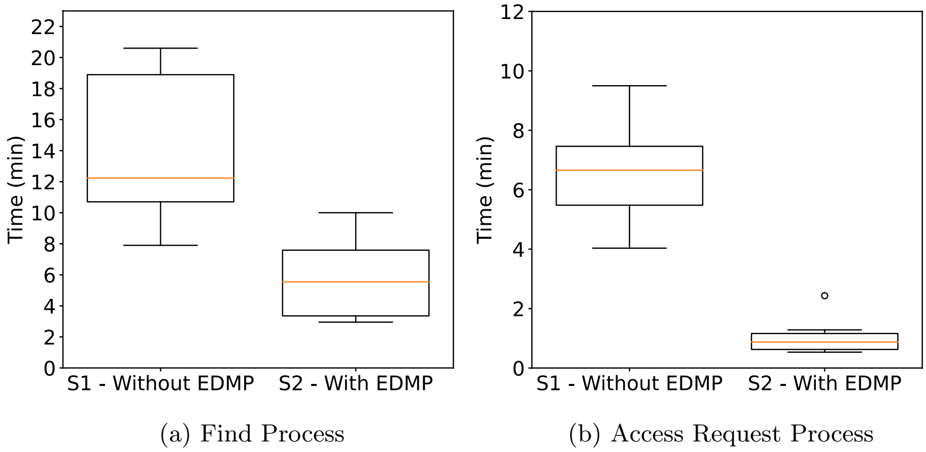
tool Apache GriFFin. In this scenario, the participants represented data scientists working for an IT department working on optimizing public transport schedules. They were given the information that daily reports on buses and trains are stored, recording the intended schedules and GPS location data. Furthermore, they were informed that their company uses a data catalog and other tools such as a business glossary and a data quality platform that contain further metadata. They were also given contact information of a colleague for questions regarding any of the above topics. The specified task in this setting involved finding and requesting access to one of the above mentioned bus or train reports. Only one quality requirement was given specifying that this dataset should be at least 95% accurate. As a oint, they were given a link to the data catalog tool together with user account details.

Figure 13 depicts the workflow they had to figure out to find and request the according data asset. The participants first had to search in the data catalog to find the according data asset based on the name and content description. After realizing that the required metadata on the data quality was not provided through the data catalog, they had to figure out that this metadata is provided through another tool, i.e., the DQ tool. They were then provided with a form to request access to the DQ tool. In companies, access to a tool often has to be granted by a supervisor, which usually takes some time. In the experiment, this was simulated by a one-minute timer after which access details were given to the participants. As, similar to a real-world setting, the metadata is not integrated across the tools, the participants had to decipher which entries in the quality tool belong with which entries in the data catalog. Based on this, they could find a data asset with the required data accuracy. Having identified the required data asset, the participants then had to work out how to request access. For this, the e-mail address of the data owner was provided in the data catalog. After writing an e-mail, the participants were sent a form through which they could request access to the chosen data asset. This scenario simulates a real-world environment in which the tools for searching, understanding, and accessing of data have not been integrated to enable a consistent workflow. As can be seen in Fig. 13 the participants had to find and use a variety of tools and forms, and were reliant on tribal knowledge of colleagues.

Scenario 2 - With the Use of an EDMP (S2): In this scenario, the participants worked with the EDMP prototype as described in Sect. Prototypical demonstration. The setting of this scenario also involved them working as a data scientist in an IT department, this time working on a predictive maintenance use case in a company that manufactures vehicles. The participants were given the information that data from various sensors is collected and error messages from the individual sensors are collected and jointly stored per day in a report for the individual production steps. The participants were also informed that the company tool landscape includes an EDMP.

Like in scenario S1, the participants were tasked with finding one of the above mentioned reports. Again, only one quality requirement was given, stating that the data should be at least 95% complete. As a starting point, the participants were given a link to the EDMP prototype. The EDMP prototype is integrated with the data catalog and DQ tool, therefore, all the required metadata was available through the marketplace. The workflow for this scenario therefore entailed three steps as shown in Fig. 14, all of which could be conducted within the EDMP prototype. The participants first had to use the marketplace search bar to identify a data asset according to the task description. Having found an appropriate data asset, they could add this to the shopping cart in the EDMP. The last step involved filling out and submitting the form to request access in the shopping cart.

Measures: To determine whether the hypothesis holds true, the three metrics, efficiency, effectiveness, and complexity had to be measured. The efficiency relates to the time required to perform the assigned task. We, therefore, logged when which step was started and completed. Based on this log, we could ascertain how long the steps for finding data and requesting data took in both scenarios. The effectiveness can be measured based on whether the correct datasets were requested within the scenarios. In order to determine the complexity of the consumer processes with and without a data marketplace, we had the participants fill out a questionnaire after each scenario with the same set of questions. After completing the second scenario they also filled out a third questionnaire comparing the two scenarios. There were three sets of questions in the scenario-specific questionnaires. The first set concerned the process for finding data, the second set, the process for requesting access, and the third set the overall process. For instance, the participants were asked to disclose whether they found the process intuitive, cumbersome, or laborious, and if it was clear which steps had to be followed for identifying the relevant dataset, or to request access to this dataset. For most of the questions a Likert scale was used to record the answer, in this case with the options: strongly agree, agree, neutral, disagree, and strongly disagree. Additionally the participants had to specify whether they asked for guidance through a yes/no question. The complexity is deduced based on a set of the above-mentioned aspects. The first being how intuitive the participants found the process and whether



(a) Find Process (b) Access Request Process

Fig. 15 Time required for the find and access request process in both scenarios

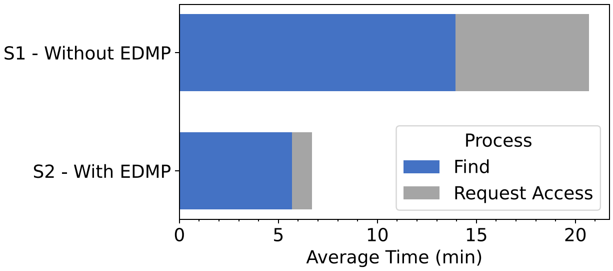


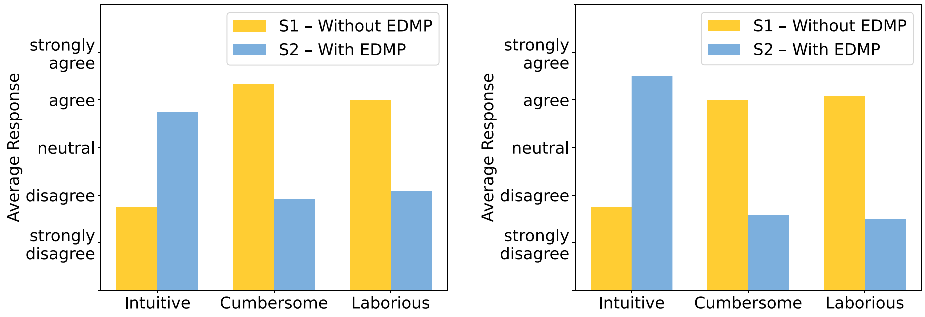
Fig. 16 Scenario runtime comparison

or not it was clear which steps had to be followed to complete the tasks. Also, if the participants found the processes cumbersome, is factored in, meaning the process might have been easy, but entailed many unnecessary steps. Similarly, how laborious they found the process, referring to whether it was resource-intensive in the sense of, e.g., time-consumption. How many participants required guidance to complete the task is also considered in the complexity metric. Lastly, aspects like the variance in time the participants required to complete the task may also indicate that people found the process more or less complex.

Results

In this section, we provide the experiment results for the three dimensions of the hypothesis: efficiency, effectiveness, and complexity. The results are discussed in the following Sect. Experiment discussion and conclusion.

Efficiency: The time it took the participants to find and request access to the data throughout the two scenarios, with and without an EDMP, is visualized in Fig. 15. As can be seen on the left in Fig. 15a, the participants identified the correct data asset in a time range from 7:54 min up to 20:36 min in scenario S1, without an EDMP. The mean therein is a duration of 12:14 min. In scenario S2, with the EDMP, the time span for identifying the data asset ranges from 2:57 min to 10:00 min, the mean therein being 5:33 min. The time required for requesting access to the identified data asset



(a) Find Process (b) Access Request Process

Fig. 17 Participants perception comparison: is the consumer process in the scenarios intuitive, cumbersome, or laborious

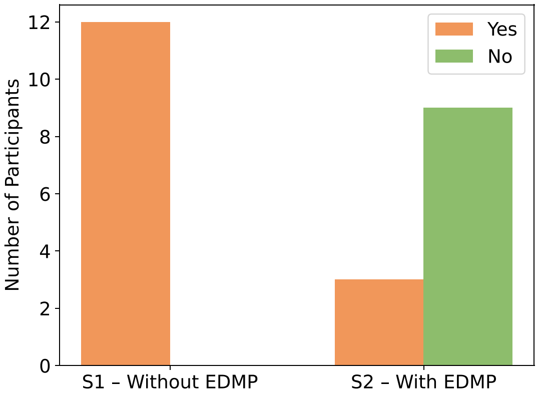


Fig. 18 A comparison of the required guidance in the two scenarios

is shown in Fig. 15b. In scenario S1, without the EDMP, this step took between 4:02 min and 9:30 min, with a mean of 6:40 min. Requesting data in S2, with the EDMP, required between 0:32 min and 1:17 min, with a mean of 0:53 min. There is one outlier, representing a participant that required 2:26 min to request the data asset. In both steps, it can also be observed that the distribution of the values for scenario S1, without an EDMP, is larger than for the processes in scenario S2, with an EDMP. This is especially pronounced in the process of requesting access.

Figure 16 displays the average times for the individual processes, finding in blue, requesting access in grey, as well as the average time it took to complete the entire process for both scenarios. The average time to complete scenario S1, without the EDMP, was 20:41 min. Therein the average time to find the data asset was 13:57 min and 6:44 min to request access to this data asset. In scenario S2, with the EDMP, the overall average time is measured at 6:42 min. Finding the data asset took an average of 5:42 min and requesting this data asset an average of 1:00 min.

Effectiveness: In terms of effectiveness, 100% of the participants requested access to the correct data assets according to the given requirements in both scenarios.

Complexity: Fig. 17 depicts the results of the two questionnaires the participants filled out after completing each scenario. The bars reflect the average answer given

Table 2 Process Simplified through EDMP

Question	Result (Mean)
Simplified finding and understanding	Agree
Simplified requesting data	Strongly agree

Options: strongly disagree, disagree, neutral, agree, strongly agree

for each question. The results were quantified by allotting each response option in the Likert scale to an according number, i.e., 1 for strongly disagree, 2 for disagree, etc. While this enables quantifying the results given throughout the questionnaires more precisely we will discuss the rounded average in the following. Regarding the statement that *the process for finding and understanding data is intuitive*, the participants disagree in scenario S1, without the EDMP, and agree in scenario S2, with the EDMP. For both the statements that *this process to find data is cumbersome or laborious*, the participants agree in scenario S1 and disagree in scenario S2. As illustrated in Fig. 17b, the results concerning the access request process are similar. The participants disagree that the access request process in scenario S1, without the EDMP, is intuitive, yet strongly agree that it is intuitive in scenario S2, with the EDMP. They also agree that the access request process was cumbersome and laborious in scenario S1, yet disagree that this is the case in scenario S2. Furthermore, the average answers given to the statement *that it is clear which steps had to be followed to complete the processes* yielded a disagree for the find and access process in scenario S1, without the EDMP. In contrast in scenario S2 with the EDMP, the participants agree concerning the finding process and strongly agree for the access request process.

Figure 18 depicts that out of a dozen participants all required and asked for guidance to find and request data in scenario S1. Not one participant completed the scenario independently. In comparison, only three participants required guidance in scenario S2, and nine were able to complete the process independently.

Having completed both scenarios the participants filled in a third questionnaire with only two questions comparing both scenarios. As before the Likert Scale results were quantified. Table 2 lists the rounded mean result of the questions regarding the simplification of the consumer process with the use of an EDMP. The participants agree that the EDMP simplified the process for finding and understanding data, and strongly agree that it simplifies requesting access to the data.

Experiment discussion and conclusion

In this section we evaluate whether the use of an EDMP improves the efficiency, effectiveness, and complexity of the consumer process and consequently, whether the hypothesis for this experiment holds true. To this end, the results of scenario S1 and scenario S2 are compared. For this comparison, the scenarios were designed as similarly as possible, involving the same task with the same requirement, identically structured sets of data, and the same participants. They differed mainly in their workflow, which is based on the use of different tools, i.e., once with and once without an EDMP.

Efficiency: With an average of 5:42 min, as opposed to 13:57 min, the process for finding data is more than twice as fast with an EDMP. Although the fastest person in

Table 3 Hypothesis evaluation: the EDMP improves the consumer process in terms of efficiency, effectiveness, and complexity

Efficiency	Effectiveness	Complexity
✓	-	✓

✓ improved, - unchanged

scenario S1, without an EDMP, is faster than the slowest person in scenario S2, with an EDMP, the smaller standard deviation in scenario S2 still indicates that people are generally faster with the EDMP. This is most likely the case as the required metadata, i.e., content descriptions and quality metrics were supplied and integrated within the EDMP. Therefore, the EDMP can also offer additional filter functionality in its discovery service, e.g., to filter for data with a specific completeness level. Furthermore, the participants only had to figure out how to operate one tool as opposed to several. The EDMP therefore not only offers a variety of metadata in one place but also provides additional search functionality and supports the workflow for finding data throughout one tool.

Similarly, the process to request access to data is faster in the EDMP with an average of 1:00 min as opposed to 6:42 min without an EDMP. Not only were all participants faster in scenario S2, but as visible in Fig. 15b, the standard deviation for scenario S2 is a lot smaller than for scenario S1, indicating that the participants were similarly fast with little deviation. This time difference is most likely due to two factors: Firstly, in scenario S2 the participants did not have to figure out how to request the data as in scenario S1, but were guided through the process. Secondly, in scenario S1 the access request workflow involved several tools and forms which were not integrated, making the process more complex and therefore more time consuming. In contrast, in scenario S2 the EDMP supported the workflow in one tool.

We can therefore deduce that the overall consumer process with the use of an EDMP is more efficient, with an average duration of 6:42 min, than without an EDMP, with the average duration of 20:41 min.

Effectiveness: Since in both scenarios the correct data assets were requested in 100% of the cases, we cannot definitively deduce with these parameters that the EDMP increases the effectiveness of the data consumer. We assume that the extended time to find data enabled the same level of effectiveness. Had there been a time constraint, we assume the participants would not have had the time to familiarize themselves with both tools and the provided metadata and, therefore, might have requested a data asset that met some but not all of the requirements. In this experiment, the choice would most likely have been based on the content description in the data catalog, whereas the quality information, which was harder to attain, would most likely have been disregarded. Therefore, with enough time, both scenarios are equally effective, but with a time constraint, we assume that the marketplace would be more effective.

Complexity: Given that in the EDMP context the participants on average agree that the process to find data is intuitive and strongly agree that the access request process is intuitive, whereas they disagree on both accounts without the EDMP, we deduce that the EDMP increases the intuitiveness of the consumer process. We argue that this is due to the fact that it integrates with the available tools and thereby offers an integrated view on metadata, and also supports the overall workflow of the data consumer process.

Thereby, the users are guided through a set of steps as opposed to having to decipher the next steps by themselves. This is also reflected in the average answers given whether it was clear which steps had to be followed to complete the processes. We assume that the access request process was especially intuitive to the participants as the data was ordered as in online shops through the usual shopping cart workflow. Based on these results we conclude that the use of an EDMP makes the consumer process more intuitive and, therefore, less complex.

As the consumer process was perceived as less cumbersome and laborious in scenario S2, the EDMP seems to decrease the complexity also in this regard. Furthermore, as all participants required guidance in scenario S1, whereas only three required guidance in scenario S2 and nine conducted the process without help, this also underlines that the process is less complex with an EDMP. Lastly, the greater standard deviations for the find and access request process for scenario S1, as illustrated in Fig. 15, suggests that the participants were challenged to varying degrees. Since the standard deviation with an EDMP is reduced, it can be argued that the marketplace reduces complexity so that the performance of the participants converges.

As the consumer process is more intuitive, less cumbersome and laborious, requires less guidance, and reduces the deviation in performance, we conclude that the consumer process becomes less complex through the use of an EDMP.

Lessons Learned: In the scope of this experiment, we have established that an EDMP makes the data consumer process more efficient and less complex. The hypothesis that this marketplace improves the process in terms of efficiency, effectiveness, and complexity, does not hold true, as only two out of these three aspects are improved, as summarized in Table 3. The effectiveness is not explicitly improved or reduced, it remains the same with and without an EDMP. Yet, we assume that given a time constraint for finding data the EDMP would be more effective. Based on the results of this experiment, we conclude that the introduction of an EDMP significantly benefits data consumers. In the context of a companies' goal to democratize their data, the consumer process relates to the data democratization dimension one, which involves the accessibility of data [37]. Based on this experiment, we therefore stipulate that the EDMP addresses the first democratization dimension by improving the data consumer process. With the prototype and the experiment we have consequently demonstrated the technical feasibility of the presented EDMP concepts and that an EDMP significantly furthers the data democratization initiative.

Evaluating the significance of an EDMP for the provider would mainly include registering data assets and products. In order to register data assets, data providers must be very familiar with these assets to be able to supply various metadata in various tools. If the providers are not familiar with the data, they are reliant on other experts to supply this information. As the simulation would have involved participants that are not actually familiar with the data, a lot of the provider's tasks would have to be realized for them, distorting the effort and process. Therefore, the scenario of registering data assets and products could not be reasonably modeled in a realistic way in an experiment. For this reason, the experiment focused on the consumer side and the evaluation of the provider side constitutes future work, for instance, by conducting a field study in a real-world environment.

Summary and conclusion

Enterprise Data Marketplaces for exchanging data within companies are becoming increasingly relevant as they support data democratization and consequently contribute to extracting more of a company's potential data value. In this paper, we have established that the EDMP is a distinct type of marketplace with specific characteristics. This was clarified by placing the EDMP in a classification framework, by distinguishing it from the related tool type of data catalogs and by highlighting a set of requirements which are specific to the EDMP. By presenting a platform architecture, discussing how this platform integrates with existent enterprise system landscapes and demonstrating these concepts through a prototype, we laid the foundations for the development of an EDMP. Moreover, based on the conducted experiment we unveiled that an EDMP makes the data consumer process of finding and requesting data more efficient and less complex. It can also be argued that given a time contraint the data consumers become more effective by using an EDMP. We therefore conclude that the use of an EDMP significantly benefits a company by improving the data consumer and provider processes and consequently, is an essential constituent in any data democratization initiative for empowering employees to find, understand, access, use, and share company data.

Implementing an EDMP prototype and an in-depth knowledge exchange with the industrial manufacturer investigating EDMPs revealed that there are still a number of challenges to be addressed when using a marketplace in the enterprise internal context. These include topics like incentivizing data providers to share their data, finding, assigning and retaining data ownership, and preventing the flooding of the EDMP with unusable data. Another of the challenges involves integrating the EDMP into the existing system landscape. In future, we intend to address this challenge, especially with regard to the topic of metadata management, by investigating how metadata from a variety of tools can be modeled and displayed in an integrated view in the EDMP. Furthermore, a detailed examination how privacy and security aspects are handled in the EDMP is also subject to future work. In closing, this work, based on both research and a practical viewpoint from industry, reveals that establishing an EDMP requires a variety of interdisciplinary perspectives, ranging from business economics to address, e.g., issues of data valuation, legal aspects to enable, e.g., legitimate data sharing and ownership issues, aspects from information systems such as the design of incentivation mechanisms for sharing data, or also more technical aspects concerning the implementation of a EDMP.

Abbreviations

IoT	Internet of Things
ERP	Enterprise Resource Planning
EU	European union
EDMP	Enterprise Data Marketplace
DMP	Data marketplace
GDPR	General data protection regulation
HDFS	Hadoop Distributed File System

Acknowledgements

Not applicable.

Author contributions

RE conducted the research, was responsible for the prototypical implementation and wrote the manuscript. CG, EH, CS, HS and BM were involved in various significant discussions regarding the research direction and activities and contributed their expertise on topics such as privacy and security or enterprise and software architecture and thereby highly

improved the quality of the manuscript. HS, CG and BM took on a supervisory role and oversaw the completion of the work. CS also edited the manuscript. All authors reviewed and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Ccompeting interests
The authors declare that they have no competing interests.

Received: 9 June 2023 Accepted: 16 October 2023
Published online: 24 November 2023

References

1. Alpers S, Betz S, Fritsch A, et al. Citizen empowerment by a technical approach for privacy enforcement. In: Proceedings of the 8th International Conference on Cloud Computing and Services Science, CLOSER' 18; 2018. pp 589–595. https://doi.org/10.5220/0006789805890595
2. Alrawahi AS, Lee K, Lotfi A. AMACoT: a marketplace architecture for trading cloud of things resources. IEEE Int Things J. 2020;7(3):2483–95. https://doi.org/10.1109/JIOT.2019.2957441.
3. Anhalt-Depies C, Stenglein JL, Zuckerberg B, et al. Tradeoffs and tools for data quality, privacy, transparency, and trust in citizen science. Biol Cons. 2019;238: 108195. https://doi.org/10.1016/j.biocon.2019.108195.
4. Ardagna CA, Cremonini M, Damiani E, et al. Location privacy protection through obfuscation-based techniques. In: Data and Applications Security XXI. Springer Berlin Heidelberg; 2007. pp 47–60, https://doi.org/10.1007/978-3-540-73538-0_4
5. Awasthi P, George J. A case for Data Democratization. In: Proceedings of the 26th Americas Conference on Information Systems, AMCIS '20. 2020; pp 23:1–23:10
6. Azcoitia SA, Laoutaris N. A survey of data marketplaces and their business models. SIGMOD Rec. 2022;51(3):18–29. https://doi.org/10.1145/3572751.3572755.
7. Cao L. Data science: a comprehensive overview. ACM Comput Surv. 2017. https://doi.org/10.1145/3076253.
8. Charness G, Gneezy U, Kuhn MA. Experimental methods: between-subject and within-subject design. J Eco Behav Organ. 2012;81(1):1–8. https://doi.org/10.1016/j.jebo.2011.08.009.
9. Clifton C, Kantarcioundefinedlu M, Doan A, et al. Privacy-preserving distributed mining of association rules on horizontally partitioned data. IEEE Trans Knowl Data Eng. 2004. https://doi.org/10.1145/1008694.1008698.
10. Dehghani Z, Fowler M. Data Mesh: Delivering Data-driven Value at Scale. O'Reilly Media.2022.
11. Driessen S, Monsieur G, van den Heuvel WJ. Data Product Metadata Management: an Industrial Perspective. In: Workshop Proceedings of the 20th International Conference on Service-Oriented Computing, ICSOC Workshops '22. 2023; pp 237–248
12. Driessen SW, Monsieur G, Van Den Heuvel WJ. Data market design: a systematic literature review. IEEE Access. 2022;10:33123–53. https://doi.org/10.1109/ACCESS.2022.3161478.
13. Eichler R, Giebler C, Gröger C, et al. Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges. In: Proceedings of the 24th International Conference on Business Information Systems, BIS '21. 2021a; pp 269–279, https://doi.org/10.52825/bis.v1i.47
14. Eichler R, Giebler C, Gröger C, et al. Modeling Metadata in Data Lakes – A Generic Model. Data & Knowledge Engineering. 2021b; 136(C):101931. https://doi.org/10.1016/j.datak.2021.101931
15. Eichler R, Göger C, Hoos E, et al. Data Shopping – How an Enterprise Data Marketplace supports Data Democratization in Companies. In: Proceedings of the 2022 CAiSE Forum on Intelligent Information Systems, CAiSE Forum '22. 2022a; pp 19–26, https://doi.org/10.1007/978-3-031-07481-3_3
16. Eichler R, Gröger C, Hoos E, et al. From Data Asset to Data Product – The Role of the Data Provider in the Enterprise Data Marketplace. In: Proceedings of the 16th Symposium and Summer School On Service-Oriented Computing. Springer, SummerSoc '22. 2022b; pp 119–138
17. European Parliament and Council of the European Union. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (Data Protection Directive). Legislative Acts L119, Official Journal of the European Union. 2016.
18. Fan L. Practical image obfuscation with provable privacy. In: Proceedings of the 2019 IEEE International Conference on Multimedia and Expo, ICME '19. 2019; pp 784–789, https://doi.org/10.1109/ICME.2019.00140
19. Fernandez RC, Subramaniam P, Franklin MJ. Data market platforms: trading data assets to solve data problems. Proc VLDB Endowment. 2020;13(12):1933–47. https://doi.org/10.14778/3407790.3407800.
20. Fruhwirth M, Rachinger M, Prlja E. Discovering business models of data marketplaces. In: Proceedings of the 53rd Hawaii International Conference on System Sciences, HICSS '20. 2020; pp 5738–5747

21. Ge C, Susilo W, Baek J, et al. Revocable attribute-based encryption with data integrity in couds. IEEE Trans Dependable Secure Compu. 2022;19(5):2864–72. https://doi.org/10.1109/TDSC.2021.3065999.
22. Giebler C, Gröger C, Hoos E, et al. A Zone Reference Model for Enterprise-Grade Data Lake Management. In: Proceedings of the 24th IEEE Enterprise Computing Conference, EDOC '20. 2020; pp 57–66, https://doi.org/10.1109/EDOC49727.2020.00017
23. Giebler C, Gröger C, Hoos E, et al. The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture. In: Proceedings der 19. Fachtagung Datenbanksysteme für Business, Technologie und Web, BTW '21. 2021; pp 351–370, https://doi.org/10.18420/btw2021-19
24. Gritti C. Publicly Verifiable Proofs of Data Replication and Retrievability for Cloud Storage. In: Proceedings of the 2020 International Computer Symposium, ICS '20. 2020; pp 431–436, https://doi.org/10.1109/ICS51289.2020.00091
25. Gritti C, Önen M, Molva R. Privacy-Preserving Delegable Authentication in the Internet of Things. In: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19. 2019; pp 861–869, https://doi.org/10.1145/3297280.3297365
26. Gröger C. There is no AI without data. Commun ACM. 2021;64(11):98–108. https://doi.org/10.1145/3448247.
27. Henderson D, Earley S, Sebastian-Coleman L, editors. DAMA-DMBOK: Data Management Body of Knowledge. Basking Ridge, NJ, USA: Technics Publications; 2017.
28. ISO, IEC 27000:2018(en,. Information Technology-Security Techniques-Information Security Management Systems-Overview and Vocabulary. International Organization for Standardization: Iso standard; 2018.
29. Jahnke N, Otto B. Data catalogs in the enterprise: applications and integration. Datenbank-Spektrum. 2023;23:89–96. https://doi.org/10.1007/s13222-023-00445-2.
30. Janssen M, Charalabidis Y, Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government. Inform Syst Manage. 2012;29(4):258–68. https://doi.org/10.1080/10580530.2012.716740.
31. Kassa YM, Gonzalez J, Cuevas A, et al. Your Data in the Eyes of the Beholders: Design of a Unified Data Valuation Portal to Estimate Value of Personal Information from Market Perspective. In: Proceedings of the 2016 11th International Conference on Availability, Reliability and Security, ARES '16. 2016; pp 701–705, https://doi.org/10.1109/ARES.2016.55
32. Koutroumpis P, Leiponen A, Thomas LDW.The (Unfulfilled) Potential of Data Marketplaces. ETLA Working Papers 53, The Research Institute of the Finnish Economy (ETLA). 2017.
33. Krishnamachari B, Power J, Kim SH, et al. I3: An IoT marketplace for smart communities. In: Proceedings of the 16th ACM International Conference on Mobile Systems, Applications, and Services, MobiSys '18. 2018; pp 498–499, https://doi.org/10.1145/3210240.3223573
34. Labadie C, Legner C, Eurich M, et al. FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs. In: Proceedings of the IEEE 22nd Conference on Business Informatics, CBI '20. 2020; pp 201–210, https://doi.org/10.1109/CBI49978.2020.00029
35. Lange J, Stahl F, Vossen G. Datenmarktplätze in verschiedenen Forschungsdisziplinen: Eine Übersicht. Informatik-Spektrum. 2018;41:170–80. https://doi.org/10.1007/s00287-017-1044-3.
36. Lawrenz S, Sharma P, Rausch A. Blockchain Technology as an Approach for Data Marketplaces. In: Proceedings of the 2019 International Conference on Blockchain Technology, ICBCT '19. 2019; pp 55–59, https://doi.org/10.1145/3320154.3320165
37. Lefebvre H, Legner C, Fadler M. Data democratization : toward a deeper understanding. In: Proceedings of the 2021 International Conference on Information Systems, ICIS '21. 2021; p 2106
38. Meisel L, Spiekermann M. Datenmarktplätze - Plattformen für Datenaustausch und Datenmonetarisierung in der Data Economy. Fraunhofer ISST: Isst-bericht; 2019.
39. Mezzetta S. Principles of Data Fabric: Become a data-driven organization by implementing Data Fabric solutions efficiently. Packt Publishing. 2023.
40. Otto B, Jürjens j, Schon J, et al. Industrial Data Space - Digitale Souveränit Über Daten. Tech. rep. 2016.
41. Otto B, Steinbuß S, Teuscher A, et al. IDS reference architecture model version 3.0. Steinbuss, S. (ed.) International Data Spaces Association. 2019.
42. Pourahmadi M. Estimation and interpolation of missing values of a stationary time series. J Time Series Anal. 1989;10(2):149–69. https://doi.org/10.1111/j.1467-9892.1989.tb00021.x.
43. Ramachandran GS, Radhakrishnan R, Krishnamachari B. Towards a Decentralized Data Marketplace for Smart Cities. In: Proceedings of the 2018 IEEE International Smart Cities Conference, ISC2 '19. 2018. pp 1–8, https://doi.org/10.1109/ISC2.2018.8656952
44. Ramosaj B, Pauly M. Predicting missing values: a comparative study on non-parametric approaches for imputation. Comput Stat. 2019;34:1741–64. https://doi.org/10.1007/s00180-019-00900-3.
45. Roman D, Stefano G. Towards a reference architecture for trusted data marketplaces: The credit scoring perspective. In: Proceedings of the 2nd International Conference on Open and Big Data, OBD '16. 2016; pp 95–101, https://doi.org/10.1109/OBD.2016.21
46. Schmid S, Bröring A, Kramer D, et al. An architecture for interoperable IoT Ecosystems. In: Proceedings of the 2nd International Workshop on Interoperability and Open-Source Solutions for the Internet of Things, InterOSS-IoT '17. 2017; pp 39–55
47. Schomm F, Stahl F, Vossen G. Marketplaces for data: an initial survey. ACM SIGMOD Record. 2013;42(1):15–26. https://doi.org/10.1145/2481528.2481532.
48. Sharma P, Lawrenz S, Rausch A. Towards Trustworthy and Independent Data Marketplaces. In: Proceedings of the 2020 2nd International Conference on Blockchain Technology, ICBCT '20. 2020; pp 39–45, https://doi.org/10.1145/3390566.3391687
49. Spiekermann M. Data marketplaces: trends and monetisation of data goods. Intereconomics. 2019;54:208–16. https://doi.org/10.1007/s10272-019-0826-z.
50. splunk (2019) The State of Dark Data. Report
51. Stach C. Data is the new oil sort of: a view on why this comparison is misleading and its implications for modern data administration. Future Int. 2023. https://doi.org/10.3390/fi15020071.

52. Stahl F, Schomm F, Vossen G, et al. A classification framework for data marketplaces. Vietnam J Comp Sci. 2016;3:137–43. https://doi.org/10.1007/s40595-016-0064-2.
53. Stahl F, Schomm F, Vomfell L, et al. Marketplaces for digital data: Quo Vadis? Comp Inform Sci. 2017;10(4):22–37. https://doi.org/10.5539/cis.v10n4p22.
54. Täuscher K, Laudien SM. Understanding platform business models: a mixed methods study of marketplaces. Eur Manage J. 2018;36(3):319–29. https://doi.org/10.1016/j.emj.2017.06.005.
55. Wells D. The Rise of the Data Marketplace: Data as a Service. Eckerson Group: Report; 2017.
56. Wells D. Dynamic Data Marketplace: Fast Data for Fast Business. Eckerson Group: Report; 2018.
57. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018. https://doi.org/10.1038/sdata.2016.18.
58. Yu J, Zhang B, Kuang Z, et al. iprivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning. IEEE Trans Inform Forensics Secur. 2017;12(5):1005–16. https://doi.org/10.1109/TIFS.2016.2636090.
59. Zaidi E, De Simoni G, Edjlali R, et al. Data Catalogs Are the New Black in Data Management and Analytics. Gartner: Gartner research; 2017.
60. Zasadzinski M, Theodoulou M, Thurner M, et al. The trip to the enterprise gourmet data product marketplace through a self-service data platform. arXiv. 2021. https://doi.org/10.48550/arXiv.2107.13212.
61. Zhu H, Yuan Y, Chen Y, et al. A secure and efficient data integrity verification scheme for cloud-IoT based on short signature. IEEE Access. 2019;7:90036–44. https://doi.org/10.1109/ACCESS.2019.2924486.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rebecca Eichler is a PhD student at the Institute of Parallel and Distributed Systems of the University of Stuttgart, Germany. She received her master's degree in computer science at the University of Stuttgart in the year 2019. Her research interests include metadata management in complex enterprise data landscapes, metadata management specifically for data lakes, and the topic of democratizing data through data marketplaces.

Christoph Gröger is enterprise architect for data analytics and a senior technical professional in the data strategy team at Bosch, a leading global supplier of technology and services. He holds a doctoral degree in computer science from the University of Stuttgart, Germany. Christoph's overall area of expertise is industrial analytics, i.e., data management and data analytics in industrial value chains, with current focus on data lakes and data strategies.

Eva Hoos is lead architect for data analytics at the Robert Bosch GmbH. She earned her doctoral degree in computer science in the area of smart engineering apps at the Graduate School of Excellence Manufacturing Engineering, University of Stuttgart, Germany. It motivates her the vision to generate value out of data. She is an expert for data analytics platforms, including amongst others data lakes and data marketplaces.

Christoph Stach received his PhD in computer science from the University of Stuttgart for his research in the area of information security and data privacy in mobile applications. From June 2020 to September 2021, he held the deputy professorship in Data Engineering at the University of Stuttgart. Today, he is head of the working area of Information Systems and Applications at the Applications of Parallel and Distributed Systems department of the University of Stuttgart. His current research focuses on concepts and tools required to enable trustworthy and demand-oriented data provisioning.

Holger Schwarz is professor at the Institute of Parallel and Distributed Systems of the University of Stuttgart, Germany. He holds a doctoral degree in computer science from the same university. His current research interests include data modelling, complex data management landscapes and metadata management. He also works on explorative data analytics with a focus on clustering techniques.

Bernhard Mitschang is professor for Database and Information Systems and head of the department 'Applications of Parallel and Distributed Systems' that is part of the Institute of Parallel and Distributed Systems at the University of Stuttgart, Germany. Both research and teaching spectra of his department cover on one hand data-intensive applications ranging from business applications to engineering systems and on the other hand fundamental data management techniques, data analytics as well as scalable data processing architectures. Since 2013, he is CEO of the Graduate School of Excellence on advanced Manufacturing Engineering and head of the Technology Partnership Lab at the university.